2017-1

Evaluating the Effectiveness of the Gestalt Principles of Perceptual Observation for Virtual Reality User Interface Design

William MacNamara
Technological University Dublin

Recommended Citation

MacNamara, W. (2016) *Evaluating the Effectiveness of the Gestalt Principles of Perceptual Observation for Virtual Reality User Interface Design*. Masters thesis, 2016.

Evaluating the Effectiveness of the Gestalt Principles of Perceptual Observation for Virtual Reality User Interface Design

William MacNamara

A dissertation submitted in partial fulfilment of the requirements of
Dublin Institute of Technology for the degree of
M.Sc. in Computing (Advanced Software Development)

2016

DECLARATION

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Advanced Software Development), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

The work reported on in this dissertation conforms to the principles and requirements of the Institute's guidelines for ethics in research.

Signed: ________________________________

William MacNamara

Date: ____/____/________

ABSTRACT

There is a lot of interest and excitement surrounding the areas of Virtual Reality and Head-Mounted Displays with the recent releases of devices such as the Oculus Rift, Sony PSVR and the HTC Vive. While much of the focus for these devices has been related to sectors of the entertainment industries, namely the cinema and video game industries, there are many more practical applications for these technologies, with potential benefits in educational, training/simulation, therapeutic and modelling/design software. Developing a set of reliable guidelines for Virtual Reality User Interface Design could play a crucial role in whether the medium successfully integrates into the mass market. The Gestalt Principles of Perceptual Organisation offer a psychological explanation of human perception, with particular reference to pattern recognition and how we subconsciously group entities together. There are seven Principles of Perceptual Organisation, nearly all of which are currently widely used in User Interface design, offering designers guidelines on what the size, shape, position and colour the different components of an interface should be. This study presents an analysis on the effects that the employment of the Gestalt Principles has on the usability and mental workloads of Virtual Reality applications.

Keywords: *User Interface, User Experience, Virtual Reality, Gestalt, Usability, Mental Workload*

ACKNOWLEDGEMENTS

I would like first to express my sincerest gratitude to Dr Luca Longo, who went above and beyond the call of duty to help me optimise and organise my experiments, despite not being my assigned thesis supervisor. Dr Longo also taught one of my modules during the second semester of my Master's Program, where he helped me and my classmates to choose and refine our thesis topics and proposals. I could not have completed my experiments without your help Luca, so I cannot thank you enough.

I would also like to thank my supervisor, Dr Deirdre Lillis who was also very helpful in setting up meetings with other researchers studying about Virtual Reality and for always being available whenever I needed help with anything.

Thanks are also due to the lecturers and tutors who taught me everything I needed to know to complete this research project; so to Ciaran Cawley, Pierpaolo Dondio, Brendan Tierney, Damien Gordon, John Gilligan, Brian Keegan, John McAuley, Andrew Hines and Andrea Curley I owe my gratitude.

Developing for Virtual Reality and running an Oculus Rift requires a powerful computer, and my brother Tom was kind enough to allow me to use his gaming PC to develop my application. He spent many evenings kicked out of his own room and even spent a couple of needs sleeping in beds other than his own to allow me to work on the software undisturbed. This project would not have been possible without Tom's selflessness, so to him I say thank you so much.

Without the participant observation experiment, this thesis would have been made completely redundant, so I would like to thank each and every one of the participants who volunteered to take part in the experiments, especially on such short notice. I would also like to extend my gratitude to Compu-Hire Ireland, who saved my thesis when they supplied me with a rental PC for the experiment after the PC I had been working on decided to have a meltdown a matter of days before the experiment was due to begin.

Lastly, to my parents who kept me sheltered and fed and offered me endless support, as well as to my friends who ensured my spirits were never *too* deflated over the past few months, it goes without saying that I will be eternally grateful for all of the support you never cease to offer me.

TABLE OF CONTENTS

TABLE OF TABLES

TABLE OF FIGURES

1 INTRODUCTION

1.1 Background

With the proliferation of Head-Mounted Displays (HMDs) such as the Oculus Rift, Sony PSVR and HTC Vive in 2016, Virtual Reality is an emerging market which is begging to make a splash in the world of computing. Facebook's acquisition of Oculus for a reported $2 billion is an indication of the perceived potential held within this new interactive medium, with experts predicting that the Virtual and Augmented Reality markets will be worth a combined total of $162 billion dollars by 2020[1]. While much of the focus for these devices has been related to sectors of the entertainment industries, namely the video game and cinema industries, there are many more practical applications for these technologies, with potential benefits in educational, training/simulation, therapeutic and modelling/design software (Burdea & Coiffet, 2003). Virtual Reality has existed in various forms since Ivan Sutherland's 1968 *Sword of Damocles* HMD was developed (Sutherland, 1968), but it has only really come to the forefront of mainstream computing in the second decade of the 21st Century. This is largely due to the immense processing power which is needed to render an experience which represents the immersion which VR promises; Oculus state that a frame rate of 75 frames per second is necessary to completely maintain the desired level of immersion (Oculus, 2015).

Due to the rapid growth of Virtual Reality in recent years, there is an increasing need to develop standardised patterns for the design of VR applications. The Gestalt Principles of Perceptual Organisation are a psychological explanation of human perception, with particular reference to pattern recognition and how we subconsciously group entities together. There are seven main Principles of Perceptual Organisation; Proximity, Similarity, Continuity, Closure, Figure/Ground, Symmetry and Common Fate. These Gestalt Principles are currently widely used in User Interface design, offering designers guidelines on what the size, shape, position and colour the different components of an interface should be (Rosson & Carroll, 2002).

---

[1] http://uk.businessinsider.com/virtual-and-augmented-reality-markets-will-reach-162-billion-by-2020-2016-8?r=US&IR=T

1.2 Research Project/problem

Virtual Reality is a relatively new technology in mainstream computing. There is a very real possibility that Virtual Reality and Head-Mounted Displays may make a significant entry into the market, especially with companies like Oculus who are, with the backing of Facebook, pouring millions of dollars into the development of such devices. The ability to develop effective, attractive and efficient User Interfaces will be a huge determining factor in whether the platform will be a success.

Much of the focus for Virtual Reality software has been relating to video games. Video game User Interfaces tend to differ drastically from those of professional software and browser based interfaces, as do the requirements for these UIs. With the relative youth of VR in terms of mainstream computing, design conventions have not yet been established for VR applications. This paper aims to discern if the Gestalt Principles of Perceptual Observation are viable as guidelines for developing applications for VR.

1.3 Research Objectives

The primary objective of this research paper is to evaluate the effectiveness of implementing the Gestalt Principles of Perceptual Observation when designing applications for Virtual Reality. This effectiveness will be measured in terms of two metrics; usability and mental workload. By testing a Virtual Reality application which has two User Interfaces – one heavily influenced by the Gestalt Principles and one not – and comparing the results of the usability and mental workload metrics, it should be possible to determine how effective the Gestalt Principles are for VR application design. As the first paper to directly investigate the relationships between the Gestalt Principles and usability/mental workload in the context of Virtual Reality, this paper will contribute to the body of knowledge in this field by testing the effectiveness of the principles in a scientific manner.

1.4 Research Methodologies

Descriptive statistics will be generated through the project experiment which involves user-based evaluation. A quantitative analysis of these statistics will provide the foundation of this project's research. Mean and Standard Deviation calculations as well as t-Tests and Pearson's Correlation Coefficients will form the basis of the statistical methods implemented to achieve the research objectives.

1.5 Scope and Limitations

The scope of this project is to examine the effectiveness of the Gestalt Principles in terms of their impact on usability and mental workload. Usability and UX/UI design are completely intertwined, with progress made in either field leading to advances in the other. Usability and mental workload are also intrinsically linked. The strong relationships between these two metrics, as well as the well-established methods to test them were at the heart of the decision to focus on these aspects of software design.

There were no limitations for the participant observation in terms of the sociodemographic distribution of the volunteers.

1.6 Document Outline

This body of this paper is comprised of four main chapters, as well as this introductory chapter and a conclusion chapter at either end.

Chapter Two - Literature Review will outline the bodies of knowledge relevant to this research project. Conclusions will be drawn based on the work of other researchers in the fields of the Gestalt Principles of Perceptual Observation, Virtual Reality, Mental Workload and Software Usability.

Chapter Three - Design & Methodology will discuss the design of the experiment, in particular the thought processes behind each aspect of the application design process and how the Gestalt Principles were applied for one User Interface and omitted in the other. The methodologies and formulas used to process the results will also be outlined in this section.

Chapter Four - Implementation & Results will analyse the actual implementation of the project's experiment and provide details on the results which the experiment generated.

Chapter Five - Evaluation & Analysis will analyse the results data which was presented in Chapter Four and attempt to accompany the results with plausible explanations which relate to the theories discussed in Chapter Two.

2 LITERATURE REVIEW

2.1 Introduction

This chapter will examine the existing literature for four topics which are at the heart of this research project. The effectiveness of the Gestalt Principles of Perceptual Observation is under scrutiny in this paper, so naturally the ideals of Gestalt Psychology will be discussed. What distinguishes this paper from the many other papers testing the Gestalt Principles as a UX/UI design tool is that this paper focuses on their effect for Virtual Reality applications. For this reason, Virtual Reality and the design principles which drive it will also be researched. Due to the relatively young age of VR, there is not a huge range of validated and distinguished papers written on the platform, and as of the time of writing, there is yet to be a field defining paper written on the subject. Thankfully that is not the case with either Mental Workload or Software Usability, which are the last two topics which will be examined.

2.2 The Gestalt Principles of Perceptual Observation

There are seven Gestalt Principles of Perceptual Observation which are commonly implemented in visual design; Proximity, Similarity, Symmetry, Continuity, Closure, Common Movement and Figure/Ground. Literature on each of these seven principles and their relationships with User Experience and User Interface (UX/UI) design will be briefly discussed, as will the use of Gestalt Psychology as a whole with regards to the field of Human-Computer Interaction (HCI).

The Gestalt Principles offer an explanation to how humans perceive each component in any given environment. This can be applied to audial, tactile and visual environments, although this paper will focus solely on Gestalt Psychology in relation to the visual environment. Having the ability to perceive the structure of the visual environment enables one to better comprehend the elements contained within that environment. This leads to us being able to react to any events which take place within that same environment quicker and more efficiently (Koffka, 1935a). Allowing users to more easily get their bearings with a system tends to produce more positive performances in terms of both statistics and perception. (Shneiderman, 2000). By applying the Gestalt Principles to a user interface, a designer can improve the overall clarity of their software solution. This is due to the fact that "visual interfaces often rely heavily on association between graphical elements, such as the placement of a label next to a checkbox, or the grouping of items in a list" (Yee, 2002). The Gestalt Principles have been widely adopted for improving software user interfaces as they allow users to subconsciously take control of any software environment because of the human species' innate ability to cluster the elements and add context to each element without need for explicit instructions or explanation (Lidwell, Holden & Butler, 2010).

2.2.1 Principle of Proximity

"*Proximity is one of the most powerful means of indicating relatedness in a design, and will generally overwhelm competing visual cues*". (Lidwell et al., 2010)

Arranging the elements in an interface so that elements which are related are located close to one another causes the user to assume that these components are related to one another. Different layouts of the elements in a user interface can portray a plethora of different kinds of relationships between these elements. As an example, elements which overlap or are in some way connected are generally perceived to share one or more common attributes, while elements which are not connected but are positioned close to each other are usually interpreted as being independent of each other while still being related. This is demonstrated in Figure 2.1 (Lidwell, Holden & Butler, 2010).

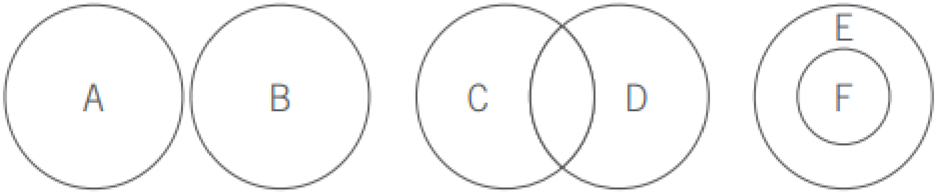


Figure 2.1The Gestalt Principle of Proximity (Lidwell, Holden & Butler, 2010)

Han et al. argue that proximity is the most useful of the Gestalt Principles for allowing users to distinguish which components of the system are the most important and/or useful, as well as making the organisation of the components easier, allowing for better overall usability. (Han, Humphreys & Chen, 1999)

2.2.2 Principle of Similarity

"*It [similarity] asserts that similar elements are perceived as a single group or chunk, and are interpreted as being more related than dissimilar elements*". (Fisher & Smith-Gratto, 1999).

Elements which are similar in their physical appearance, be that through size, shape, colour or texture, are most likely to be perceived as being members of the collective group. Efficient employment of the Principle of Similarity can be used to highlight any element or group of elements (Fisher & Smith-Gratto, 1999). Similarity makes interfaces simpler, easier to navigate and more aesthetically pleasing while also emphasising the relationships (or lack thereof) between a collection of elements. Size and colour are the best tools for portraying connectedness, with shape and texture being more useful for illustrating differences (Lidwell, Holden & Butler, 2010).

The Principles of Proximity and Similarity are intrinsically linked to one another. Chang et al. state their reasoning for selecting these two principles as the primary focus of their study by saying "the principles of similarity and proximity have been selected for the experiment because they are two important and commonly-used Gestalt principles that have been identified as higher level principles for the design of multi-sensory displays" (Chang, Nesbitt & Wilkins, 2007). In combination, the use of the Principles of Proximity and Similarity is the most widely adopted implementation of Gestalt Psychology in UX/UI design (Weiser, 2006).

2.2.3 Principle of Symmetry

"*Its [symmetry's] great value to design lies in its power of transition or movement from one form to another in the system. It produces the only perfect modulating process in any of the arts*" (Hambridge, 2012).

Elements which are placed or shaped symmetrically tend to be easier to remember - in terms of both recognition and recall - due to the fact that symmetry lends itself to simpler design and the fact that humans tend to be better at remembering designs which are considered beautiful rather than designs perceived as being ugly or uninteresting (Lidwell, Holden & Butler, 2010). If an element of an interface (or even the interface as a whole) does not have some aspect of symmetry and/or balance in its design - be it through reflection, rotation or translation - there is a tendency for it to appear to the end user that an aspect of the interface is missing. This can lead to confusion or disillusionment for the user, thereby negatively affecting the usability of the system (Chang, Nesbitt & Wilkins, 2007).

2.2.4 Principle of Closure

"*The principle of closure enables designers to reduce complexity by reducing the number of elements needed to organize and communicate information. ... [closure] not only reduces its complexity, but it makes it more interesting to look at — viewers subconsciously participate in the completion of its design*" (Lidwell et al., 2010).

The Principle of Closure explains why instead of seeing a random collection of individual elements strewn across the screen, we can relate each of these elements as being part of a pattern or interconnected system (Lidwell, Holden & Butler, 2010). A good example of this can be seen in the logo of IBM, where instead of seeing a collection of independent white lines on a blue background, we instead see the letters 'I', 'B' and 'M', even though the letters are not fully drawn, as can be seen in Figure 2.2 (Chang, Dooley & Tuovinen, 2002).

Figure 2.2 The IBM logo - a good example of closure in design

Using closure allows UI designers to decide on the appropriate locations for every component when used in conjunction with other Gestalt Principles, most notably the Principles of Proximity, Symmetry, Continuity and Figure/Ground. While the implementation of the Principle of Closure does not necessarily result in greater usability, examples where Closure is omitted tend to cause problems with the integration of visuals and text (Moore, 1993).

2.2.5 Principle of Continuity

“Elements arranged in a straight line or a smooth curve are perceived as a group, and are interpreted as being more related than elements not on the line or curve” (Lidwell et al., 2010).

When sections of a perceived linear pattern are covered or obscured, our brains tend to mentally continue along the path the line would have taken and we expect to see the line re-emerge and an approximate location. For this reason, when developing a user interface, linear continuity should be at the forefront of the designer’s thoughts when placing the components of the interface (Lidwell, Holden & Butler, 2010).

2.2.6 Principle of Common Fate

The Principle of Common Fate explains how we perceive moving elements of an interface. Elements that move in the same direction are perceived to be related to one another and to be independent of elements which are either stationary or moving in a different direction (Lidwell, Holden & Butler, 2010). The Principle of Common Fate has traditionally not been widely implemented in traditional User Interface Design due to the fact that the majority of UIs since the introduction of the Graphical User

Interface (GUI) have been static, with little or no moving parts. However, with recent developments in UX/UI and the emergence of platforms such as AngularJS and other dynamic UI development tools, the Principle of Common Movement has begun to become more important in modern UX/UI design (Johnson, 2013).

2.2.7 Principle of Figure/Ground (Prägnanz)

"Changes in the figure/ground relationship are important and provide valuable feedback to the user by visually informing the user that action will occur when the link is clicked." (Graham, 2008).

The Principle of Figure/Ground demonstrates how when multiple interpretations of a design are available, each individual will first perceive the interpretation which is simplest or most familiar to them (Lidwell, Holden & Butler, 2010). Users can interpret changes in the Figure/Ground of an environment, making this a powerful tool for user feedback. A common example of this in software and web development is the 'hovered' functionality, whereby the cursor image changes or some other visual indication is given to the user that they are hovering over any given component (Graham, 2008). Norman lists this type feedback as one of the six principles of design (Norman, 1983).

2.3 Virtual Reality

One of the biggest problems facing developers who are looking to move to Virtual Reality development is that User Interfaces for Virtual Reality and Head Mounted display software do not yet have any design conventions. Whether they be adaptations of more established patterns or the development of newer design patterns, finding and cementing design principles is likely to be a decisive factor as to whether the Virtual Reality platform manages to integrate into the wider market. (Alger, 2015).

Many of the newer design philosophies surrounding Virtual Reality revolve around the idea of immersion. Bowman and McMahan argue that high levels of immersion and realism are the keys to the successful implementation of Virtual Reality, citing the success of VR applications in the fields of phobia therapy and military training. The paper argued that their experiments suggest that VR applications with highly immersive UIs result in more efficient task performance and a greater level of user understanding (Bowman & McMahan, 2007). By contrast, Morie and Williams make the argument that too much emphasis is being places on photo and audio-realism to enhance immersion, with not enough focus on how these inputs are perceived and processed by the user and how they affect the overall user experience (Morie & Williams, 2003). Alger argues that the most important factor in building a successful UI for VR is to correctly “zone” each area of the interface (i.e.) ensure that each component is at the appropriate depth and in the appropriate region of the screen (Alger, 2015).

Testing and evaluation methods for Virtual Reality User Interfaces will also need to be established for the VR platform to grow successfully. Sutcliffe and Gault expanded upon Nielsen’s Heuristics to create a model to evaluate the usability of Virtual Reality interfaces. Twelve heuristics are outlined in the paper, two of which are solely applicable to VR applications, with the other ten being closely linked to the existing guidelines outlined in Nielsen’s Heuristics (Sutcliffe & Gault, 2004). Sawyerr, Brown, and Hobbs have outlined a hybrid method of evaluation for three dimensional interfaces which is used to effectively highlight usability issues with VR applications, specifically those which fully utilise the full 360° displays and the depth of field that VR can provide (Sawyerr, Brown, & Hobbs, 2013).

While Virtual Reality applications may cause more frustration than a traditional desktop application, the differences between the mental workloads of both platforms was insignificant in the case of their experiments. The increased frustration is likely caused by the lack of experience of the average user when it comes to using a Virtual Reality application as opposed to carrying out the same task on a more traditional platform (Stefanidis et al., 2006).

In terms of design patterns or principles, there are varying opinions on how to approach VR design. Sherman argues that a standard design approach is the best possible design philosophy when designing a VR app. He does state however, that this is based on the idea that it is more beneficial to have an idea which you believe VR can provide a solution to, rather than starting with the idea of building an app in Virtual Reality and searching for a possible application of the technology. If a suitable problem is identified, Sherman encourages an Agile Development cycle, stating

> "*Once a decision is made to explore virtual reality as a means to attain a goal, standard design techniques should be followed. In particular, involve users. Get their ideas up front and get feedback from them at each stage of the development. Iterate over the design. Continually refine the experience by implementing, testing, and analysing the work as it progresses and be willing to throw away bad ideas, no matter how much effort was put into them*" (Sherman, 2002).

Alger argues that design principles from other mediums will need to be adopted in order to fully unlock the potential which VR holds, citing areas such as web design, architecture, interior design, theatre, motion graphics and print design as areas which can provide better design principles than the traditional software development methods (Alger, 2015).

2.4 Mental Workload

Mental workload is a multidimensional concept borrowed from psychology which is difficult to define uniquely (Longo, 2014) (Longo, 2015a) (Rizzo, Dondio, Delany, & Longo, 2016). "Mental workload is described in terms of time, distance to desired goal, and effort". Mental workload has a profound effect on the stress induced by the software and thus also affects the performances of its users as much as the actual usability of the system. The mental workload required can be a determinate factor in the overall performances of complex tasks (Hancock, 1988). Having the ability to estimate the mental workload imposed by an interface allows designers to add or remove complexity as needed to optimise the demands of the system. By being able to identify the perceived mental workload of any given interface at an early stage in development, both time and money can be saved in the development process (Wu & Liu, 2009). Application of the concept of mental workload are various. For instance, (Longo, 2011) proposed to adopt mental workload to contribute to the assessment of cognitive engagement in the World Wide Web, of for adaptive and personalised web-design (Longo, 2012). The same author also investigated its relation with the concept of usability (Longo, Rusconi, Noce, & Barrett, 2012) (Longo & Dondio, 2015) and applied it in the context of health-care and medical systems (Longo, 2015b) (Longo, 2015c), (Longo, 2016), (Longo, Dondio, & Barrett, 2010).

An interface which is perceived as being too simple, creating a mental underload, can be as detrimental to a system as the mental overload caused by an overly complex interface. Creating an interface which is challenging without being overbearing or which is simple without being uninteresting or tedious is an ideal mental workload design goal (Xie & Salvendy, 2000). The levels of mental workload which fall into the categories of underload, overload or optimal load vary from person to person. This is based largely on what Jex refers to as our "meta-controller devices". By this, he is referring to each individual's ability to adapt to the difficulty of the task, choose an appropriate strategy to handle a given problem, interpret the environment of the system and other such cognitive tools (Jex, 1988). To gauge the perceived mental workload of a system, testing of large and diverse groups needs to be set in place. The most widely implemented tool for testing the mental workload of a system is the NASA Task Load Index (NASA-TLX), a multidimensional subjective rating system developed at the National Aeronautics and Space Administration (NASA) over a number of years in the mid-1980s. The NASA-TLX involves a questionnaire which contains six main dimensions; Mental Demand, Physical Demand, Temporal Demand, Overall Performance, Frustration and Effort (Moroney, 1992). The paper released by NASA to demonstrate the effectiveness of the Task Load Index model provides data which indicates that using this standardised format can contribute greatly towards developing a system with an optimal workload. The data provided is based on the metrics of task performance, a weighted workload (WWL) score and "experimental variables reported elsewhere" (Hart & Staveland, 1988). Xiao et al. reported that, from their experiments at least, the NASA-TLX proved to be highly retestable, with high correlations and Cronbach's Alphas between participants who were tested using the same unchanged system on multiple occasions (Xiao, Wang, Wang & Lan, 2005). In their study, which compared NASA-TLX to two other mental workload assessment tools - Subjective Workload Assessment Technique (SWAT) and Workload Profile - Rubio et al found that NASA-TLX is the optimal choice for predicting user performance (Rubio, Díaz, Martín & Puente, 2004).

In recent years, there has been a strong movement for an adaptation of the NASA-TLX model known as Raw Task Load Index (RTLX). The RTLX provides a similar framework to the NASA-TLX, but no weighting is placed on any of the dimensions. Cao et al. report that there is a high correlation between test groups who are tested using NASA-TLX and RTLX. RTLX allows the researcher to set questions more specific to the system being tested (Cao, Chintamani, Pandya & Ellis, 2009). Bustamante argues that this shorter, condensed version of NASA-TLX improves the validity of the questionnaires as compared to the full version (Bustamante, 2008). Hart states that across 29 different implementations of RTLX that there were examples of it outperforming, being outperformed by and performing equally as well as the full version of NASA-TLX (Hart, 2009).

2.5 Software Usability

According to the International Organization for Standardization (ISO), usability can be defined as "*The extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use*".

Nielsen expands on this definition, stating that usability is an umbrella term which can be used to describe many different aspects of an interface including the efficiency at which tasks can be completed, how easy the system is to learn, how memorable the system is and how errors are handled (Nielsen, 1992). Abran et al. state that usability is "no longer a luxury", but is now a necessity in a market which is constantly evolving and becoming more competitive (Abran, Khelifi & Suryn, 2003).

There are three main categories for doftware usability evaluation; user-based evaluation, expert-based evaluation and model-based evaluation. User-based evaluation involves target users testing and evaluating the system. Expert-based involves hiring a UX/UI expert to assess and evaluate the application's usability.

Model-based evaluation implements a number of criteria which the system must meet in order to predict the performances of the end users (Dillon, 2001). The experiment outlined in this paper will be implementing a user-based evaluation. There are numerous tools available to developers for user-based evaluation of software usability. The Post-Study System Usability Questionnaire (PSSUQ), the User Experience Questionnaire (UEQ) and the Software Usability Measurement Inventory (SUMI) are all examples of popular usability evaluation tools. However, the Software Usability Scale (SUS) has established itself as the most popular method to establish a system's usability based on user feedback. In particular, it has emerged as a suitable evaluation tool for smaller sample sizes, which makes it preferable to the PSSUQ for this research project. Bangor et al. found that the SUS had a Cronbach's Alpha coefficient of .91; Cronbach's Alpha being a method to determine the reliability and consistency of a data set (Bangor, Kortum and Miller, 2008).

2.6 Conclusion

The Gestalt Principles have been utilised to great effect across a plethora of mediums for much of the last century and have played an important role in the formation of many of the UX/UI design patterns commonly implemented today. In particular, the Principles of Proximity, Similarity, Symmetry and Figure/Ground have led to the formation of many design conventions many designers take for granted today. The implementation of these principles tends to have a profound effect on the usability of any interface. Many of the lessons taught by Gestalt Psychology should be transferable to the Virtual Reality platform since they are general design principles, rather than patterns which are specific to the design of two-dimensional user interfaces. This is important because as of yet, no Virtual Reality design conventions have been established, with some authors believing that borrowing from other areas of design will be crucial to the survival of VR and to unlocking the true potential that it holds.

Most who have studied VR will agree that immersion is another factor which will also play an important part in the success or failure of the platform. The ability of researchers and designers to figure out what levels of immersion help to develop highly usable VR applications will undoubtedly be fundamental to whether commercial, professional and industrial uses of Virtual Reality come to fruition.

With any emerging technology, new users will need time to adapt to the novel challenges and nuances of a new system, so maintaining a manageable mental workload will likely influence the initial impressions of those using VR for the first time.

3 DESIGN AND METHODOLOGY

3.1 Introduction

This chapter will outline the design of the experiments carried out as part of this research project. The experiment was devised with the intention of discerning the value of the Gestalt Principles of Perceptual Observation when designing User Interfaces for Virtual Reality software. This was to be achieved through the development of a Virtual Reality application with two different User Interfaces; one of which is intended to measure the effectiveness of the Gestalt Principles by strongly exhibiting features of Gestalt Psychology in its design, the other which acts as a control with little or no thought of the Gestalt principles when designed. This would then be followed by having participants test and evaluate the usability and perceived mental workloads of the application with both UIs. The participants would fill out a pre-questionnaire before attempting each task assigned to them in order to gain an insight into their mental state prior to undertaking each task. Upon completing each task, the user will then complete a post-questionnaire comprised of two parts; a mental workload questionnaire based on the Raw Task Load Index (RTLX) model and the System Usability Scale (SUS) questionnaire for measuring the system's usability. These questionnaires are designed to give subjective information to process.

During the act of completing these tasks, the application will be recording data such as length of time taken to complete each task and the number of mistakes made by the user. This data will provide objective information, which can be used in conjunction with the data from the mental workload and SUS results.

By triangulating the subjective and objective feedback data, it should be possible to provide an insight into the effectiveness of the Gestalt Principles for a basic VR application. It was intended that the application be designed in a way that no prior knowledge or experience with Virtual Reality was necessary for the participants. It was also intended that the participants be from diverse backgrounds in terms of age, gender, nationality and experience/comfort with computers and high technologies.

In this chapter, detailed accounts of the software design and development processes will be outlined, as well as the thinking behind the quantitative methods of evaluation which were used and the approach to handling the results of the experiments. Also to be discussed in this chapter will be the technologies (in terms of both hardware and software) which were used for the development and evaluation of the experiments.

3.2 Hardware and Software Implemented

3.2.1 Hardware Implemented

The application was built to run on an Oculus Rift Head Mounted Display (HMD) as the Oculus Rift is the most established Virtual Reality platform on the market currently and has become the go-to platform for VR development since the release of the Oculus Rift DK1 in March 2013, even more so since the release of the DK2 in July 2014. While the Oculus Rift is less powerful than its major competitor the HTC Vive, the Oculus Rift has better developer support than the HTC Vive (Alger, 2015).

For the experiment, the participants were required to wear an Oculus Rift HMD. An XBOX One Gamepad was to be used as the user’s input device. The XBOX One Gamepad was chosen as the input device because of its compatibility with both Windows and the Oculus platform, its easy integration through the engine in which the software was built, its portability/mobility compared to a keyboard or mouse and the lack of availability of any motion-control or haptic I/O devices for this research project.

3.2.2 Software Implemented

The application itself was built using a game engine developed by Epic Games called Unreal Engine 4 (UE4). UE4 is the second most popular open-source game engine on the market behind Unity Technology’s Unity3D 5 engine when it comes to developing desktop applications. While both engines have integrated support for Virtual Reality application development, UE4 has emerged as the leading development tool for creating VR games due to a more intuitive design and through its blueprints system (Shah, 2014).

The application was developed using a combination of UE4's baked-in tools for VR development, UE4's Blueprints functionality for programming and scripting, as well as personally written C++ code. Unreal Engine 4 implements C++ version 11, so this is the C++ version in which all the code for this project is written.

For the results processing, a combination of the Python Pandas data analysis library, RStudio 3.3.2 and Google Spreadsheets was used.

3.3 Application Design

3.3.1 Application Design - Tasks

Two tasks were to be developed for the participants to carry out during the experiment. The tasks were designed to differ in the levels of complexity and the mental workload required to accomplish each assignment. One of these tasks would be a very simple rudimentary task which would not be mentally taxing for the user. The second task would require a greater level of concentration and would necessitate the participant having to use their short-term memory to complete the task correctly and efficiently. The reasoning for this is to test the effectiveness of the Gestalt Principles for both the ease of use for mundane tasks which the user can complete on autopilot as well as measuring the ease of use for tasks which require split concentration when the Gestalt Principles are applied. This follows the recommendations outlined by Nielsen who states that it would be beneficial for testers to complete both "typical tasks" and tasks which involve "cumbersome steps, steps that would not be natural for users to try" (Nielsen, 1994).

That being said, it is important to ensure the balance of the more mentally taxing task's level of difficulty. While it is important to create a task which demands a relatively high mental workload, it is equally as important to avoid creating a task which is overly difficult as doing wo may lead to a possibility of the results being warped by the difficulty of the task at hand, relying more on the abilities of the participants and less on the effectiveness of the user interface's design (Chen & Vertegaal, 2004).

Both tasks designed for the experiment were based around the manipulation of the cells in a grid which the user is greeted with as soon as the experiment begins. The grid contained nine cells, in a three by three pattern, as is outlined in the prototype diagram displayed in Figure 3.1.

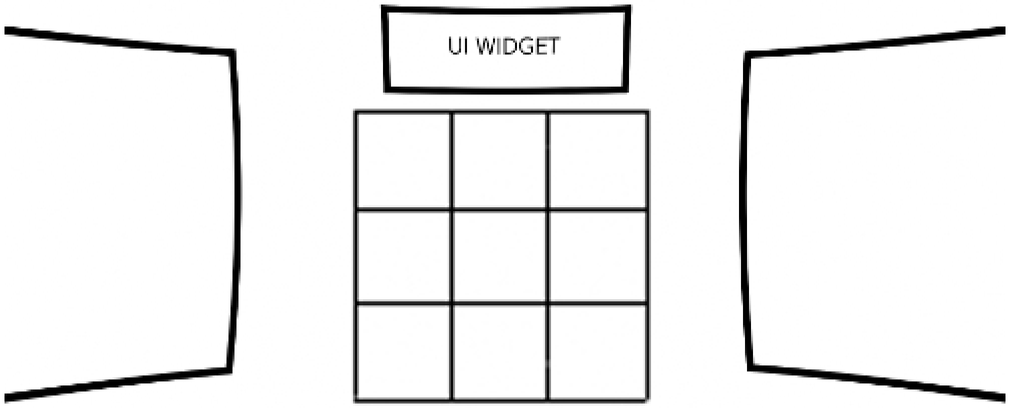


Figure .3 A low fidelity prototype for the front facing view of the software, with the grid in the centre and the functional panels on either side.

As can be seen in this low fidelity prototype, there are four main elements on the screen when the user starts the application; the grid in the centre, panels on either side of the grid, and a User Interface Widget above the grid. In the panel on the left are a selection of tools for the user to implement; a directory containing a selection of patterns for the user to apply to the grid, a fill tool which fills one of the cells to a different colour, as well as undo and redo buttons. The panel on the right-hand side contains a selection of colours for the user to choose from. Above the grid is a User Interface Widget which displays to the user which tool and colour is currently selected. The arrangement, positioning and layout of the components on each of these elements of the UI will be at the heart of the difference between the two different interfaces, with implementation of the Gestalt Principles being the driving force behind the actualisations of those properties in one interface, and a lack thereof being the primary design goal for the other.

Both tasks will be started remotely by the observing party once the participant has informed them that they are ready to begin. As soon as the task has started and all the components have rendered, the software begins tracking an array of objective information which will be processed for the experiment's results. The information being tracked is as follows:

- **The time taken to complete the task -** a task being completed faster may indicate a greater ease of use. Participants will be told to complete the task at their own leisurely pace. As the participants will be observed during the experiment, the speed at which the participants attempt to complete the task will be determined by the observer and a rating of fast, regular or slow will be assigned to the user to weight the results of the time taken appropriately.
- **The total number of clicks -** a lower number of clicks needed to complete a task may indicate a better level of usability as it would indicate that the user has less required of them in order to complete a given task. Results indicating a higher number of clicks needed to complete a task may indicate inefficient interface design. As both interfaces will share the exact same functionality, the minimum number of clicks necessary to complete the task will be the same for both UIs.
- **The total number of errors** - for the experiment, errors are defined as any time the user makes a mistake related to completing the task (e.g.) selecting the wrong tool or incorrectly manipulating one of the grid cells.
- **The total number of misclicks** - for this system, a misclick will be defined as any time a user clicks on an area of the screen which has no functionality. The number of misclicks will be tallied and a proportion of the misclicks will be removed from the total number of clicks metric, as these clicks will not fairly represent the data which the total number of clicks in attempting to compile. The same proportion of misclicks which is deducted from the total number of clicks will be added to the total number of errors metric.
- **Field of View (FOV) utilised by the participant-** this metric is specific to the Virtual Reality platform, as the 360° design space allows for different opportunities when designing an interface for a VR HMD. The FOV covered by the two interfaces differ slightly. Results which show that the participant utilised a FOV similar to the total FOV which the interface covers would

> indicate an efficient use of the three-dimensional space, whereas if a participant utilises a FOV which is greatly higher or lower than that of the interface, this may indicate that the interface was implemented inefficiently considering the platform.

Upon completion of the given task, calculations to derive the total number of mistakes are performed based on the total number of errors and misclicks. Also calculated in the percentage differential between the FOV utilised by the user and the FOV implemented by the interface. The values stored for the time taken, the number of clicks, the number of mistakes and a the FOV differential are then output.

For the first, easier task, the user is given an overlay on top of the grid, telling them which colour to fill each cell with. The user will be told to change the colours of each tile in the grid, but will not be instructed on how to do so. The colour pattern the user is instructed to colour each panel is the same for all participants.

To complete this task, the user must select the Fill tool which is indicated by the paint bucket icon which has become synonymous with a function such as this from other Graphical/Image Manipulation software such as Microsoft Paint or Adobe Photoshop. After having selected the Fill tool, the user must then select the appropriate colour from the colours panel and apply it to the relevant cell. Once the system recognises that each cell has been correctly filled with the colour corresponding to the overlay, it finishes recording the data previously outlined and writes the result to a log file, along with a participant ID integer and binary values indicating the task which was undertaken and on which interface it was completed.

For the second task, the user was greeted with a similar scene to that of the first task. As with the simple assignment, when the application is launched the user will be greeted with a grid with an overlay telling them which colours to fill each cell with. On this occasion, some of the colours listed will not be available to the user from the colours panel. Participants will instead have to navigate through the directory to find an image which corresponds to the overlay and apply this image to the grid. This will require the user to remember the pattern they have been asked to recreate while trying to navigate through a file system. With the added mental workload of having to retain this information, this task should be a decent indication of how usable the two directory UIs are for complex, mentally taxing goal-oriented exercises. The directory UIs for both interfaces will differ significantly based on their levels of adherence to the principles of Gestaltist Psychology.

3.2.3 Application Design - Interfaces

There are three main components to the User Interface for the application. Firstly, there are the buttons for selecting the tools and colours. The second main component is the directory system implemented in the second task. The final facet of the UI is what is known as a Heads-Up Display (HUD), which displays information to the user without them having to make an effort to look for this information (i.e.) the user should not have to move their FOV to find this information, it should be visible to them at all times. Each of these three components would require different designs in order to amplify the effects of the presence or absence of the Gestalt Principles of Perceptual Observation. This section will outline the design philosophies for both interfaces across the three UI components.

For the purpose of distinction between the two User Interfaces implemented in this experiment, from hence forth the interface which exhibits the features of the Gestalt Principles will be referred to as Interface A, whereas the interface which lacks the Gestalt Principles in its design will be referred to as Interface B. Likewise, the simpler task will from this point be referred to as Task A, with the more complex task being referred to as Task B.

3.2.3.1 Scene and Component Layout

With regards to the overall layout of the scene and each of its components, the Gestalt Principles which were followed most closely for Interface A were the principles of Proximity, Similarity and Symmetry. By contrast, these principles were either entirely ignored or purposely contradicted when designs were being created for Interface B.

The Principle of Proximity states that we perceive objects which are positioned closely to one another as being somehow related to one another (Koffka, 1935a). For this reason, the tools and colours panels were placed at opposite sides of the scene, with the user needing to look left to view the tools panel, whereas to view the colour selection panel they would have to look to their right. The purpose of this was to have the user perceive these two panels of being independent of one another. A selection of a button on either panel does not directly affect the selection of a button on the other panel. By contrast, the buttons contained within each panel do immediately affect the other buttons on that panel, so they were position next to each other to convey to the user that these buttons are all related.

For Interface B, the placement of each component within the scene (with the exception of the grid which remained in its central location) was entirely randomised. With such disorganisation amongst these components, the positions of the buttons on this interface represent the antithesis of the Gestalt Principle of Proximity.

The Principle of Similarity states that we perceive objects which are similar in their physical appearances to be related (Koffka, 1935a). Adherence to the Principle of Similarity drove the decision to make each button the same size as the other buttons on its own panel, but different to the buttons on the other panel. The reasoning for this is the same as was described when discussing the Principle of Proximity. The panels themselves are identical in both their appearance (except for their contents) and their dimensions. The intention of this is to inform the user that they serve the same purpose.

As was the case with the Principle of Proximity, the design behind the shape and appearance of each element of the interface was designed to go against all that the Principle of Proximity teaches. Each of the components were warped from their original shape, creating an array of components of different shapes, dimensions, rotations and scales. In doing so, this should not give the user any perception of the components being related to each other.

The Principle of Symmetry states that when we perceive objects, we perceive them around a central point, also sometimes referred to as mirror line (Koffka, 1935a).

Objects which are equidistant from this point, or mirror one another across this point may be perceived as being related. In the case of this application, there are a few different instances of central points around which objects may be perceived. First and foremost, the position of the camera within the scene itself acts as a central point, with all scene components surrounding the user. This means that the components themselves should be located symmetrically along each of the three axes in relation to the camera. For example, supposing the camera is positioned at the origin (0,0,0), a component at a position of 50,100,0 could have a corresponding component at -50,100,0, mirroring the first component along the X-axis. One of the biggest difference between developing interfaces for traditional platforms and developing for Virtual Reality is the introduction of a Z-axis, which introduced the semblance of depth to the application's interface. Another example of a centre point is at the centre of each panel, with the buttons for the tools and colours being positioned based upon the dimensions, rotations and location of each panel and the positions of the other buttons relative to the centre as determined by these parameters.

With Interface A, major every component is symmetrically aligned with another component it shares some characteristics with. The camera is located at the origin within the scene. The tools panel has a location of -270,100,35 and a rotation of 335°. The colours panel mirrors the tools panel along the X-axis, with a location of 270,100,35 and mirrors the tools panel's rotation with a rotation of 25°. Likewise, as was previously mentioned, the two panels are identical in terms of their dimensions. This results in each panel being equidistant from the camera and having a perfect symmetry from the user's perspective.

As was mentioned earlier, the locations and rotations of the components were completely randomised for Interface B. This obviously had a major negative impact on the symmetry of the elements within the scene itself. Whereas in Interface A each element is perfectly symmetrical with another element or is positioned in a way so that it is mirrored along the same axis as the camera, this is not the case whatsoever with Interface B.

3.2.3.2 Directory Systems

The directory system is implemented to allow the users to pick a pattern to colour the grid based on a series of clues given to them. By using a more user-friendly, Gestaltist interface, it is hoped that the user will be able to focus more on the task at hand and following the clues properly, rather than being distracted by a clunky or unintuitive UI. Yet again, the employment/negligence of the principles of Proximity, Similarity and Symmetry were central to the designs of this facet of the two interfaces. The directory was implemented as a 2-Dimensional panel in both interfaces. This is largely down to a lack of time to properly design and implement a 3D directory interface. Both directory UIs feature five buttons; three face buttons for selecting clues, a back button and an apply button. There are three layers to the directory, with any clue leading to three more clues until the grid is correctly coloured in, at which point they can press an ‘Apply’ button, which will end the task.

For Interface A, each of the three buttons were equal in dimensions, shape and were positioned symmetrically relative to the centre of the directory panel. Each of the three buttons were also all coloured the same. This meant that all three of the buttons were in line with the principles of Proximity, Similarity and Symmetry. Likewise, the Back and Apply buttons were the same in terms of dimensions, shape and were positioned at opposite ends of the panel, equidistant from the centre. The Apply and Back buttons were coloured differently to the three main face buttons to distinguish them from the buttons which had a different functionality. The two buttons were also different colours from one another; the Back button was coloured red whereas the Apply button was coloured blue. This was designed to further enforce the differences between these two buttons.

For Interface B, the buttons were all different in terms of dimensions. No two of the face buttons were the same size, whereas one of the face buttons was the same size and shape as the Back button. The buttons were also scattered across the panel in a non-symmetrical fashion.

3.2.3.3 Heads Up Displays (HUDs)

Displayed on the HUD were two snippets of information for the user to consume; the currently selected tool and the currently selected colour. Whenever the user selects a tool or colour, the HUD is automatically updated. The HUD was attached to the camera, so the user could see this information always, regardless of which part of the screen they were looking at. The HUD could be opened or closed by pressing a button on the XBOX One Gamepad which was to be used as the primary input device.

For Interface A, the HUD was positioned along the top of the user's FOV, in a central area at the top. The HUD was attached to a spring arm component, which meant that it followed the camera, so the HUD was always in the same position as far as the participant's FOV was concerned (i.e.) if they panned the camera to the left, the HUD would still be in the same area of the visible screen. This complies with two of the Gestalt Principles that have so far not yet been discussed; the Principle of Common Movement and the Principle of Figure/Ground.

Because the participant can always see the HUD no matter where they are looking, and due to the fact that it is the only actor which they perceive to be static, they immediately perceive that it is a different component to all other components. This comes from the Principle of Common Movement, which states that we comprehend objects perceived to be moving together to be related to one another.

The Gestalt Principle of Figure/Ground would suggest that as the HUD is perceived as being in a fixed position which is always visible, the user tends to focus on the other screen components.

All components of the HUD were sized identically and all the text was the same colour. They are also spaced evenly and positioned in a symmetrical manner. In this way, the design of the HUD for Interface A also complies with the Principles of Proximity, Similarity and Symmetry.

When it came to designing the HUD for Interface B, none of the principles outlined above were implemented. While the HUD remained in the user's FOV at all times, it's position on the screen was not fixed. If the user moved the camera to the right, the

HUD would remain in the same area of the screen it was at the start of the task, but once it reached left edge it would stay there (and vice versa for when they moved their head to the left). By making it appear that the HUD is moving in tandem with the other on-screen components, Gestalt Psychology would suggest that the user dismisses it as being part of the scene and not part of the UI. In this way, the Principles of Common Movement and Figure/Ground are not followed.

The Principles of Proximity, Similarity and Symmetry are also ignored in the HUD design for Interface B, with each HUD component differing in size and colour, and being positioned randomly inside the HUD widget component.

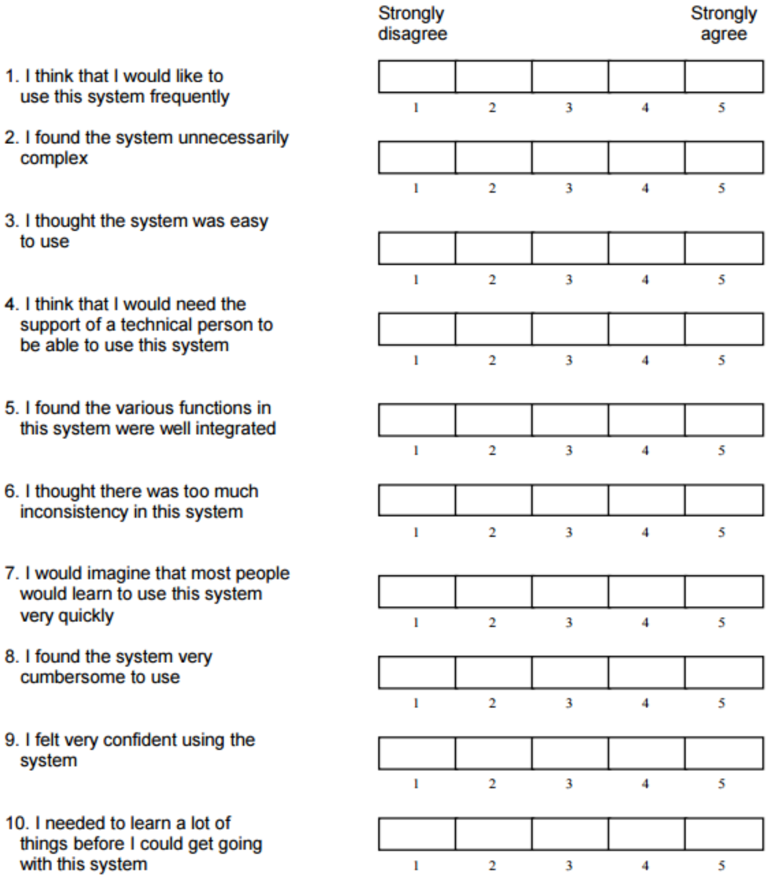
3.3 Result Processing

The results processing can be broken down into three distinct sections; objective results processing, subjective results processing and data triangulation. The first two of these sections involve a quantitative analysis of the data sets provided by the participants during the participant observations and the pre-test and post-test questionnaires. These two data sets will then be triangulated to provide an overall metric by which to judge the usability and required mental workloads of the two Interfaces.

The subjective data will be used to measure two different aspects of the system. The first metric to be measured is the system’s usability. In order to measure the usability of both User Interfaces, a tool developed by John Brooke at Digital Equipment Corporation in 1986 called the System Usability Scale (SUS) will be implemented. The other aspect under investigation is the mental workload imposed on the participants in attempting to complete their assigned tasks on each interface. The interfaces’ mental workloads will be examined using an adaptation of the NASA Task Load Index (NASA-TLX) tool which was developed at the National Aeronautics and Space Administration (NASA) in 1986. This adapted version of the NASA-TLX tool, known as Raw TLX (RTLX) has been optimised to measure the mental workload requirements of a software solution, in particular one which implements Virtual Reality technologies.

3.3.1 System Usability Scale (SUS)

The feedback regarding the usability of the interfaces from this project's experiment is collected through the use of a well-established tool for usability testing called the System Usability Scale (SUS). The System Usability Scale is a ten-question survey, with alternating positive and negative statements, each of which the participant indicates on a five-point scale how much they agree with the statement.



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
2. I found the system unnecessarily complex	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
3. I thought the system was easy to use	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
4. I think that I would need the support of a technical person to be able to use this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
5. I found the various functions in this system were well integrated	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
6. I thought there was too much inconsistency in this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
7. I would imagine that most people would learn to use this system very quickly	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
8. I found the system very cumbersome to use	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
9. I felt very confident using the system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
10. I needed to learn a lot of things before I could get going with this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

Figure 3.4 The System Usability Scale questionnaire (Brooke, 1986)

Several methods for subjective evaluation were considered, most notably instruments such as the Post-Study System Usability Questionnaire (PSSUQ), the User Experience Questionnaire (UEQ) and the Software Usability Measurement Inventory (SUMI). The SUS has, over the past three decades, proven to be an effective method for post-test evaluation of both hardware and software systems. This longevity, along with its lack of a licensing fee, were arguably the two most important factors in choosing SUS as the method for collecting the subjective user feedback.

3.3.2 Raw Task Load Index (RTLX)

The tool selected to measure the perceived mental workload of each interface was an adaptation of the NASA Task Load Index (NASA-TLX) called the Raw Task Load Index (NASA-TLX). Like the SUS, NASA-TLX has been well established as a leading tool in its field for three decades, with the RTLX adaptation gaining popularity in more recent years. Bustamante argues that this shorter, condensed version of NASA-TLX improves the validity of the questionnaires as compared to the full version (Bustamante, 2008). Hart states that across 29 different implementations of RTLX that there were examples of it outperforming, being outperformed by and performing equally as well as the full version of NASA-TLX (Hart, 2009). The main benefit of RTLX over NASA-TLX is that it is simpler and less time consuming to implement.

The questions asked in the questionnaires can be found in the appendices.

3.3.3 Objective Results Processing

The processing of the objective results was relatively straightforward for this experiment and did not follow any guidelines or theoretical models in particular, but rather would apply a descriptive statistical analysis on the data collected by the application during the experiment’s runtime. The time taken to complete the task, the total number of clicks performed by the participant, the total number of mistakes made by the participant and the Field of View (FOV) differential are the metrics which are measured and compared during the objective results processing phase.

A number of formulae are required to calculate some of these metrics. For example, to calculate the total number of mistakes made by the participant during any task, there are multiple factors which the system considers. Firstly, definitions for what counts as an error need to be established. An error is recorded on every occasion that the participant performs an action which does one of two things:

- Any action which does not progress the participant towards the completion of the experiment
- Any action which is performed incorrectly

An example of the first error would be if the participant selects the wrong tool. If, for example, the participant was asked to perform Task A, which involves colouring each square individually and they select any tool other than the ‘Fill’ tool, this would be perceived as an action which does not further their progress towards the completion of the task, so an error is recorded. Likewise, if the user repeatedly clicked to select a tool or colour which was already selected, this would be perceived as an error by the system.

The second error type is recorded when the participant attempts the correct function, but does so incorrectly. An example of this would be if they were attempting Task A and they coloured any of the squares the wrong colour (e.g.) all participants are asked to colour the top left square of the grid blue, so an error would be recorded if they filled the top square with any colour other than blue. Similarly with Task B, where participants are instructed to select the correct statement based on the pattern shown to them, if the participant selects an untrue statement an error is recorded.

The second statistic which was necessary to calculate the total number of mistakes was the number of times either the ‘Undo’ or ‘Redo’ tools were utilised by the participant. The idea behind recording this is that by hitting either of these buttons, the participant is displaying an awareness of the error they had made. For this reason, each time these buttons were selected and an action was undone/redone, a variable which was to be deducted from the total number of errors was incremented.

The last statistic to be calculated in order to output the number of mistakes made was the total number of misclicks. A misclick was recorded on every instance the participant clicked on an area of the screen which has no functionality (i.e.) they clicked an empty space. The number of misclicks is divided by the total number of clicks (as outlined in the previous section). This number is then multiplied by the total from the number of undo/redos deducted from the number of errors to produce the equations output.

$$(Errors - (Undo \vee Redo)) \times \frac{Mislicks}{Clicks}$$

Equation 1: The Mistakes Calculation Formula

Note: The output of this algorithm was supposed to be a floating-point number, but due to a programming error, it was actually output as an integer. Because of this, the results which were output have been rounded to the nearest whole number.

Another algorithm was required to create a measurable output for the FOV differential between the FOV viewed by the participant and the FOV utilised by the components of each interface. In order to calculate the differential, the following four statistics needed to be tracked by the application during the experiments:

- **Highest X-Axis value viewed** - translates to how far to their right the participant looked during the experiment.
- **Lowest X-Axis value viewed** - translates to how far to their left the participant looked during the experiment.
- **Highest Y-Axis value viewed** - translates to how far up the participant looked during the experiment.
- **Lowest Y-Axis value viewed** - translates to how far down the participant looked during the experiment.

These statistics would return a floating-point value, which could be compared to co-ordinates of the actors at the extremities of each interface. The formula to put a numerical value to both FOV metrics to allow this comparison is outlined in Fig 3.5.

$$((LowX \times -1) + HighX) + ((LowY \times -1) + HighY)$$

Equation 2: FOV calculation formula

Because the camera is in a fixed position, with only rotations allowed and not transformation, the Z-Axis (depth) values are always determined by which actor in the scene the participant is looking at. For this reason, the Z-Axis values did not need to be included as part of the calculations.

The output for the FOV differential is the percentage difference between the number for the FOV viewed by the participant and the FOV utilised by the interface. It is calculated by dividing the difference between the two values stored for the FOVs by the value of the Interface FOV. For example, if the FOV utilised by the interface was set at a value of 100, if the user viewed a FOV of either 95 or 105 the percentage differential would be 5%.

As for the other two metrics - time and clicks - neither of these required any algorithms or formulae to have their values calculated. A timer begins as soon as the party observing the experiments and is automatically stopped when the application detects that its “win condition” has been met. Regarding the clicks metric, a variable with a default value of 0 is instantiated whenever a task is started by the observing party and its value is incremented whenever the application receives a certain input.

To prepare the objective data for the triangulation process, an overall score had to be generated. This was done so using the formula devised below, in Figures 3.6 and 3.7

$$Time \rightarrow 100 - (\frac{T}{t} \times 100) \qquad Clicks \rightarrow 100 - (\frac{C}{c} \times 100)$$

$$Mistakes \rightarrow 100 - (\frac{M}{m} \times 100)$$

Equation 3: Formulae for calculating the objective variables Time, Clicks and Mistakes

$$ObjectiveScore \rightarrow \frac{Time + Clicks + Mistakes}{3}$$

Equation 4: Overall objective score calculation formula

In the above formula, T represents the time taken by the individual participant for that particular task performance, whereas t represents the average time taken to complete than task on that interface. Likewise, C represents the number of clicks performed during that task attempt with c representing the average clicks required to complete the task on that interface. M represents the individual's mistakes with m representing the average number of mistakes.

For each of the variables, the actual times, clicks are mistakes of each participant are divided by the average of the relevant metric. This number is then subtracted from 100. This creates a metric on a 100-point scale (although it is possible for a task performance to receive a negative score). The average of each of the three variables is calculated to provide the objective score.

3.3.4 Data Triangulation

Denzin describes triangulation as "the combination of methodologies in the study of the same phenomenon." (Denzin, 1978). In the case of this project, the methodologies being mixed are the objective quantitative research and the subjective quantitative research.

By triangulating the subjective and objective data, an Overall Triangulated Score (OTS) which encompasses the usability and perceived mental workload of the application's two interfaces can be formed. The formula used to triangulate the data can be seen below in Figure 3.7.

$$\frac{Objective + (100 - RTLX) + SUS}{3}$$

Equation 5: Overall Triangulated Score (OTS) Calculation Algorithm

To calculate the OTS, three variables were required. The first of these was the overall objective score which was calculated using the formula which can be seen in Figure 3.6. The second variable is calculated by subtracting the RTLX score from 100. This calculation is performed because the RTLX scores work on a reverse 0-100 scale, with a lower score being preferable to a higher score. By subtracting the score from 100, a positive RTLX score receives a higher value which can be processed in the triangulation algorithm. The final variable to be taken into account is the SUS score. The average of these three variables is then calculated to produce the overall score. No weighting is used, as each variable was deemed to be no more or less important than any of the other variables.

3.4 Design Limitations

There were several limitations to the design of this experiment. Most of these limitations were imposed due to time or resource constraints, while some were due to the inexperience of the researcher.

Ideally, testing the tasks on more than two User Interfaces would have been better. The initial plan for the experiment was to develop three interfaces; one interface strongly exhibiting the Gestalt Principles of Perceptual Observation, one of which completely goes against the teachings of Gestaltism, and a third control interface which meets these two designs somewhere in between. Unfortunately, this target had to be abandoned when it became clear that developing a third interface would have been too time-consuming and infeasible. This initial plan also involved creating a third task for the participants to perform, but this was also discontinued due to the lack of the time necessary to develop a third task. By consequence of these limitations, the experiments will produce a two by two results matrix, rather than a three by three results matrix which would have allowed for a better understanding of the effects of Gestaltism over a wider variety of tasks and other such situations.

Another limitation of the experiment is the fact that an XBOX One Gamepad had to be used, rather than the ideal scenario of motion controls being implemented. Unfortunately, the Oculus Rift Developer Kit 2 (which is the Virtual Reality Head-Mounted Display (HMD) which was used for the development process) does not include the Oculus Touch, a recently released peripheral sold separately to the Oculus Rift which the user holds in each hand, allowing the system to track the position of their hands relative to the HMD. Using a game console gamepad, which utilises a button mimicking the click action of a mouse, takes away from the overall immersion that any Virtual Reality application is striving to achieve. Motion controls help to cement this immersive experience and make using a VR application entirely distinct from traditional desktop computing. For this reason, having access to motion controls would have allowed for more meaningful and interesting research.

4 IMPLEMENTATION AND RESULTS

4.1 Introduction

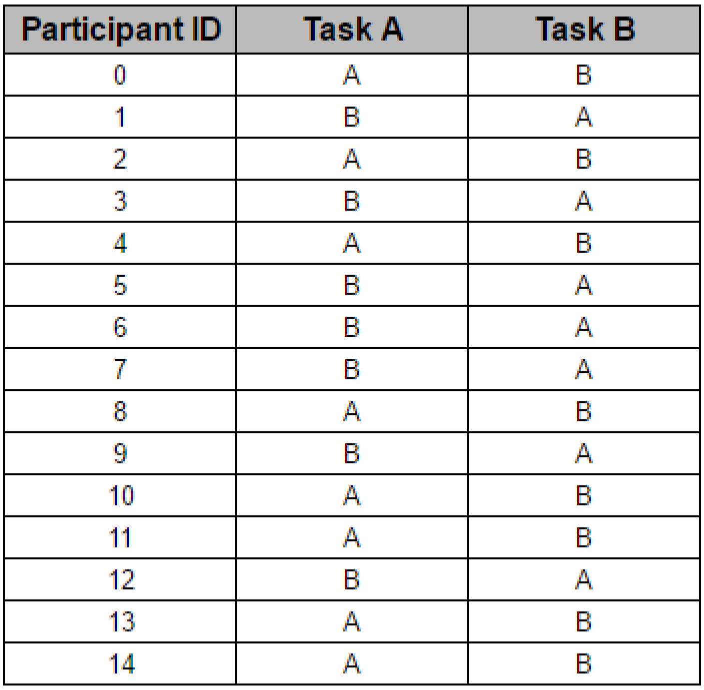
This chapter will outline the implementation of the experiments described in Chapter Three. Also to be discussed and highlighted are the results of the experiment.

4.2 Participant Observation

The Participant Observations took place over two days in December 2016 at the Dublin Institute of Technology's Kevin Street Campus. Each participant was assigned a 45 minute slot to complete the experiment. Each participant began by reading and signing an ethics/consent form. Also given to the participants upon entry was a document outlining the purpose of the experiment and Frequently Asked Questions (FAQ) about the experiment. Copies of both the consent form and the study information can be found in the appendix.

The participant observation consisted of each participant completing two tasks using the Virtual Reality software whilst wearing an Oculus Rift, with a pre-questionnaire to be filled out before completing each task and a post-questionnaire to be filled out after the task has been completed. Participants were given a short break of roughly five minutes between each task. Each participant completed both tasks and used both interfaces. No participants used the same interface twice, nor were any of them asked to complete the same task on more than one occasion.

The participants were divided into two groups, Group A and Group B. The participants in Group A were assigned Task A on Interface A and Task B on interface B, whereas the participants in Group B performed Task A on Interface B and Task B on Interface A.



Participant ID	Task A	Task B
0	A	B
1	B	A
2	A	B
3	B	A
4	A	B
5	B	A
6	B	A
7	B	A
8	A	B
9	B	A
10	A	B
11	A	B
12	B	A
13	A	B
14	A	B

Table 4.1 Participant Group Distribution

As can be seen in Table 4.1, there were eight participants in Group A and seven participants in Group B. In total, there were sixteen participants observed. One participant completed the pre-test questionnaire and completed the first task, but did not complete a post-test questionnaire or partake in the second half of the experiment, so that participant’s results will not be counted. Because of this, the dataset contains the experiment data for the remaining fifteen participants who fully completed both tasks and filled in both pre-test and post-test questionnaires.

4.2.1 Participant Demographics

The demographics of the participants was quite diverse, with seven different nationalities from four different continents represented. Of the fifteen participants, eight were native English speakers. The gender distribution was 86% male to 14% female. The age of the participants ranged from 19 years old to 44 years old, with the majority of participants being between the ages of 24 and 28. The distribution of the participants ages can be seen in Figure 4.1.

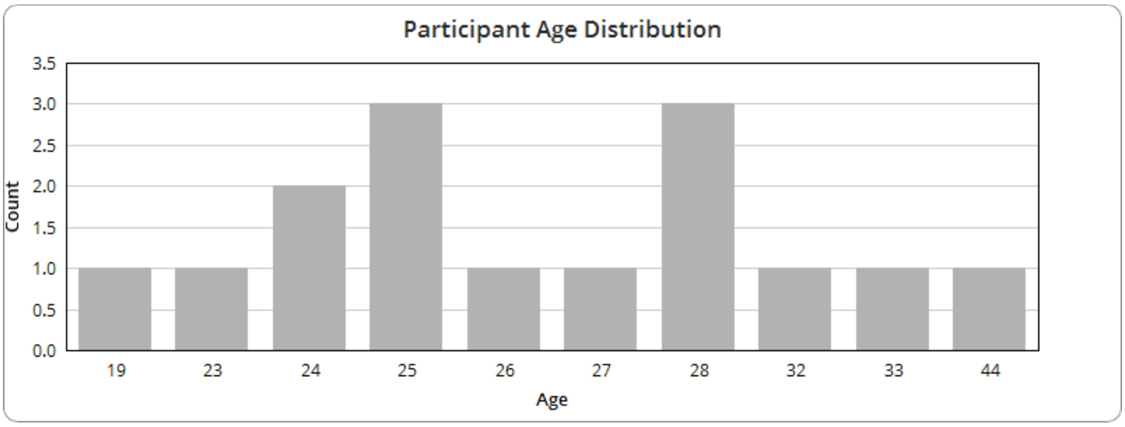


Figure 4.5 Participant Age Distribution

Out of the fifteen participants who took part in the experiment, three needed to wear glasses at all times, with another six participants requiring reading glasses. As a Virtual Reality application is by nature an entirely visual experience, it was possible that limited vision may have hindered overall performances when attempting to complete any of the assigned tasks. For this reason, the participants who required glasses were spread across the two groups as evenly as possible in order to protect the integrity of the results. None of the participants were colour-blind, which is important as Task A relies heavily on the participants' abilities to identify colours.

Another important aspect of the participants' backgrounds was their familiarity with using a gaming console gamepad, as experience using such an input device allowed for easier instruction and consequently faster performances of the task. Of the fifteen participants, four had never used a gamepad of any sort before, whereas two had used gamepads other than an XBOX controller but did not associate themselves as people who played video games or used such a controller often. Of the remaining nine participants, seven regularly used a gamepad other than the XBOX controller with the other two regularly using XBOX gamepads.

During each participant observation, only the participant and the researcher overseeing the experiments were present in the room. During the observation of two of the participants, the experiment was disturbed by another participant entering the room briefly.

4.3 Results

For the purpose of distinguishing between the two User Interfaces implemented in this experiment, the interface which exhibits the features of the Gestalt Principles will once again be referred to as Interface A, whereas the interface which lacks the Gestalt Principles in its design will again be referred to as Interface B. Task A will refer to the simpler task of manually colouring in each square, while Task B will refer to the more demanding task of applying a pattern to the grid based on a series of clues.

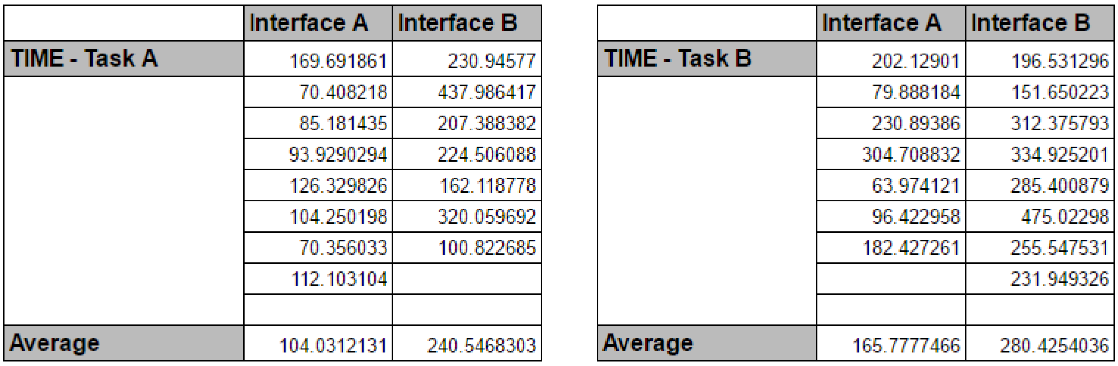
4.3.1 Objective Results

When each task was started and all components had rendered, the software began tracking objective data about each participant's attempt to complete the task they were assigned. Upon completion of the task, this data was written to file. The metrics output were the time taken to complete the task, the total number of clicks by the user in completing the task, the total number of mistakes made by the participant while attempting the task and the percentage differential between the Field of View (FOV) observed by the participant and the FOV utilised by the interface.

4.3.1.1 Time

The first objective metric to be examined is the time taken by each participant to complete each task. A global float variable was instantiated at the beginning of each task and was output to an accuracy six decimal places.

The times taken (in seconds) to complete each task across the two interfaces can be seen in the tables below. It is important to note that some of the participants were not native English speakers and that they needed to have the tasks explained to them on more than one occasion, which naturally will have slowed down their progress in completing the tasks. However, this limitation applied to tasks being performed on both User Interfaces across both tasks, and thus should not influence the results of the experiment in any meaningful way.



	Interface A	Interface B
TIME - Task A	169.691861	230.94577
	70.408218	437.986417
	85.181435	207.388382
	93.9290294	224.506088
	126.329826	162.118778
	104.250198	320.059692
	70.356033	100.822685
	112.103104	
Average	104.0312131	240.5468303

	Interface A	Interface B
TIME - Task B	202.12901	196.531296
	79.888184	151.650223
	230.89386	312.375793
	304.708832	334.925201
	63.974121	285.400879
	96.422958	475.02298
	182.427261	255.547531
		231.949326
Average	165.7777466	280.4254036

Table 4.2Time

As can be seen in Table 4.2, the average time taken to complete each task is vastly lower for Interface A than the time taken to complete the same task on Interface B. Task A took on average 104 seconds to be completed on Interface A, whereas on Interface B this same task needed 241 seconds to be completed. This represents a huge difference, with the tasks on Interface B taking on average 231% longer to complete. As well as having a shorter average completion time, four of the five fastest completions of Task A were achieved on Interface A, including all three of the fastest recorded times. At the other end of the scale, all five of the slowest completion times which were recorded were when the Task was being attempted on Interface B. The slowest attempt on Interface A lasted 170 seconds, whereas on Interface B there were five attempts which took over 200 seconds to finish successfully.

It was much the same story with Task B. The average completion time for Task B on Interface A was approximately 166 seconds, compared to taking 280 seconds on Interface B. This represents an increase of 169% to complete the task on Interface B in comparison to the completion time for the same task on Interface A. As was the case with Task A, four of the five fastest recorded times for Task B were when the participant was attempting to complete the task on Interface A. Likewise, the bottom end of the scale was dominated by attempts which were made on Interface B, with six of the seven slowest attempts being recorded on the Interface which omitted the Gestalt Principles of Perceptual Observation.

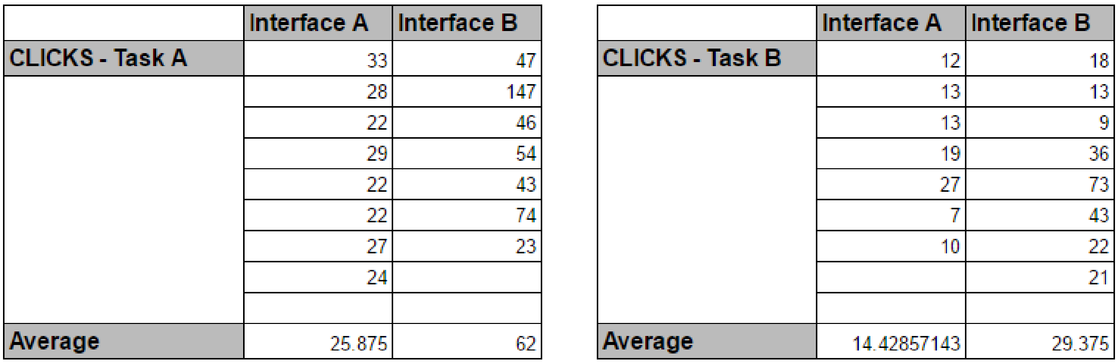
4.3.1.2 Clicks

The next metric to be examined is the number of clicks required by the user to complete the task. A global integer variable named clickCount was instantiated at the beginning of each task. This variable was updated upon certain events, as will be outlined in the coming paragraphs. The value stored for clickCount was output upon completion of each task and the value reset to its default value of 0.

When undertaking each task, the participants only had two buttons on the XBOX Gamepad which offered any functionality. The first of these was the ‘A’ Face Button on the Gamepad, which was used to select whichever actor was being hovered over - an equivalent to the left click function on a typical mouse I/O system. A click is recorded on every occasion when the participant hits the ‘A’ Face Button on the XBOX One Gamepad.

The second button on the gamepad which offered functionality to the user was the left trigger button. When this button was pressed, whichever scene component was at the centre of the screen - therefore being the actor which the participant would select if they were to hit the ‘A’ Face Button - would highlight, acting as a guide to the user so that they could get a better understanding of how the camera and focusing worked with a Head Mounted Display. This comes from the ideas outlined in the Gestalt Principle of Figure/Ground. As this was more of a guide button and not a button which progressed the completion of the task, presses of the left trigger button did not increment the value stored for the number of clicks. None of the other buttons on the gamepad had any effect on the clickCount variable.

The minimum possible number of clicks required to complete Task A was 19. For Task B, the minimum number of clicks the participants needed to make to successfully complete the task was five clicks. Both numbers were true for completing the tasks on both Interface A and Interface B.



	Interface A	Interface B
CLICKS - Task A	33	47
	28	147
	22	46
	29	54
	22	43
	22	74
	27	23
	24	
Average	25.875	62

	Interface A	Interface B
CLICKS - Task B	12	18
	13	13
	13	9
	19	36
	27	73
	7	43
	10	22
		21
Average	14.42857143	29.375

Table 4.3 Clicks

The results for the actual number of clicks taken by each participant across both tasks indicate a big difference between Interface A and Interface B.

With Task A, the average number of clicks performed by each participant when observed using Interface A was 25.875 clicks. This represents a margin of 6.875 clicks more than the absolute minimum number of clicks required, which equates to a total number of clicks roughly 36% higher than the best possible performance. When the participants were undergoing the task on Interface B, the average number of clicks shoots up enormously to a value of 62. This margin of 43 extra clicks equates to an increase of 226%. With the exception of one well performing participant on Interface B, all the attempts on Interface B recorded more clicks than every attempt on Task A.

However, it should be noted that one participant struggled greatly to complete the task on Interface B. This participant's total of 147 clicks is an extreme outlier and brings the average number of clicks for this test group up to 62, whereas without this participant's data included, the average number of clicks needed to complete Task A on Interface B drops significantly down to 47.8333 clicks. This lower margin of 28.8333 clicks equates to a 151% increase on the best possible performance.

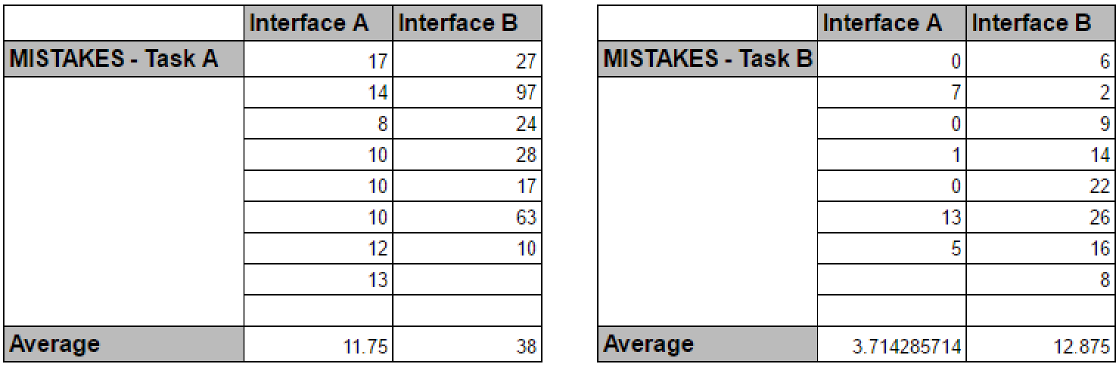
The results of Task B mirrored those of Task A. The average number of clicks performed in completing Task B on Interface A worked out at 14.43 clicks. This represents an increased click rate of 189% compared to the best possible performance of this task in terms of clicks. When performed on Interface B, the average number of clicks rises to 29.375, roughly double the average number of clicks required on Interface A. Like the results of the time metric, four of the five “worst” performances were attempts made on Interface B, while four of the five “best” performances were recorded on Interface A.

4.3.1.3 Mistakes

> **Note:** The output for this metric was supposed to be a floating-point number, but due to a programming error, it was actually output as an integer. As a result of this, the results which were output have been rounded to the nearest whole number.

The mistakes metric was calculated with an algorithm based on a number of statistics which were recorded during the experiment process. The algorithm used can be seen in Fig 3.4.

The tables in Table 4.4 show the results for the total number of mistakes made by each participant on each task.



	Interface A	Interface B
MISTAKES - Task A	17	27
	14	97
	8	24
	10	28
	10	17
	10	63
	12	10
	13	
Average	11.75	38

	Interface A	Interface B
MISTAKES - Task B	0	6
	7	2
	0	9
	1	14
	0	22
	13	26
	5	16
		8
Average	3.714285714	12.875

Table 4.4 Mistakes

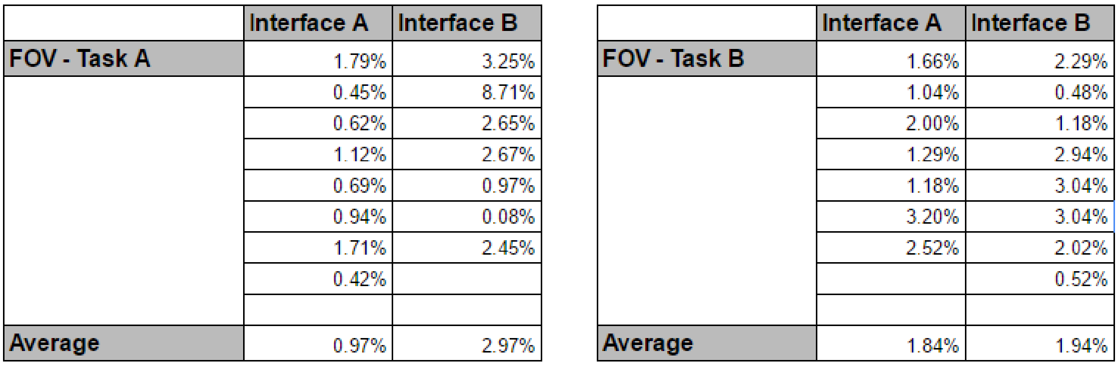
For Task A, the number of mistakes made ranged from a minimum of eight to a maximum of seventeen on Interface A, with the average number of mistakes made working out at 11.75 mistakes per participant. This is in stark contrast with the results for Task A when performed on Task B. While the lowest number of mistakes made only increased by two, up to ten mistakes, the highest number of mistakes made by participants attempting Task A on Interface B was recorded at 97, 80 mistakes more than any participant attempting the same task on Interface A. It will come as no surprise that this total of 97 mistakes was made by the same participant identified earlier as being an outlier in previous categories. However, all bar one of the participants attempting to complete Task A on Interface B recorded at least seventeen mistakes. One participant recorded just ten mistakes on Interface B, one participant equalled the highest number of mistakes made on Interface A, but after those two participants, all of the others recorded more mistakes than the worst performing Interface A attempt.

Three participants who attempted Task B on Interface A successfully completed the task without making a single mistake. Naturally this resulted in a low average mistake count for this task on Interface A, with the mean number of mistakes being calculated as roughly 3.714 mistakes per participant. On the other hand, with Interface B, all participants made at least two mistakes, with seven of the eight participants making more mistakes than the average number of mistakes made on Interface A for the same task. The average number of mistakes made when attempting Task B on Interface B rose to 12.875 mistakes per user. This represents 246% more mistakes made on Interface B for Task B when compared to Interface A.

4.3.1.4 Field of View (FOV) Differential

The Field of View (FOV) differential refers to the amount of the 3-Dimensional space which the participant viewed/utilised during the task completion compared to the actual space which is filled by the Interface.

The results of the FOV differentials can be seen below in Table 4.5.



	Interface A	Interface B
FOV - Task A	1.79%	3.25%
	0.45%	8.71%
	0.62%	2.65%
	1.12%	2.67%
	0.69%	0.97%
	0.94%	0.08%
	1.71%	2.45%
	0.42%	
Average	0.97%	2.97%

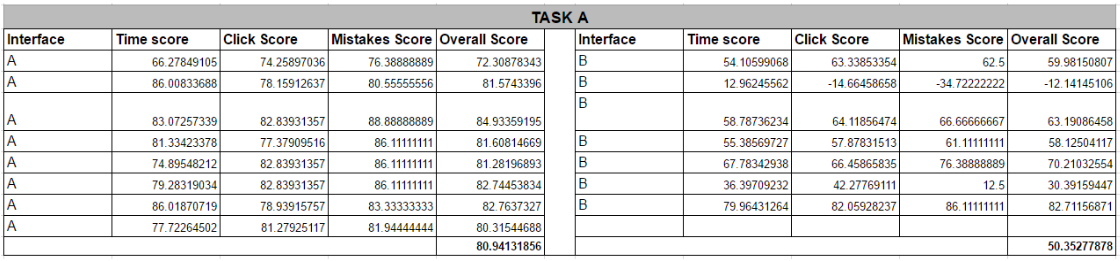
	Interface A	Interface B
FOV - Task B	1.66%	2.29%
	1.04%	0.48%
	2.00%	1.18%
	1.29%	2.94%
	1.18%	3.04%
	3.20%	3.04%
	2.52%	2.02%
		0.52%
Average	1.84%	1.94%

Table 4.5 FOV Differential

When participants attempted Task A on Interface A, the average FOV differential came to 0.97%, meaning that the FOV viewed by the participants was almost identical to the FOV utilised by the User Interface. The difference between the FOVs for Task A on Interface B was also minimal, clocking in at 2.97%. Once again, there is one outlier in the dataset, as one of the participants had a differential of 8.71%, a differential 5.46% higher than the next highest percentage.

The FOV differentials for the two interfaces were much closer for Task B. Participants who attempted Task B on Interface A averaged a differential of 1.84%, with those who performed the same task on Interface B averaging 1.94%.

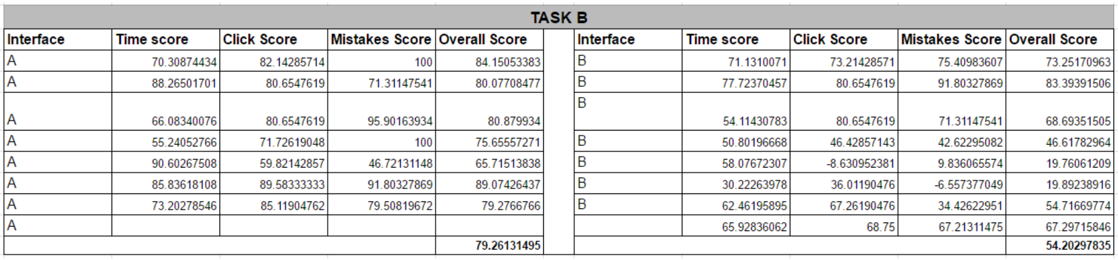
Table 4.6 displays the results from when the formula to determine each of the overall objective scores was applied to each performance of Task A.



TASK A										
Interface	Time score	Click Score	Mistakes Score	Overall Score		Interface	Time score	Click Score	Mistakes Score	Overall Score
A	66.27849105	74.25897036	76.38888889	72.30878343		B	54.10599068	63.33853354	62.5	59.98150807
A	86.00833688	78.15912637	80.55555556	81.5743396		B	12.96245562	-14.66458658	-34.72222222	-12.14145106
A	83.07257339	82.83931357	88.88888889	84.93359195		B	58.78736234	64.11856474	66.66666667	63.19086458
A	81.33423378	77.37909516	86.11111111	81.60814669		B	55.38569727	57.87831513	61.11111111	58.12504117
A	74.89548212	82.83931357	86.11111111	81.28196893		B	67.78342938	66.45865835	76.38888889	70.21032554
A	79.28319034	82.83931357	86.11111111	82.74453834		B	36.39709232	42.27769111	12.5	30.39159447
A	86.01870719	78.93915757	83.33333333	82.7637327		B	79.96431264	82.05928237	86.11111111	82.71156871
A	77.72264502	81.27925117	81.94444444	80.31544688						
				80.94131856						**50.35277878**

Table 4.6 Task A Overall Objective Results

The average score for the task performances attempted on Interface A was 80.94 out of a possible 100, with scores ranging from a lowest score of 72.3 up to a highest score of 94.93. By contrast, the average score for Task A performances on Interface B was down to 50.35, with a worst score of -12.14 and the best performance earning a score of 82.71.



TASK B										
Interface	Time score	Click Score	Mistakes Score	Overall Score		Interface	Time score	Click Score	Mistakes Score	Overall Score
A	70.30874434	82.14285714	100	84.15053383		B	71.1310071	73.21428571	75.40983607	73.25170963
A	88.26501701	80.6547619	71.31147541	80.07708477		B	77.72370457	80.6547619	91.80327869	83.39391506
A	66.08340076	80.6547619	95.90163934	80.879934		B	54.11430783	80.6547619	71.31147541	68.69351505
A	55.24052766	71.72619048	100	75.65557271		B	50.80196668	46.42857143	42.62295082	46.61782964
A	90.60267508	59.82142857	46.72131148	65.71513838		B	58.07672307	-8.630952381	9.836065574	19.76061209
A	85.83618108	89.58333333	91.80327869	89.07426437		B	30.22263978	36.01190476	-6.557377049	19.89238916
A	73.20278546	85.11904762	79.50819672	79.2766766		B	62.46195895	67.26190476	34.42622951	54.71669774
A							65.92836062	68.75	67.21311475	67.29715846
				79.26131495						**54.20297835**

Table 4.7 Task B Overall Objective Results

As can be seen in Table 4.8, the results for Task B tell a similar story. The average score for Interface A stays at roughly the mark, dropping marginally down to a score of 79.26 out of a possible 100 for Task B. The best score recorded on Interface A while attempting Task B was 89.07, with the worst score registering at 65.72. The average scores for Interface B do improve slightly from the first task, but they once again have a notably lower mean score, this time averaging 54.2. The best score of all Interface B attempts of Task B was 83.39, whereas the lowest score recorded was an alarmingly low 19.76.

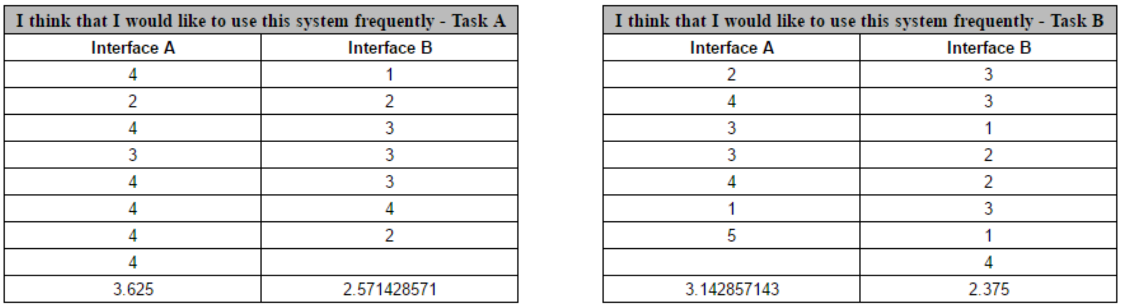
4.3.2 Subjective Results

The participants filled out questionnaires prior to attempting each task as well as after having completed the tasks, meaning all participants filled out four questionnaires; two pre-test questionnaires and two post-test questionnaires. These questionnaires were designed to receive the participants' feedback on two different aspects of the software; the usability of the software and the effect the application had on each participant regarding the mental workload required. The answers from the pre-test questionnaire would provide data for the mental workload analysis, whereas the post-test questionnaire would provide data for both the mental workload and system usability analyses.

4.3.2.1 Usability - System Usability Scale (SUS)

To measure the usability of the application, a tool developed by John Brooke at Digital Equipment Corporation in 1986 called the System Usability Scale (SUS) was implemented. The SUS is comprised of ten statements, alternating between positive and negative statements, which the participant indicates on a five point Likert Scale to what degree they agree with. These statements can be seen in Fig 3.1. This section will briefly review the results of each participant's answer to each statement as well as reviewing the overall SUS score for each Interface based on the Task assigned to the participant. Note that in the following results tables a 1 signifies that the participant indicated that they "Strongly Disagree" with the statement, whereas a 5 indicates that they "Strongly Agree" with it.

The first statement of the SUS which the participants are asked to react to reads "I think that I would like to use this system frequently".



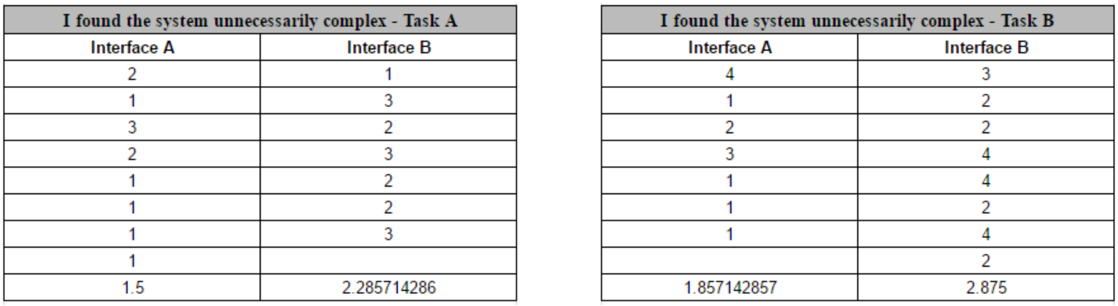
I think that I would like to use this system frequently - Task A	
Interface A	Interface B
4	1
2	2
4	3
3	3
4	3
4	4
4	2
4	
3.625	2.571428571

I think that I would like to use this system frequently - Task B	
Interface A	Interface B
2	3
4	3
3	1
3	2
4	2
1	3
5	1
	4
3.142857143	2.375

Table 4.8 SUS Q1 Results

For both tasks, Interface A received a more positive overall score than Interface B, with Interface A averaging higher than three for both tasks and Interface B averaging lower than three for both tasks.

The second statement reads "I found the system unnecessarily complex".



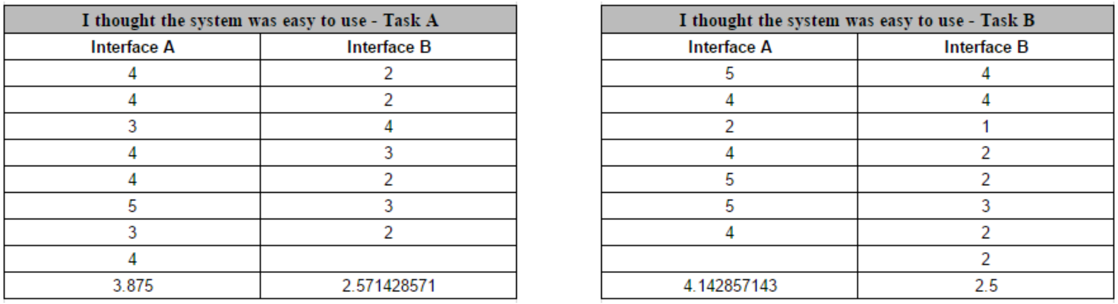
I found the system unnecessarily complex - Task A	
Interface A	Interface B
2	1
1	3
3	2
2	3
1	2
1	2
1	3
1	
1.5	2.285714286

I found the system unnecessarily complex - Task B	
Interface A	Interface B
4	3
1	2
2	2
3	4
1	4
1	2
1	4
	2
1.857142857	2.875

Table 4.9 SUS Q2 Results

Interface A also receives more positive overall scores for the second statement, with lower scores indicating disagreement that the systems were too complex for Interface A on both tasks and higher scores for Interface B indicating an agreement with the statement.

The third statement reads “I thought the system was easy to use”.



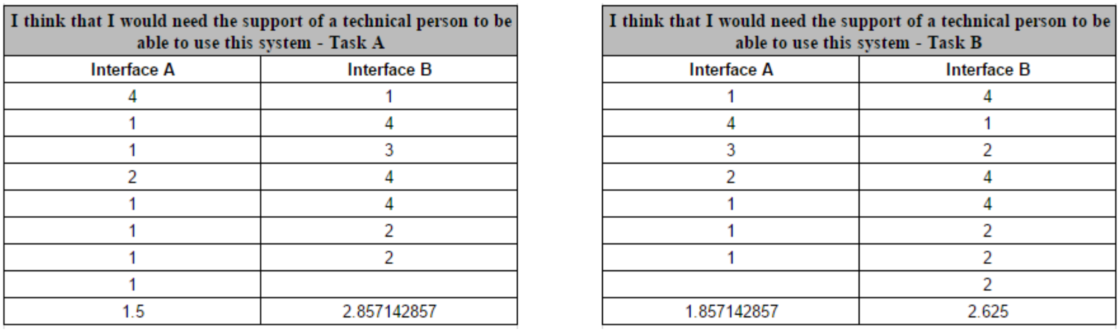
I thought the system was easy to use - Task A	
Interface A	Interface B
4	2
4	2
3	4
4	3
4	2
5	3
3	2
4	
3.875	2.571428571

I thought the system was easy to use - Task B	
Interface A	Interface B
5	4
4	4
2	1
4	2
5	2
5	3
4	2
	2
4.142857143	2.5

Table 4.10 SUS Q3 Results

The results for this question indicate that the participants agreed with the notion that the system was easy to use on Interface A, with average scores of 3.875 and 4.143 out of 5 for their ease of use for Task A and Task B respectively. For Interface B, these numbers drop, both to in and around the 2.5 out of 5 mark.

The fourth statement reads “I think that I would need the support of a technical person to be able to use this system”.



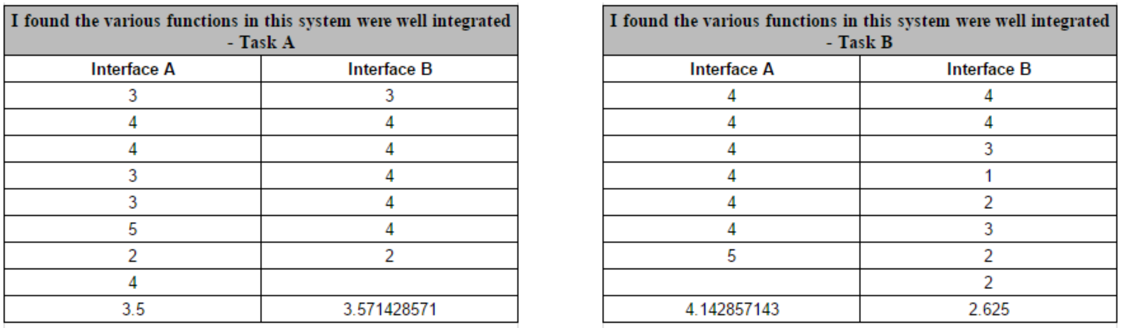
I think that I would need the support of a technical person to be able to use this system - Task A	
Interface A	Interface B
4	1
1	4
1	3
2	4
1	4
1	2
1	2
1	
1.5	2.857142857

I think that I would need the support of a technical person to be able to use this system - Task B	
Interface A	Interface B
1	4
4	1
3	2
2	4
1	4
1	2
1	2
	2
1.857142857	2.625

Table 4.11 SUS Q4 Results

The reactions to this statement were generally of disagreement for tasks performed on Interface A and of mild agreement for tasks performed on Interface B. This is indicated by the fact that both Interface A averages were less than two whereas the Interface B averages were both in the region of 2.5-3 out of 5.

The fifth statement reads “I found the various functions in this system were well integrated”.



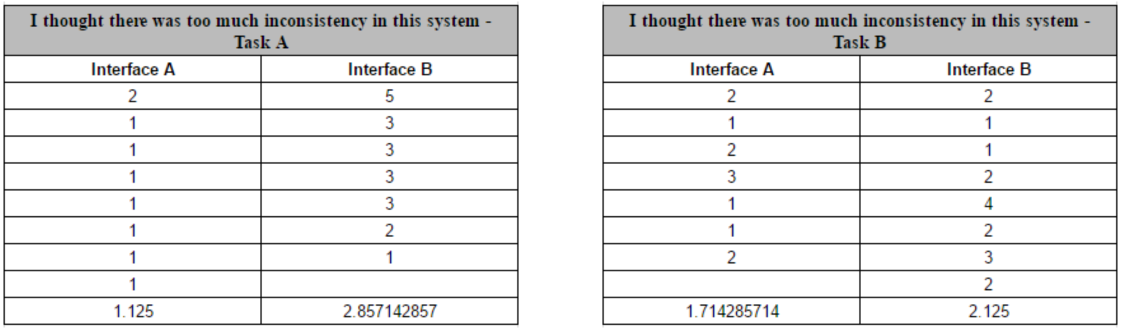
I found the various functions in this system were well integrated - Task A	
Interface A	Interface B
3	3
4	4
4	4
3	4
3	4
5	4
2	2
4	
3.5	3.571428571

I found the various functions in this system were well integrated - Task B	
Interface A	Interface B
4	4
4	4
4	3
4	1
4	2
4	3
5	2
	2
4.142857143	2.625

Table 4.12 SUS Q5 Results

This is the first metric where there is an instance of Interface B outperforming Interface A. For Task A, the average result was 3.5, with Interface B receiving a marginally higher average score of 2.571. By contrast, with Task B, Interface A performed significantly better than Interface B, with the interfaces scoring 4.143 and 2.625 respectively.

The sixth statement reads “I thought there was too much inconsistency in this system”.



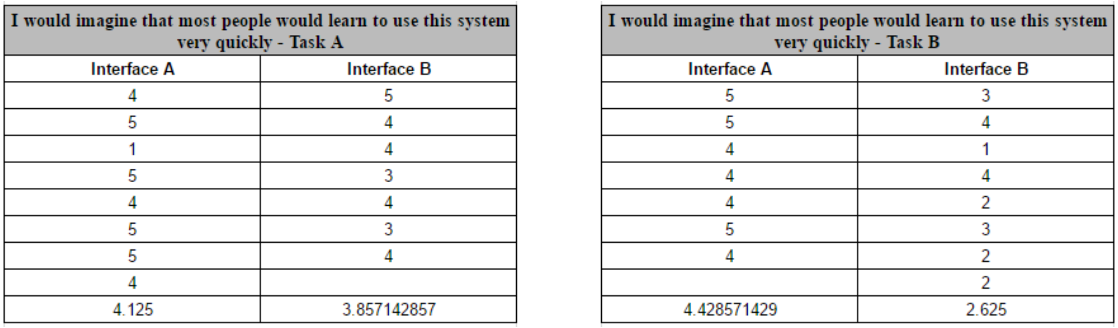
I thought there was too much inconsistency in this system - Task A	
Interface A	Interface B
2	5
1	3
1	3
1	3
1	3
1	2
1	1
1	
1.125	2.857142857

I thought there was too much inconsistency in this system - Task B	
Interface A	Interface B
2	2
1	1
2	1
3	2
1	4
1	2
2	3
	2
1.714285714	2.125

Table 4.13 SUS Q6 Results

The results across both tasks indicate that the participants felt there was a good level of consistency in Interface A, with the average scores for Interface A on both task both being calculated to a value less than two. The average scores for Interface B were both higher that their Interface A counterparts, with the averages from both tasks working out to be greater than two.

The seventh statement reads “I would imagine that most people would learn to use this system very quickly”.



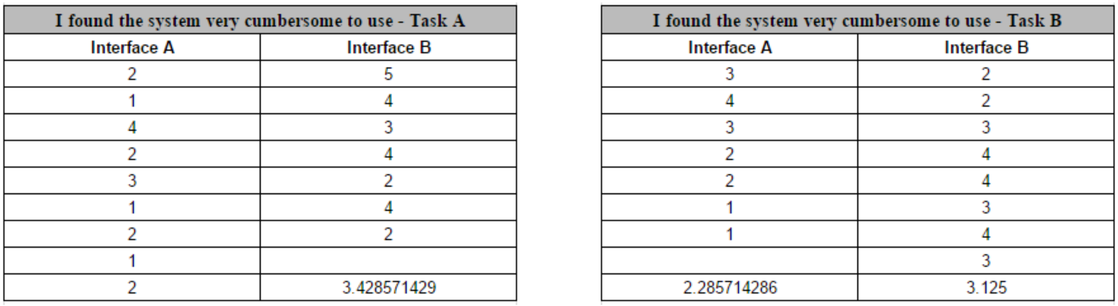
I would imagine that most people would learn to use this system very quickly - Task A	
Interface A	Interface B
4	5
5	4
1	4
5	3
4	4
5	3
5	4
4	
4.125	3.857142857

I would imagine that most people would learn to use this system very quickly - Task B	
Interface A	Interface B
5	3
5	4
4	1
4	4
4	2
5	3
4	2
	2
4.428571429	2.625

Table 4.14 SUS Q7 Results

The average scores for Interface A across both tasks were higher than four for this statement, meaning that the participants Strongly Agreed that they felt Interface A was easy to learn quickly. For Interface B, the scores for Task A maintained a high average score of 3.857, which is unsurprising due to the simple nature of Task A. For the more complex Task B, the average score dropped from 4.429 on Interface A to just 2.625 on Interface B.

The eighth statement reads “I found this system very cumbersome to use”.



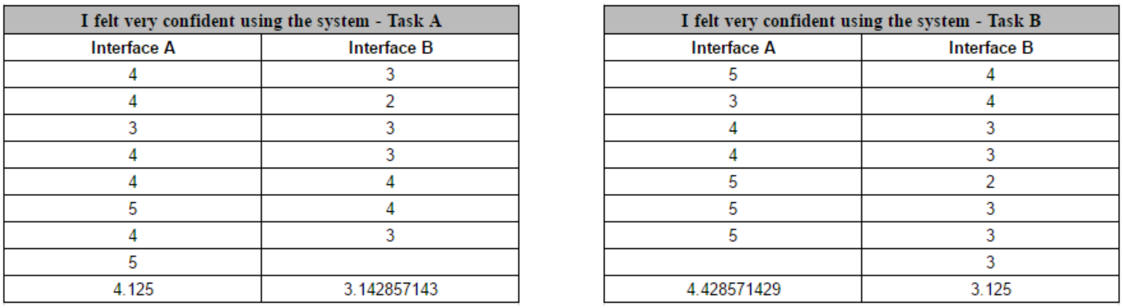
I found the system very cumbersome to use - Task A	
Interface A	Interface B
2	5
1	4
4	3
2	4
3	2
1	4
2	2
1	
2	3.428571429

I found the system very cumbersome to use - Task B	
Interface A	Interface B
3	2
4	2
3	3
2	4
2	4
1	3
1	4
	3
2.285714286	3.125

Table 4.15 SUS Q7 Results

With scores hovering around the two mark, the participants indicated that they did not find Interface A particularly cumbersome to make use of for either Task A or Task B. When questioned about Interface B, the scores indicated that the users found that UI to be clunkier than Interface A with both tasks averaging a score greater than three on that interface.

The ninth statement reads “I felt very confident using the system”.



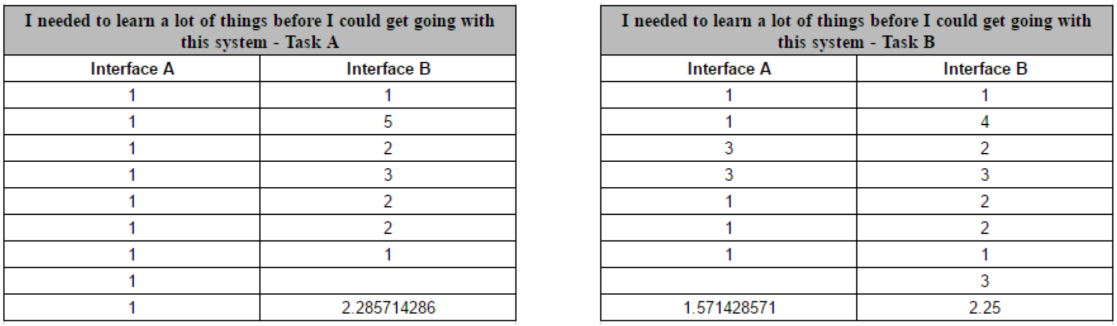
I felt very confident using the system - Task A	
Interface A	Interface B
4	3
4	2
3	3
4	3
4	4
5	4
4	3
5	
4.125	3.142857143

I felt very confident using the system - Task B	
Interface A	Interface B
5	4
3	4
4	3
4	3
5	2
5	3
5	3
	3
4.428571429	3.125

Table 4.16 SUS Q9 Results

Once again, the results for both tasks on Interface A earned average scores above four, giving an indication that the participants strongly believed that in their own abilities to complete their assignments on the given UI. The scores drop to roughly 3.1 for both tasks on Interface B, implying that while the participants still felt comfortable and confident using Interface B, they did not feel the same level of comfort as those who performed the tasks on the Gestaltist UIs did.

The tenth and final statement reads “I needed to learn a lot of things before I could get going with this system”.



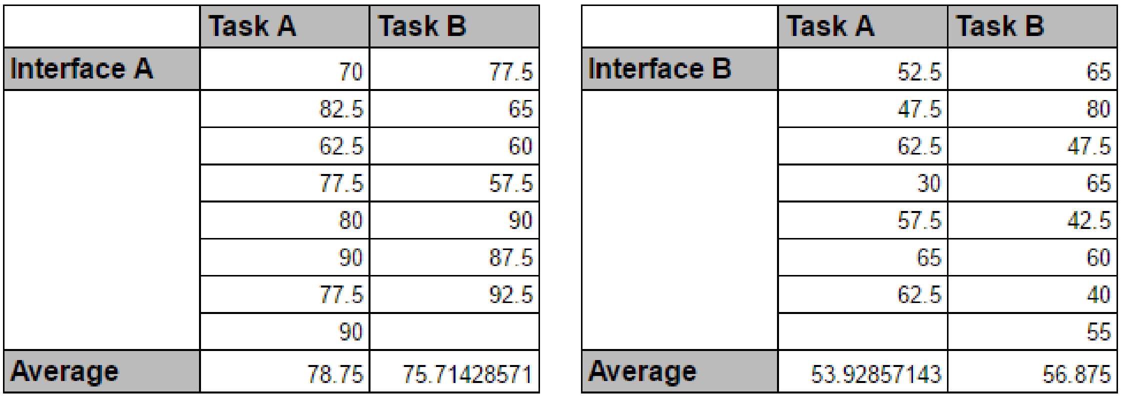
I needed to learn a lot of things before I could get going with this system - Task A	
Interface A	Interface B
1	1
1	5
1	2
1	3
1	2
1	2
1	1
1	
1	2.285714286

I needed to learn a lot of things before I could get going with this system - Task B	
Interface A	Interface B
1	1
1	4
3	2
3	3
1	2
1	2
1	1
	3
1.571428571	2.25

Table 4.17 SUS Q10 Results

All participants who completed Task A on Interface A replied to the final question of the questionnaire to say that they Strongly Disagreed that they felt they needed to learn a lot of things before being able to find their feet with this system. Similarly for Task B, the average score of 1.571 indicates that the other set of participants found it easy to dive straight into the application, despite being assigned the more taxing task. With Interface B, both tasks achieved an average score of roughly 2.25, suggesting that the participants felt they had to learn slightly more before being able to get going on the control interface.

After having examined each of the individual metrics of the SUS, next to be processed was the overall SUS scores given by each participant to each Interface for the tasks they performed on them. The results of the SUS scoring formula being applied to the result sets can be seen in Table 4.19 below. For both Task A and Task B, Interface A averages significantly higher results than Interface B does. The average score for Task A on Interface A is a very respectable 78.75 out of a possible 100. By contrast, the results for performances of Task A on Interface B are mediocre, with an average score of just 53.93 out of the maximum 100 points. These results are consistent with the findings of Task B where Interface A greatly outperformed Interface B. With the more complex task, Interface A again scored admirably, with an average result of 75.71. As was the case with Task A, Interface B’s performance left much to be desired, averaging a score of 56.875.



	Task A	Task B
Interface A	70	77.5
	82.5	65
	62.5	60
	77.5	57.5
	80	90
	90	87.5
	77.5	92.5
	90	
Average	78.75	75.71428571

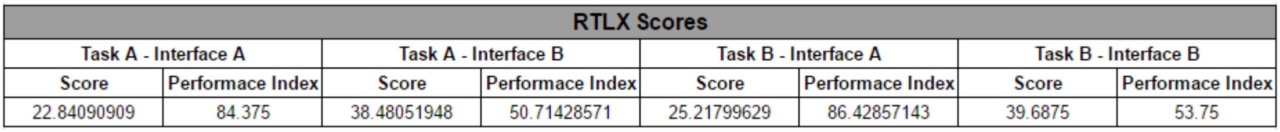
	Task A	Task B
Interface B	52.5	65
	47.5	80
	62.5	47.5
	30	65
	57.5	42.5
	65	60
	62.5	40
		55
Average	53.92857143	56.875

Table 4.18 Overall SUS Scores

The average SUS score across all recorded task attempts works out at 66.4166667. Out of the fifteen tasks performed on Interface A, eleven of the SUS scores are higher than the overall average mark. On the other hand, out of the fifteen task performances on Interface B, only one participant gave the system a usability score higher than the overall average mark. The top seven scores were all recorded on Interface A, whereas the eight lowest scores were all taken from tasks performed on Interface B.

4.3.2.2 Mental Workload - Task Load Index

With regards to the post-test questionnaires, there were two statistics whose examination of were most important. These were the overall Raw TLX (RTLX) score and the averages of each participant's own assessment of their performances. The results of these metrics should give a decent idea of just how mentally taxing each interface was for each task.



RTLX Scores							
Task A - Interface A		Task A - Interface B		Task B - Interface A		Task B - Interface B	
Score	Performace Index	Score	Performace Index	Score	Performace Index	Score	Performace Index
22.84090909	84.375	38.48051948	50.71428571	25.21799629	86.42857143	39.6875	53.75

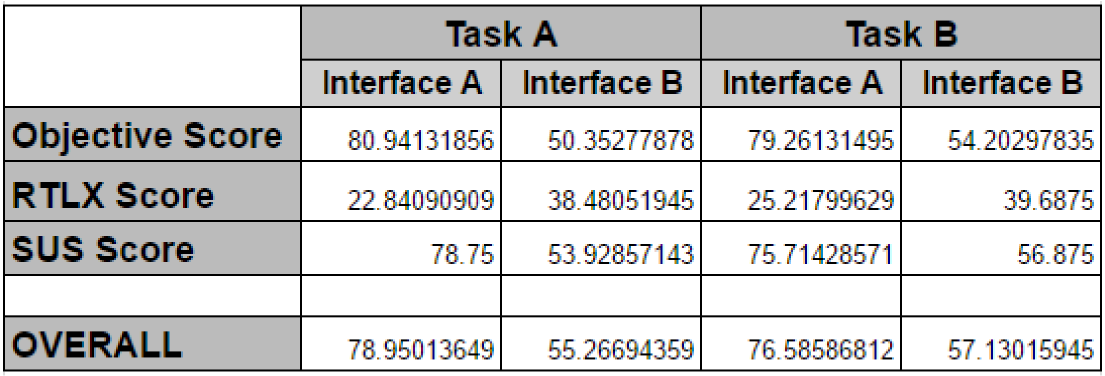
Table 4.19 The average overall RTLX and self-assessed performance scores

Table 4.19 gives us a good indication of the mental workloads of each interface for each task, also hinting at a correlation between the two statistics. Under the score heading is the average RTLX score for the given interface on the given task. The score is marked on a 0-100 scale, with 0 denoting a low mental workload and 100 indicating a mentally taxing system. The performance index is also marked on a 0-100 scale as each participant was asked to rate their own performance of the task out of 100 on the post-test questionnaire. A lower score indicates that the participants felt they performed poorly, whereas a high score indicates that they felt confident that they had performed the task with aplomb.

The RTLX score averages for Interface A across both Tasks are both quite low, coming in at 22.84 for Task A and 25.22 on Task B. These two low mental workloads were accompanied with excellent performance ratings of 84.375 and 86.43 respectively. Conversely, the perceived mental workloads for Interface B were significantly higher. The average RTLX score for the Interface B implementations of Task A is 38.48, a 15.64% increase on its Interface A counterpart. Task B also had a noteworthy increase in perceived mental workload on Interface B, with the average score clocking at 39.6875. The performance indexes decrease dramatically when the participants were using Interface B, with scores of 50.71 for Task A and 53.75 for Task B, both representing a roughly a 33.5% decrease from the Interface A performances.

4.3.3 Triangulated Results

Having compiled the results of the RTLX, SUS and Objective metrics for each participant, the results were ready to be triangulated. The average scores for both interfaces over the two tasks can be seen below in Table 4.21. The formula to calculate the Overall Triangulated Score (OTS) can be viewed in Figure 3.7



	Task A		Task B	
	Interface A	Interface B	Interface A	Interface B
Objective Score	80.94131856	50.35277878	79.26131495	54.20297835
RTLX Score	22.84090909	38.48051945	25.21799629	39.6875
SUS Score	78.75	53.92857143	75.71428571	56.875
OVERALL	78.95013649	55.26694359	76.58586812	57.13015945

Table 4.20 Average OTS results for each Interface on each task

As would be expected having seen the results leading up to this point, the OTS results for Interface A are far more positive than those of Interface B. This is most evident when comparing the OTS results for Task A. For performances of Task A on Interface A, the average OTS result was 78.95 out of a possible 100. When this simpler task was performed on Interface B, the average OTS drops to 55.27.

The numbers for Task B also indicate a superiority across this sample group for Interface A over Interface B. The average OTS for Interface A in this instance was 76.59 compared to an average score of 57.13 for Interface B.

5 EVALUATION AND ANALYSIS

5.1 Introduction

The purpose of this chapter is to dissect, analyse and evaluate the results outlined in the previous chapter. This chapter will aim to question why the results turned out the way they did and to discuss the significance of the results with regards to the research question. As well as the descriptive statistics provided in the previous chapter, additional data analytics tools such as t-Tests will be applied to further test the difference between the two interfaces across each task as well as testing the validity of the data.

The primary purpose of a t-Test is to test a null hypothesis. Consequently, establishing the null hypothesis being tested for this section is of utmost importance. The null hypothesis can be equated to the following statement:

"*The application of the Gestalt Principles of Perceptual Observation has no effect - positive or negative - on the usability or perceived mental workload of a Virtual Reality User Interface.*"

Each of the four main measurable metrics outlined in the previous chapter (Objective, Usability, Mental Workload and Overall Triangulated Score) will be examined. The use of statistical tools such as a t-Test was particularly important for Objective results set and the Overall Triangulated Score (OTS) results as these results were generated by formulas I developed myself, rather than tried and tested formulas.

Due to the fact that there were fifteen total participants spread across two groups, the sample sizes for each task-interface combination were uneven. Thus, a paired t-Test could not be performed. Instead, the t-Tests performed for the data analysis in this chapter are homoscedastic independent two-sample t-Tests.

5.2 Objective Results

Firstly, we will examine the overall objective results scores, which was calculated by taking the time, clicks and mistakes from each task performance as arguments for the algorithm outlined in Figure 3.7. Wang states that one of the goals of any user interface is to allow its users to command and control the application in as comfortable a way as possible (Wang, 1998). Examining the objective metrics associated with each task performance will give as clear an insight as possible into how the participants could command and control both interfaces for each task.

The objective metrics returned positive results for backing up the hypothesis that the application of Gestalt Principles of Perceptual Observation is beneficial to the usability of a Virtual Reality User Interface. The mean overall objective scores for both tasks when performed on Interface A were notably higher than those of Interface B. When both tasks were combined, the mean score for Interface A was 80.101, whereas for Interface B the average score across both tasks was 52.278. This represents approximately a 35% reduction in the objective performance of the participants across both tasks when the UI which did not exhibit the Gestalt Principles was being used. It is important to note at this time that all participants completed tasks on both interfaces, so the likelihood that this drop in performance is due to the skill/confidence levels of the participants is minimal.

Interestingly, there was a positive correlation between the objective scores recorded and the participant's self-assessments during the post-test questionnaires. When a Pearson's Coefficient was applied between these two metrics, a coefficient of 0.8118 calculated, which indicates a strong positive correlation. This indicates that the participants were aware of their own level of performance when trying to complete the task. As Jung Schneider and Valacich point out, this is a positive attribute for any User Interface to have, as it users to properly gauge how much time they will need to dedicate to becoming comfortable with the system, which oftentimes can be a decisive factor in whether a person decides to continue using a piece of software (Jung, Schneider & Valacich, 2010)

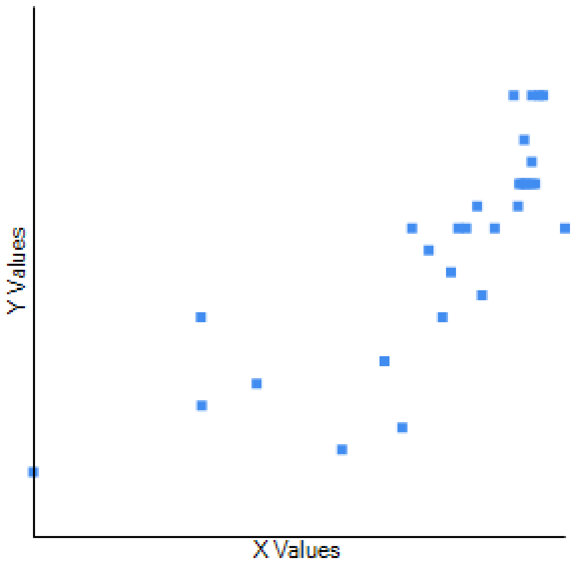


Figure 5.6 Graphical Representation of a Pearson's Coefficient between the Objective Results (X-Axis) and Participant's Self-Assessment of their Performances (Y-Axis)

Applying a t-Test to the result sets from both Tasks also yielded positive results, with both returning a value of ≤ 0.02, indicating a highly significant difference between the two User Interfaces. The results of both t-Tests indicate that the null hypothesis is incorrect, at least from the perspective of objective performance. This is hardly surprising, given the significant differences in the mean scores as well as the differences between each individual objective metric. The averages across every objective metric were widely varied, with the Gestaltist Interface posting more impressive numbers in the vast majority of cases.

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	8	80.941	3.748				
B	7	50.352	31.789				
Total	15	65.647	17.769	2.7147	13	0.0177	6.246 - 54.931

Table 5.21 Objective Results t-Test (Task A)

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	7	79.261	7.303				
B	8	54.203	23.957				
Total	15	66.732	15.630	2.6506	13	0.0200	4.635 - 45.482

Table 5.22Objective Results t-Test (Task B)

This was arguably most evident with regards to the time metric. Task A was a very simple task, with the users simply asked to fill in each square in a 3x3 grid with the colour written on that square. The fact that the average times taken to complete this rudimentary task on the two interfaces differed by 140 seconds indicates that the differences between the two interfaces had a fundamental impact on each of the participants' ability to perform the task. This is further strengthened by the fact that Task B was performed on average 115 seconds faster on Interface A than it was on Interface B. When one considers that the interface built with the Gestalt Principles in mind outperformed the control interface to the extent that it did would suggest that not only is the null hypothesis incorrect, but it also offers credibility to the idea that the implementation of the Gestalt Principles when designing a UI can improve the overall usability of a Virtual Reality application, which supports the primary hypothesis of this research project.

5.3 System Usability Scale (SUS)

While the objective metrics can inform us on the usability in terms of statistical performance, measuring the system's perceived usability through the System Usability Scale (SUS) allows us to form a much better idea of how the usability of the system affected the participants' opinions of the interfaces. Flavián, Guinalíu and Gurrea argue that the perceived usability of a system directly impacts the overall user satisfaction, which in turn acts as a catalyst for breeding user loyalty (Flavián, Guinalíu & Gurrea, 2006).

Considering Interface A received more positive average scores on all ten questions across both tasks, there is no real need to cross-examine the results of each question individually. Instead, only the overall SUS scores will be studied thoroughly. The SUS results follow a similar pattern to the objective metrics. Again, we see a significant difference between the results of the Interface A and Interface B performances of both tasks. This is highlighted by the t-Test, which again points to a significant difference between the datasets and an invalidation of the null hypothesis. With the two-tailed P values equalling 0.0007 and 0.0227 for Task A and Task B respectively, we can determine that there is enough of a significant difference between the SUS results of the two data sets to make the argument that the Gestalt Principles have a positive impact on the usability of a Virtual Reality application.

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	8	78.750	9.354				
B	7	53.929	12.235				
Total	15	66.340	10.7945	4.4489	13	0.0007	12.768 - 36.875

Table 5.23 SUS t-Test (Task A)

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	7	75.714	14.840				
B	8	56.875	12.235				
Total	15	66.295	14.1265	2.5835	13	0.0227	3.085 - 34.593

Table 5.24 SUS t-Test (Task B)

There were two metrics which the SUS scores were to have a Pearson Correlation Coefficient test against; the participant’s assessment of their own performance and the objective results. A correlation between the SUS scores and the self-assessments would indicate that the participants who deemed the tested interface to be highly usable would also have rated their own performances highly, with low usability scores corresponding to lower performance assessments. By testing the correlation between the SUS results and the objective results, we can see if the perceived usability of the interfaces matches the actual performances of each participant. The correlation between the SUS scores and self-assessments of each participant was calculated to be R=0.7151, indicating a moderate positive correlation. This tells us that the participants who thought that they performed the task well also felt that they were doing so on a user interface with a positive usability, with the users at the opposite end of the performance-assessment scale feeling that their execution of the task was held back by an interface with poor usability.

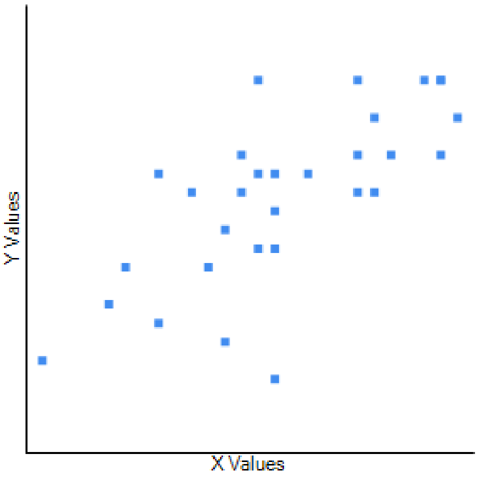
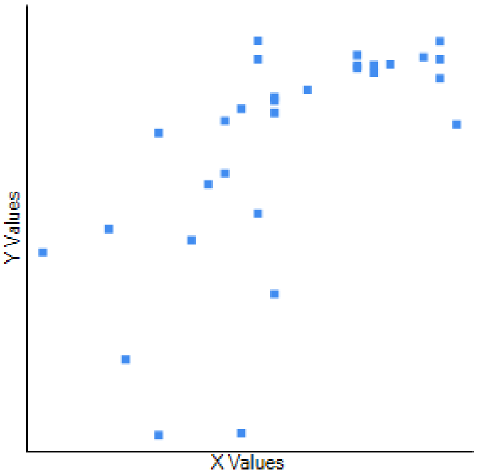


Figure 5.7 Graphical Representation of a Pearson's Coefficient between the SUS Results (X-Axis) and Participant's Self-Assessment of their Performances (Y-Axis)



.

Figure 5.8 Graphical Representation of a Pearson's Coefficient between the SUS Results (X-Axis) and Objective Results (Y-Axis)

This is important because, as Johnson points out, users like to feel that they are good at using an application, which in turn leads to a higher perceived usability for that software solution (Johnson, 2013). A moderate positive correlation between the SUS results and the objective scores was also calculated, with 0.6419 being the correlation coefficient for these two data sets. considering the positive correlation between the objective results and the performance assessment results, it was expected that this would also provide another moderate positive correlation.

With Interface A receiving an average SUS score of 77.232 across both tasks and Interface B receiving an average of 55.402, we are again given an indication that the implementation of the Gestalt Principles of Perceptual Observation is indeed beneficial for the usability of Virtual Reality applications.

5.4 Raw Task Load Index (RTLX)

As well as testing the usability of the two interfaces, another purpose of this research project was to measure the differences between the perceived mental workloads of the two UIs across each task performance.
The results from the t-Test point towards a very significant difference between the two interfaces in terms of mental workload. Upon processing the RTLX statistics, the t-Test returned with results of $t(13) = 3.1069$, $p = 0.0034$ for Task A and $t(13) = 3.2775$, $p = 0.0021$ for Task B. With p values well below 0.05 for both tasks, we are given another clear indication that the null hypothesis is likely to be invalid.

As was the case in the previous two sections, not only are the results sets significantly different, but the differences highlight a superiority for Interface A. Interface A was determined to have a low overall mental workload, with a result of 24.029 out of 100. Interface B was proven to be more mentally taxing for the participants as they attempted to perform their assigned tasks. The average mental workloads across both tasks for Interface B was a moderate 39.084 out of 100, It is interesting to note that the gaps between the interfaces perceived mental workloads actually shortened on the task which was designed to be more mentally taxing. Whereas on the simpler Task A the difference between the average scores was 15.64, that gap was narrowed to 14.47 between the means of the Task B performances. This could possibly indicate that the Gestalt Principles are slightly more effective for less mentally taxing tasks. Another possible (and more likely) explanation is that the users became more focused on the task at hand when filling out the RTLX post-test questionnaire after Task B, rather than the interface which they performed the task.

Another interesting aspect of the results was the differences between the pre-test scores depending on which interface the first task was performed on. One of the questions the participants were asked during the pre-test questionnaires was “How irritated, stressed and annoyed are you versus content, relaxed and complacent are you?”. While the results did not change much for the participants who performed their first task on Interface A, the figures for this metric were markedly different for the participants who first performed a task on Interface B. For the Group A participants, the initial pre-test questionnaire returned an average of 30 for the Frustration metric, with a mean of 28.75 for the second pre-test Frustration results. For test Group B, the initial frustration average was calculated to a value of 29.29, but prior to undertaking the second task, their average frustration had risen to 37.14. Pearson’s Coefficient tests found no correlations between any of the pre-test results and the performances of each task execution.

The results from the RTLX pre-test and post-test questionnaires provide yet more evidence which backs up the primary hypothesis of this research project that the Gestalt Principles of Perceptual Observation are beneficial to create Virtual Reality applications with better usability and lower mental workload and cognitive load requirements.

5.5 Overall Triangulated Score (OTS)

The Overall Triangulated Score (OTS) results are intended to give a comprehensive final verdict on the overall usability of the two interfaces based on the objective and subjective data provided by the participant observations. Considering the results of the previous three sections, all of which provide the data used to calculate the OTS results, it should come as no surprise that the OTS scores tell the same story as the previously discussed metrics. As expected, the t-Test once again returns with p values which indicate a significant difference between the two data sets across both tasks. The OTS results indicate that the null hypothesis is invalid.

Both tasks returned similar results in terms of the average mark for each interface. The average OTS score for all Interface A performances was 80.15731688, with a standard deviation of 5.53499388. For Interface B, the average score worked out to be 52.40621855 while the standard deviation of the Interface B results across both tasks was 26.90772694.

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	8	80.941	3.749				
B	7	50.353	31.789				
Total	15	65.647	10.7945	2.7147	13	0.0177	6.246 - 54.931

Table 5.25 OTS t-Test (Task A)

Interface	n	Mean	SD	t	df	p	95% Confidence Interval
A	7	79.261	7.303				
B	8	54.203	23.957				
Total	15	66.732	15.630	2.6506	13	0.0200	4.635 - 45.482

Table 5.26 OTS t-Test (Task B)

The low standard deviation value for Interface A highlights the fact that the clear majority of performances on the Gestalt-influenced version went very smoothly. For Interface B, the much higher standard deviation indicates that some users struggled much more than others. This was to be expected as there was one major outlier in the Interface B performances which skew the results slightly. However, even with the results of this outlier omitted, the results still favour Interface A across all metrics.

5.6 Conclusion

All of the results across each of the four topics of Objective scores, SUS results, RTLX results and OTS scores indicate that the employment of the Gestalt Principles of Perceptual Observation is highly useful for developing Virtual Reality applications. The results of this experiment indicate that perceived mental workload was reduced and usability was improved simply through the implementations of the Gestalt Principles. At this point it is quite clear that the evidence from this experiment backs up the research project’s hypothesis quite strongly. Every t-Test performed indicated that the null hypothesis was incorrect, suggesting that the Gestalt Principles have an impact on both the usability and the mental workload of a Virtual Reality application and the statistics presented suggest that this impact is a positive one.

These positive results give credence to the idea that the Gestalt Principles can be used as an effective guideline for Virtual Reality developers and designers. As Alger states, "What's particularly interesting about this section of time is that digital volumetric interfaces do not yet have established conventions. Where writing, film, television, radio, theatre, graphic design, etc. have expected elements, head-mounted displays remain conceptually open-ended. As a community, we are discovering the medium's unexpected strengths and weaknesses" (Alger, 2015). Considering the relative youth of the field of Virtual Reality and the subsequent lack of previous work in this new medium, establishing the Gestalt Principles of Perceptual Observation as a viable design convention is a positive outcome for this experiment. Because of the novelty involved with VR and the many differences it has with traditional desktop or even mobile computing, having the ability to develop applications which are usable for a plethora of different user groups could prove to be a decisive factor in the success or failure of the platform.

6 CONCLUSION

6.1 Research Overview

This research project examined the effectiveness of the Gestalt Principles of Perceptual Observation with regards to the usability and mental workloads of Virtual Reality applications when these Principles are implemented in their design. This was achieved by developing an application to be used on the Oculus Rift with two separate interfaces, one of which strongly exhibited the Gestalt Principles and one which did not. An experiment was carried out whereby participants were observed attempting two tasks on the application, performing one task on each interface. The participants filled out questionnaires which helped to determine the perceived usability and mental workloads of each interface, while performance data was being recorded during each task execution which recorded objective data. By triangulating the data from the subjective and objective datasets provided by the experiments and comparing the results of each interface for both tasks, this project would contribute information regarding how effective the Gestalt Principles are for VR designers. This is the first paper to directly research the benefits of Gestalt Psychology for Virtual Reality design.

6.2 Findings

Through a combination of primary research and the results of the experiment, this research paper has supplied evidence to support the hypothesis that the Gestalt Principles of Perceptual Observation are beneficial for Virtual Reality designers and developers. In terms of both the objective performance statistics and the subjective performance analyses of the participants, all the data gathered from the experiment indicates that the Gestalt Principles significantly improve the usability of Virtual Reality applications. Developing applications with excellent usability is becoming ever important in an industry in which User Experience is quickly developing into one of the most important aspects companies look at when designing software. By identifying a design pattern which has been proven effective in the past as a viable design convention for VR, this paper has contributed to the ever-growing body of knowledge in an exciting and rapidly expanding area of Human-Computer Interaction.

The research also suggests that the perceived mental workloads of Virtual Reality applications can be reduced by designing a user interface which follows the guidelines set in place by Gestalt Psychology. Using a new technology can be quite daunting. The fact that VR Head-Mounted Displays cover the user's vision of their immediate environment in order to better immerse them in the virtual world can also lead to stress for some users. By establishing a design convention which the evidence suggests can lower the mental workload of a VR application, no additional technostress need be instigated by mental over- or under-loads as a result of poor interface design.

6.3 Limitations

There were several limitations which significantly impacted upon the design and execution of the experiment. Three months is a very short time to have to learn how to develop a Virtual Reality application, study a set of psychological principles, build a full application which will be ready for a participant observation, carry out the experiments, process the results and write a paper about all of this. In this way, a more feasible project should probably have been chosen for this Master's Dissertation. The time constraints led to having to create a very basic application with two relatively simple tasks. The tasks which were created for this project do not serve a practical purpose other than allowing for differentiation between the two interfaces. Time constraints also meant that only fifteen participants were observed as part of the experiment. The initial hope was to have a sample size of at least 30 participants. Having such a small sample group has likely undermined much of the project's credibility, although the fact that each task performance created three datasets, measuring different aspects of the software's usability and mental workload, does help to negate the negative effects of having a smaller number of participants.

6.4 Future Work & Recommendations

This study has examined the effectiveness of the Gestalt Principles by comparing the results of all seven principles being used in tandem in one interface versus the omission of many of the principles in another. While evidence was generated to suggest that the Gestalt Principles are beneficial to designers and developers, it does not supply information as to which of these seven principles are the most impactful. This sort of information could be attained by creating a similar version of this experiment but with many more than just two interfaces, each with varying levels of each of the seven principles. In this way, we could attain a better understanding of which principles are the most beneficial and which principles can afford a lower priority for designers. Having a better understanding of the effects each principle has on an application's usability and mental workload would certainly provide strong guidelines for creating very efficient Virtual Reality applications. A project of this undertaking would also require a much larger sample size than was used for the experiments of this project. Only having fifteen participants made the sample size of this experiment unsatisfactorily small, but a lack of time and resources meant that this was as large a sample size as was achievable for this project. Ideally for future experiments, hundreds of participants would take part to provide a much larger and more significant sample group. This would allow for more diversity within the test groups and more consistent data in general.

Another positive step would be to replace the XBOX One gamepad which was used with motion controls for the experiment. The future of VR almost certainly lies with motion controls. Using a gamepad takes away from the immersive nature of VR and reminds the users that they are not actually part of the environment their visual senses are telling them they are. Motion controls certainly help to further enhance the immersive experience. Changing the main input method would undoubtedly influence the system's usability, especially when motion controls are so different from what the majority of the general population are used to when interacting with a computer. Seeing how the motion controls are affected by the Gestalt Principles (and vice versa) would make for interesting research. It would also allow for the introduction of another side of Gestalt Psychology. This paper focused on the Gestalt Principles within the visual spectrum, but the Gestalt Principles are also applicable when it comes to haptic perception. Incorporating tactile feedback into the research would create interesting data which could prove useful to designers as we move towards ubiquitous Virtual Reality devices and software. Audial perception could also be added to create an all-encompassing methodology for creating ergonomic and satisfying VR apps.

For such large undertakings, it would also be beneficial to develop tasks which are more practical and more appropriate to the platform. The tasks which were implemented in the application developed for this research project were chosen largely because of the feasibility of developing them in a small window of time, while still allowing for a decent amount of variance between the levels of Gestaltism in the two UIs. Developing software which could serve a commercial or industrial purpose, such as a Computer Aided Design (CAD) or medical training application would provide far more relevant data than an application as simple as was used for this experiment.

Bibliography

Abran, A., Khelifi, A., Suryn, W., & Seffah, A. (2003). Usability meanings and interpretations in ISO standards. *Software Quality Journal,* 11(4), 325-338.

Alger, M. (2015). Visual Design Methods for Virtual Reality. *Moving Image* September 2015.

Agarwal, C., & Thakur, N. (2014). The Evolution and Future Scope of Augmented Reality. *International Journal of Computer Science Issues (IJCSI)*, 11(6), 59.

Bailey, B. P., & Iqbal, S. T. (2008). Understanding changes in mental workload during execution of goal-directed tasks and its application for interruption management. *ACM Transactions on Computer-Human Interaction (TOCHI)*, *14*(4), 21.

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction*, *24*(6), 574-594.

Bowman, D. A., Kruijff, E., LaViola Jr, J. J., & Poupyrev, I. (2001). An introduction to 3-D user interface design. Presence: *Teleoperators and virtual environments,* 10(1), 96-108.

Bowman, D. A., & McMahan, R. P. (2007). Virtual reality: how much immersion is enough?. *Computer,* 40(7), 36-43.

Boyer, S. (2009). A virtual failure: Evaluating the success of Nintendo's Virtual Boy. *The Velvet Light Trap,* (64), 23-33.

Brooke, J. (1996). SUS-A quick and dirty usability scale. *Usability evaluation in industry*, *189*(194), 4-7.

Brooke, J. (2013). SUS: a retrospective. *Journal of usability studies*, *8*(2), 29-40.

Brooks Jr, F. P. (1999). What's real about virtual reality?. *Computer Graphics and Applications, IEEE,* 19(6), 16-27.

Burdea, G., & Coiffet, P. (2003). Virtual reality technology. Presence: *Teleoperators and virtual environments,* 12(6), 663-664.

Cakmakci, O., & Rolland, J. (2006). Head-worn displays: a review. *Display Technology, Journal of, 2*(3), 199-216.

Cao, A., Chintamani, K. K., Pandya, A. K., & Ellis, R. D. (2009). NASA TLX: Software for assessing subjective mental workload. *Behavior research methods*, 41(1), 113-117.

Chang, D., Dooley, L., & Tuovinen, J. E. (2002, July). Gestalt theory in visual screen design: a new look at an old subject. *In Proceedings of the Seventh world conference on computers in education conference on Computers in education: Australian topics-Volume 8* (pp. 5-12). Australian Computer Society, Inc..

Chang, D., Nesbitt, K. V., & Wilkins, K. (2007, January). The Gestalt principles of similarity and proximity apply to both the haptic and visual grouping of elements. *In Proceedings of the eight Australasian conference on User interface-Volume 6*4 (pp. 79-86). Australian Computer Society, Inc..

Chang, D., & Nesbitt, K. V. (2006, April). Developing Gestalt-based design guidelines for multi-sensory displays. *In Proceedings of the 2005 NICTA-HCSNet Multimodal User Interaction Workshop-Volume 57* (pp. 9-16). Australian Computer Society, Inc..

Chen, D., & Vertegaal, R. (2004, April). Using mental load for managing interruptions in physiologically attentive user interfaces. In *CHI'04 extended abstracts on Human factors in computing systems* (pp. 1513-1516). ACM.

Collaborative, A. (2001). COLLAGEN. *AI magazine*, 22(4).

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *psychometrika*, *16*(3), 297-334.

Cuevas, A., Febrero, M., & Fraiman, R. (2004). An ANOVA test for functional data. *Computational statistics & data analysis,* 47(1), 111-122.

Denzin, N. K. (1978). The research act: A theoretical introduction to sociological methods (2nd ed.). New York: McGraw-Hill.

Desurvire, H., & Wiberg, C. (2009). Game usability heuristics (PLAY) for evaluating and designing better games: the next iteration. *In Online communities and social computing* (pp. 557-566). Springer Berlin Heidelberg.

Deterding, S., Sicart, M., Nacke, L., O'Hara, K., & Dixon, D. (2011, May). Gamification. using game-design elements in non-gaming contexts. *In CHI'11 Extended Abstracts on Human Factors in Computing Systems* (pp. 2425-2428). ACM.

Dillon, A. (2001). The evaluation of software usability. *Encyclopedia of Human Factors and Ergonomics*.

Doshi, A., Cheng, S. Y., & Trivedi, M. M. (2009). A novel active heads-up display for driver assistance. Systems, Man, and Cybernetics, Part B: *Cybernetics, IEEE Transactions on,* 39(1), 85-93.

Ellis, W. D. (1999). A source book of Gestalt psychology (Vol. 2). Psychology Press.

Emma-Ogbangwo, C., Cope, N., Behringer, R., & Fabri, M. (2014). Enhancing user immersion and virtual presence in interactive multiuser virtual environments through the development and integration of a gesture-centric natural user interface developed from existing virtual reality technologies. *In HCI International 2014-Posters' Extended Abstracts* (pp. 410-414). Springer International Publishing.

Fisher, M., & Smith–Gratto, K. (1999): Gestalt theory: a foundation for instructional screen design. *Journal of Educational Technology Systems* 27(4): 361–371.

Flavián, C., Guinalíu, M., & Gurrea, R. (2006). The role played by perceived usability, satisfaction and consumer trust on website loyalty. *Information & Management, 43(1)*, 1-14.

Gallace, A., & Spence, C. (2011). To what extent do Gestalt grouping principles influence tactile perception?. *Psychological bulletin*, 137(4), 538.

Genaro Motti, V., & Caine, K. (2014, September). Understanding the wearability of head-mounted devices from a human-centered perspective. In Proceedings of the 2014 *ACM International Symposium on Wearable Computers* (pp. 83-86). ACM.

Graham, L. (2008). Gestalt theory in interactive media design. *Journal of Humanities & Social Sciences*, 2(1).

Hambidge, J. (2012). The elements of dynamic symmetry. Courier Corporation.

Han, S., Humphreys, G. W., & Chen, L. (1999). Uniform connectedness and classical Gestalt principles of perceptual grouping. *Perception & psychophysics*, 61(4), 661-674.

Hancock, P. A., & Chignell, M. H. (1988). Mental workload dynamics in adaptive interface design. *IEEE Transactions on Systems, Man, and Cybernetics*, *18*(4), 647-658.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. *Advances in psychology*, 52, 139-183.

Jex, H. R. (1988). Measuring mental workload: Problems, progress, and promises. *Advances in Psychology*, 52, 5-39.

Jick, T. D. (1979). Mixing qualitative and quantitative methods: Triangulation in action. *Administrative science quarterly*, *24*(4), 602-611.

Johnson, D., & Wiles, J. (2003). Effective affective user interface design in games. Ergonomics, 46(13-14), 1332-1345.

Johnson, J. (2013). Our Vision is Optimized to See Structure, Designing with the mind in mind: simple guide to understanding user interface design guidelines (pp. 13-28). Elsevier.

Jung, J. H., Schneider, C., & Valacich, J. (2010). Enhancing the motivational affordance of information systems: The effects of real-time performance feedback and goal setting in group collaboration environments. *Management Science*, 56(4), 724-742.

Koffka, K (1935), Visual Organization and its Laws, Principles of Gestalt Psychology (pp. 106-177), Oxon, Routledge.

Koffka, K (1935), Tri-Dimensional Space and Motion, Principles of Gestalt Psychology (pp. 265-305), Oxon, Routledge.

Lewis, J. R., & Sauro, J. (2009, July). The factor structure of the system usability scale. In *International Conference on Human Centered Design* (pp. 94-103). Springer Berlin Heidelberg.

Lewis, J. R. (2002). Psychometric evaluation of the PSSUQ using data from five years of usability studies. *International Journal of Human-Computer Interaction*, *14*(3-4), 463-488.

Lidwell, W., Holden, K., & Butler, J. (2010). Universal principles of design, revised and updated: 125 ways to enhance usability, influence perception, increase appeal, make better design decisions, and teach through design. Rockport Pub.

Longo, L. (2011). Human-Computer Interaction and Human Mental Workload: Assessing Cognitive Engagement in the World Wide Web. In *INTERACT (4)* (pp. 402–405).

Longo, L. (2012). Formalising Human Mental Workload as Non-monotonic Concept for Adaptive and Personalised Web-Design. In *UMAP* (pp. 369–373).

Longo, L. (2014). *Formalising Human Mental Workload as a Defeasible Computational Concept*. Trinity College Dublin.

Longo, L. (2015a). A defeasible reasoning framework for human mental workload representation and assessment. *Behaviour and Information Technology*, *34*(8), 758–786.

Longo, L. (2015b). Designing medical interactive systems via assessment of human mental workload. In *Int. Symposium on Computer-Based Medical Systems* (pp. 364–365).

Longo, L. (2016). Mental Workload in Medicine: Foundations, Applications, Open Problems, Challenges and Future Perspectives. In *2016 IEEE 29th International Symposium on Computer-Based Medical Systems (CBMS)* (pp. 106–111). https://doi.org/10.1109/CBMS.2016.36

Longo, L., & Dondio, P. (2015). On the Relationship between Perception of Usability and Subjective Mental Workload of Web Interfaces. In *IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology, WI-IAT 2015, Singapore, December 6-9, Volume I* (pp. 345–352).

Longo, L., Dondio, P., & Barrett, S. (2010). Enhancing Social Search: A Computational Collective Intelligence Model of Behavioural Traits, Trust and Time. *Transactions on Computational Collective Intelligence*, *2*, 46–69.

Longo, L., Rusconi, F., Noce, L., & Barrett, S. (2012). The importance of human mental workload in web-design. In *8th International Conference on Web Information Systems and Technologies* (pp. 403–409).

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile. In L. Iliadis & I. Maglogiannis (Eds.), *Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016, Proceedings* (pp. 215–229). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-44944-9_19

Mayer, R. E. (1992). Thinking, problem solving, cognition . WH Freeman/Times Books/Henry Holt & Co.

Mesbah, A., Van Deursen, A., & Lenselink, S. (2012). Crawling Ajax-based web applications through dynamic analysis of user interface state changes. *ACM Transactions on the Web (TWEB)*, 6(1), 3.

Moore, P., & Fitz, C. (1993). Using Gestalt theory to teach document design and graphics. *Technical communication quarterly*, 2(4), 389-410.

Morie, J. F., & Williams, J. (2003). The gestalt of virtual environments. In Presence Conference Aalborg, Denmark. Paper available at http://www. ict. usc. edu/publications/GestaltA. pdf. [last retrieved 23 December 2016]

Moroney, W. F., Biers, D. W., Eggemeier, F. T., & Mitchell, J. A. (1992). A comparison of two scoring procedures with the NASA task load index in a simulated flight task. *In Aerospace and electronics conference, 1992*. NAECON 1992., proceedings of the IEEE 1992 national (pp. 734-740). IEEE.

Mullis, T., Sonnenfeld, N. A., Meyers, M., & Rincon, F. (2016). Next-Gen Virtual Reality: A Comparative Study of Immersive Tendency and Differential Presence.

Nance, R., Moose, R., & Foutz, R. (1987). A statistical technique for comparing heuristics: an example from capacity assignment strategies in computer network design. *Communications of the ACM*, *30*(5), 430-442.

Newell, A., & Card, S. K. (1985). The prospects for psychological science in human-computer interaction. *Human-computer interaction*, 1(3), 209-242.

Nielsen, J. (1994, April). Usability inspection methods. In *Conference companion on Human factors in computing systems* (pp. 413-414). ACM.

Norman, D. A. (1983, December). Design principles for human-computer interfaces. In Proceedings of the SIGCHI conference on Human Factors in Computing Systems (pp. 1-10). ACM.

Norman, D. A. (2013). The design of everyday things: Revised and expanded edition. Basic books.

Oculus, V. R. (2013). Oculus rift development kit.

Poupyrev, I., Tan, D. S., Billinghurst, M., Kato, H., Regenbrecht, H., & Tetsutani, N. (2002). Developing a generic augmented-reality interface.*Computer,* 35(3), 44-50.

Preece, J., Rogers, Y., Sharp, H. (2002), Interaction Design: Beyond Human-Computer Interaction, New York: Wiley

Raudys, S. J., & Jain, A. K. (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Transactions on pattern analysis and machine intelligence*, *13*(3), 252-264.

Rauschenberger, M., Schrepp, M., Cota, M. P., Olschner, S., & Thomaschewski, J. (2013). Efficient measurement of the user experience of interactive products. How to use the user experience questionnaire (ueq). example: spanish language version. IJIMAI, 2(1), 39-45.

Reeves, L. M., Lai, J., Larson, J. A., Oviatt, S., Balaji, T. S., Buisine, S., ... & McTear, M. (2004). Guidelines for multimodal user interface design. *Communications of the ACM*, 47(1), 57-59.

Rizzo, L., Dondio, P., Delany, S. J., & Longo, L. (2016). Modeling Mental Workload Via Rule-Based Expert System: A Comparison with NASA-TLX and Workload Profile. In L. Iliadis & I. Maglogiannis (Eds.), *Artificial Intelligence Applications and Innovations: 12th IFIP WG 12.5 International Conference and Workshops, AIAI 2016, Thessaloniki, Greece, September 16-18, 2016, Proceedings* (pp. 215–229). Cham: Springer International Publishing. Retrieved from http://dx.doi.org/10.1007/978-3-319-44944-9_19

Rosson, M. B., & Carroll, J. M. (2002). Information Design, Usability engineering: scenario-based development of human-computer interaction (pp. 109-158). Morgan Kaufmann.

Rubio, S., Díaz, E., Martín, J., & Puente, J. M. (2004). Evaluation of subjective mental workload: A comparison of SWAT, NASA- TLX, and workload profile methods. *Applied Psychology*, 53(1), 61-86.

Sawyerr, W., Brown, E., & Hobbs, M. (2013). Using a Hybrid Method to Evaluate the Usability of a 3D Virtual World User Interface. *International Journal of Information Technology & Computer Science (IJITCS)*, 8(2).

Shah, R. (2014). *Mastering the Art of Unreal Engine 4-Blueprints*. Lulu. com.

Sherman, W. R., & Craig, A. B. (2003). Understanding Virtual Reality—Interface, Application, and Design. Presence: Teleoperators and Virtual Environments, 12(4), 441-442.

Shackel, B., & Richardson, S. J. (1991). Human factors for informatics usability. Cambridge university press.

Sherman, W. R., & Craig, A. B. (2002). Understanding virtual reality: Interface, application, and design. Elsevier.

Shneiderman, B. (2000). Creating creativity: user interfaces for supporting innovation. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 7(1), 114-138.

Stefanidis, D., Haluck, R., Pham, T., Dunne, J. B., Reinke, T., Markley, S., ... & Scott, D. J. (2007). Construct and face validity and task workload for laparoscopic camera navigation: virtual reality versus videotrainer systems at the SAGES Learning Center. *Surgical endoscopy*, 21(7), 1158-1164.

Sutcliffe, A., & Gault, B. (2004). Heuristic evaluation of virtual reality applications. *Interacting with computers*, 16(4), 831-849.

Sutherland, I. E. (1968, December). A head-mounted three dimensional display. In *Proceedings of the December 9-11, 1968, fall joint computer conference, part I* (pp. 757-764). ACM.

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and instruction*, 4(4), 295-312.

Thórisson, K. R. (1994, August). Simulated perceptual grouping: An application to human-computer interaction. *In Proceedings of the Sixteenth Annual Conference of the Cognitive Science Society* (pp. 876-881).

Wang, W., & Felsenstein, L. (1998). U.S. Patent No. 5,832,296. Washington, DC: U.S. Patent and Trademark Office.

Weintraub, D. J. (1992). Human factors issues in head-up display design: The book of HUD (No. SOAR-92-2). DAYTON UNIV OH RESEARCH INST.

Weiser, M. (2006). Ubiquitous computing. *Ubiquitous Computing Homepage*. March, 17(1996), 12.

Wu, C., & Liu, Y. (2009). Development and evaluation of an ergonomic software package for predicting multiple-task human performance and mental workload in human–machine interface design and evaluation. *Computers & Industrial Engineering*, 56(1), 323-333.

Xiao, Y. M., Wang, Z. M., Wang, M. Z., & Lan, Y. J. (2005). [The appraisal of reliability and validity of subjective workload assessment technique and NASA-task load index]. *Chinese journal of industrial hygiene and occupational diseases*, 23(3), 178-181.

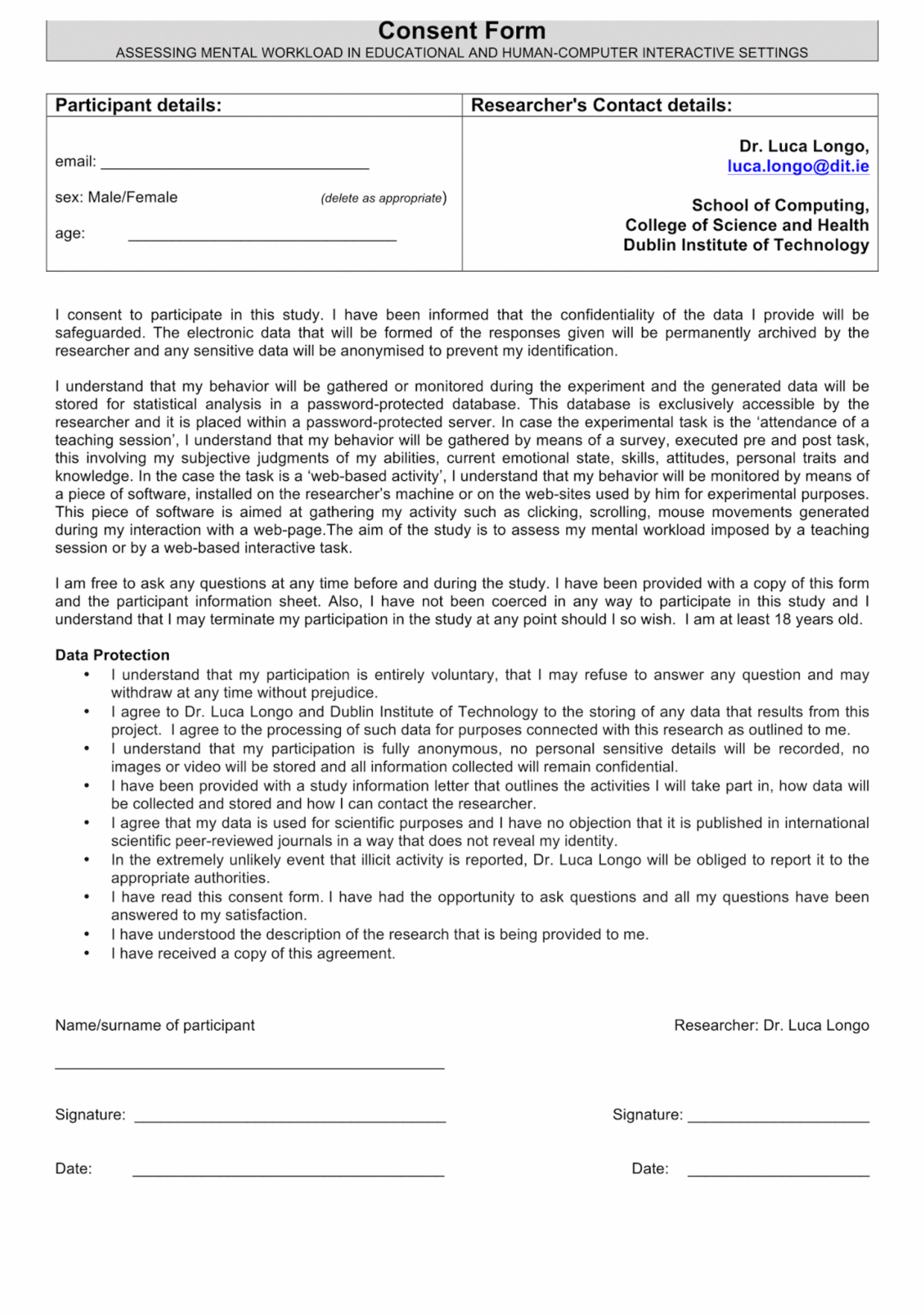
Xie, B. & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single and multi-task environments. *Work and Stress*, 14(1):74–99

Yee, K. P. (2002, December). User interaction design for secure systems. *In International Conference on Information and Communications Security* (pp. 278-290). Springer Berlin Heidelberg.

Yoon, J. W., Jang, S. H., & Cho, S. B. (2010, August). Enhanced user immersive experience with a virtual reality based FPS game interface. I*n Computational intelligence and games (cig), 2010 ieee symposium on* (pp. 69-74). IEEE.

APPENDICES

Consent Form



Consent Form

ASSESSING MENTAL WORKLOAD IN EDUCATIONAL AND HUMAN-COMPUTER INTERACTIVE SETTINGS

Participant details:	**Researcher's Contact details:**
email: ______________________________ sex: Male/Female *(delete as appropriate)* age: ______________________________	**Dr. Luca Longo,** **luca.longo@dit.ie** **School of Computing,** **College of Science and Health** **Dublin Institute of Technology**

I consent to participate in this study. I have been informed that the confidentiality of the data I provide will be safeguarded. The electronic data that will be formed of the responses given will be permanently archived by the researcher and any sensitive data will be anonymised to prevent my identification.

I understand that my behavior will be gathered or monitored during the experiment and the generated data will be stored for statistical analysis in a password-protected database. This database is exclusively accessible by the researcher and it is placed within a password-protected server. In case the experimental task is the 'attendance of a teaching session', I understand that my behavior will be gathered by means of a survey, executed pre and post task, this involving my subjective judgments of my abilities, current emotional state, skills, attitudes, personal traits and knowledge. In the case the task is a 'web-based activity', I understand that my behavior will be monitored by means of a piece of software, installed on the researcher's machine or on the web-sites used by him for experimental purposes. This piece of software is aimed at gathering my activity such as clicking, scrolling, mouse movements generated during my interaction with a web-page.The aim of the study is to assess my mental workload imposed by a teaching session or by a web-based interactive task.

I am free to ask any questions at any time before and during the study. I have been provided with a copy of this form and the participant information sheet. Also, I have not been coerced in any way to participate in this study and I understand that I may terminate my participation in the study at any point should I so wish. I am at least 18 years old.

Data Protection

- I understand that my participation is entirely voluntary, that I may refuse to answer any question and may withdraw at any time without prejudice.
- I agree to Dr. Luca Longo and Dublin Institute of Technology to the storing of any data that results from this project. I agree to the processing of such data for purposes connected with this research as outlined to me.
- I understand that my participation is fully anonymous, no personal sensitive details will be recorded, no images or video will be stored and all information collected will remain confidential.
- I have been provided with a study information letter that outlines the activities I will take part in, how data will be collected and stored and how I can contact the researcher.
- I agree that my data is used for scientific purposes and I have no objection that it is published in international scientific peer-reviewed journals in a way that does not reveal my identity.
- In the extremely unlikely event that illicit activity is reported, Dr. Luca Longo will be obliged to report it to the appropriate authorities.
- I have read this consent form. I have had the opportunity to ask questions and all my questions have been answered to my satisfaction.
- I have understood the description of the research that is being provided to me.
- I have received a copy of this agreement.

Name/surname of participant

____

Signature: ______________________________

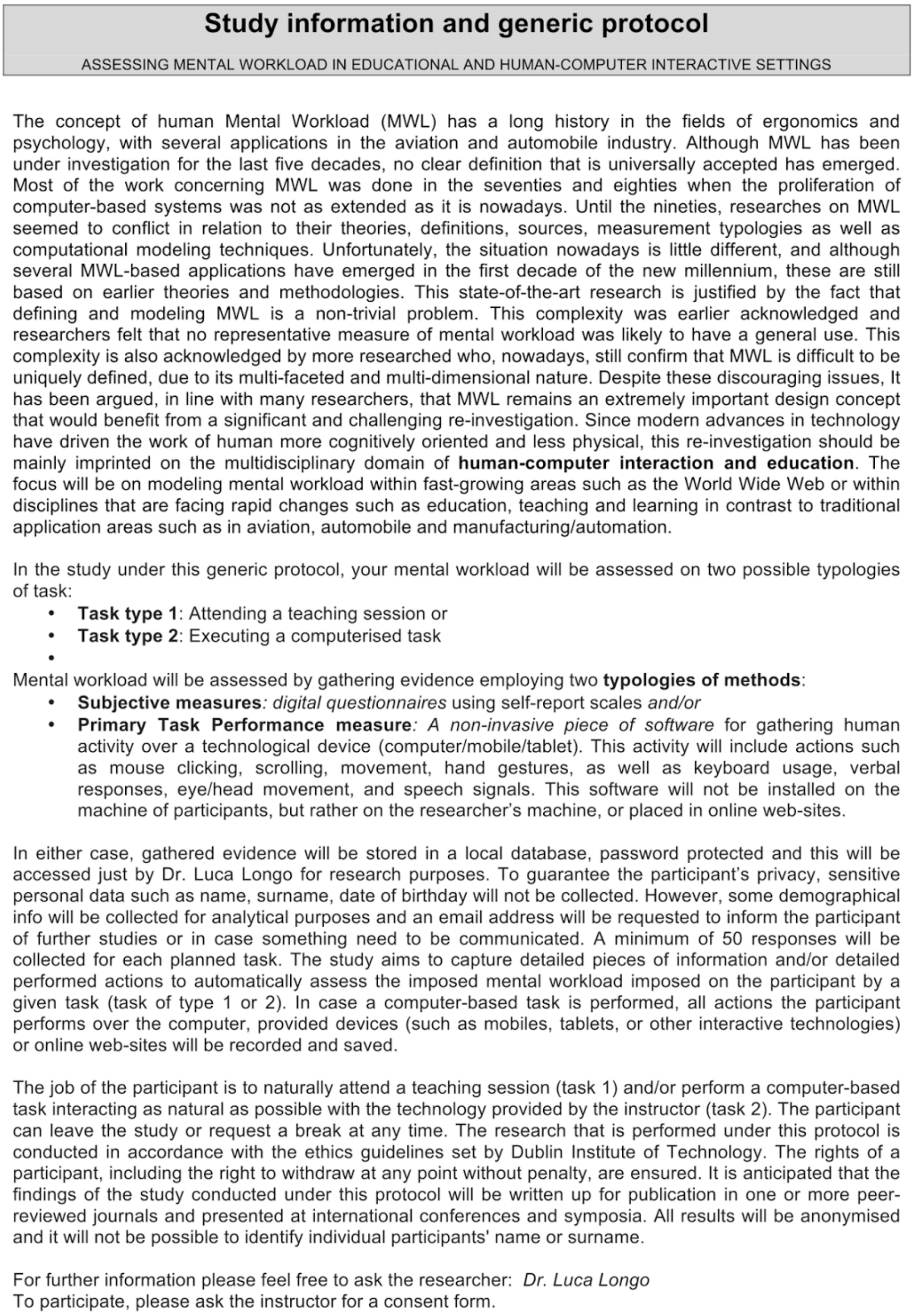
Date: ______________________________

Researcher: Dr. Luca Longo

Signature: ____________________

Date: ____________________

Study Information



Study information and generic protocol

ASSESSING MENTAL WORKLOAD IN EDUCATIONAL AND HUMAN-COMPUTER INTERACTIVE SETTINGS

The concept of human Mental Workload (MWL) has a long history in the fields of ergonomics and psychology, with several applications in the aviation and automobile industry. Although MWL has been under investigation for the last five decades, no clear definition that is universally accepted has emerged. Most of the work concerning MWL was done in the seventies and eighties when the proliferation of computer-based systems was not as extended as it is nowadays. Until the nineties, researches on MWL seemed to conflict in relation to their theories, definitions, sources, measurement typologies as well as computational modeling techniques. Unfortunately, the situation nowadays is little different, and although several MWL-based applications have emerged in the first decade of the new millennium, these are still based on earlier theories and methodologies. This state-of-the-art research is justified by the fact that defining and modeling MWL is a non-trivial problem. This complexity was earlier acknowledged and researchers felt that no representative measure of mental workload was likely to have a general use. This complexity is also acknowledged by more researched who, nowadays, still confirm that MWL is difficult to be uniquely defined, due to its multi-faceted and multi-dimensional nature. Despite these discouraging issues, It has been argued, in line with many researchers, that MWL remains an extremely important design concept that would benefit from a significant and challenging re-investigation. Since modern advances in technology have driven the work of human more cognitively oriented and less physical, this re-investigation should be mainly imprinted on the multidisciplinary domain of **human-computer interaction and education**. The focus will be on modeling mental workload within fast-growing areas such as the World Wide Web or within disciplines that are facing rapid changes such as education, teaching and learning in contrast to traditional application areas such as in aviation, automobile and manufacturing/automation.

In the study under this generic protocol, your mental workload will be assessed on two possible typologies of task:

- **Task type 1**: Attending a teaching session or
- **Task type 2**: Executing a computerised task
-

Mental workload will be assessed by gathering evidence employing two **typologies of methods**:

- **Subjective measures**: *digital questionnaires* using self-report scales *and/or*
- **Primary Task Performance measure**: *A non-invasive piece of software* for gathering human activity over a technological device (computer/mobile/tablet). This activity will include actions such as mouse clicking, scrolling, movement, hand gestures, as well as keyboard usage, verbal responses, eye/head movement, and speech signals. This software will not be installed on the machine of participants, but rather on the researcher's machine, or placed in online web-sites.

In either case, gathered evidence will be stored in a local database, password protected and this will be accessed just by Dr. Luca Longo for research purposes. To guarantee the participant's privacy, sensitive personal data such as name, surname, date of birthday will not be collected. However, some demographical info will be collected for analytical purposes and an email address will be requested to inform the participant of further studies or in case something need to be communicated. A minimum of 50 responses will be collected for each planned task. The study aims to capture detailed pieces of information and/or detailed performed actions to automatically assess the imposed mental workload imposed on the participant by a given task (task of type 1 or 2). In case a computer-based task is performed, all actions the participant performs over the computer, provided devices (such as mobiles, tablets, or other interactive technologies) or online web-sites will be recorded and saved.

The job of the participant is to naturally attend a teaching session (task 1) and/or perform a computer-based task interacting as natural as possible with the technology provided by the instructor (task 2). The participant can leave the study or request a break at any time. The research that is performed under this protocol is conducted in accordance with the ethics guidelines set by Dublin Institute of Technology. The rights of a participant, including the right to withdraw at any point without penalty, are ensured. It is anticipated that the findings of the study conducted under this protocol will be written up for publication in one or more peer-reviewed journals and presented at international conferences and symposia. All results will be anonymised and it will not be possible to identify individual participants' name or surname.

For further information please feel free to ask the researcher: *Dr. Luca Longo*
To participate, please ask the instructor for a consent form.

Frequently Asked Questions

1) Is the study anonymous?

   Yes, the studies conducted under this protocol are totally anonymous, collected data will not be linked to the participant's identity.

2) Will the participant's experience while attending a teaching session or while executing a computer-based task be altered by any monitoring technology?

   No, the participant's experience while attending a teaching session or while executing a computer-based task will not change. The monitoring technology, if applied, will be completely invisible and non-invasive.

3) Will data the participant inputs online such as logins and passwords be captured and stored somewhere?

   No, data entered online such as form input, logins, passwords, or email addresses, will not be recorded.

4) How the participant's privacy is guaranteed?

   The participant's personal data will not be stored. Logins, password, and any data entered into forms or over social networks, wikis, blogs, or other resources will never be recorded.

5) Will recorded data be linked to the participant?

   No, recorded data will never be linked to the participant. Employed software will randomly generate a code to identify the participant's interactions with provided technologies or devices over time, but this code can never be associated with the participant's personal data, computer IP, or computer MAC address as such information is never stored.

6) Is the captured data stored in a public database?

   No, the captured data will be stored in a local password-protected database behind proxy machines and firewalls within the School of Computing at Dublin Institute of Technology.

7) Who will have access to stored participants' data, and what about confidentiality?

   Only the researcher of the studies conducted under this protocol will have access to your interaction data, exclusively for research purposes. The researcher will never be able to associate any stored data with the identity of a specific participant, as this information is never stored. No one else will have the right to access any stored information.

8) What does the monitoring software look like?

   The monitoring software employed in the studies conducted under this protocol is totally transparent. It comes as either a background application – such as a plug-in/add-on for the browsers (Firefox or Chrome) that is installed only once, or a non-invasive and hidden tracking software. Nothing further is required.

System Usability Scale

	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
2. I found the system unnecessarily complex	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
3. I thought the system was easy to use	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
4. I think that I would need the support of a technical person to be able to use this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
5. I found the various functions in this system were well integrated	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
6. I thought there was too much inconsistency in this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
7. I would imagine that most people would learn to use this system very quickly	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
8. I found the system very cumbersome to use	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
9. I felt very confident using the system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5
10. I needed to learn a lot of things before I could get going with this system	☐ 1	☐ 2	☐ 3	☐ 4	☐ 5

RTLX - Pre-test Questionnaire

0 Pre-questionnaire

Frustration

How irritated, stressed and annoyed versus content, relaxed, and complacent are you?

0	25	50	75	100
very relaxed	relaxed	normal	annoyed	very frustrated

Motivation

How motivated are you to execute the next task?

0	25	50	75	100
not at all	slightly	moderately	considerably	extremely

Arousal

Are you aroused? In other words, are you sleepy/tired or fully activated/awake?

0	25	50	75	100
extremely sleepy	sleepy	normally awake	activated	fully activated

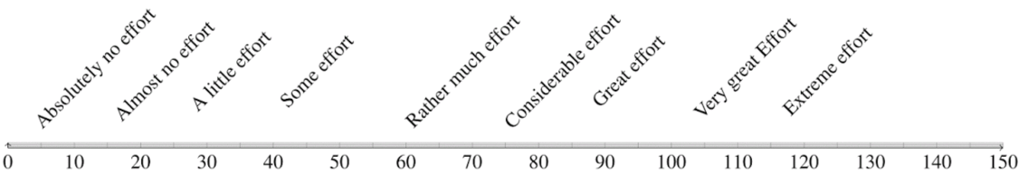
Emotional state

Is some thought or something that occurred recently in your life mentally bothering you right now?

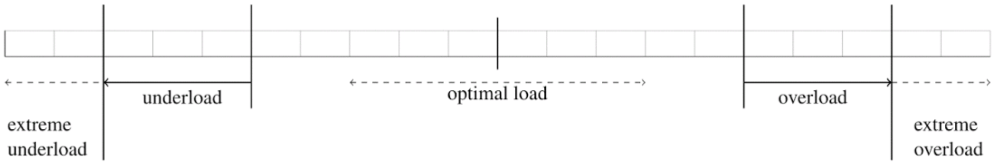
0	25	50	75	100
not at all	slightly	moderately	considerably	extremely

RTLX - Post-test Questionnaire

1 Please indicate, by marking the horizontal axis below, how much effort it took for you to execute the task you have just completed.



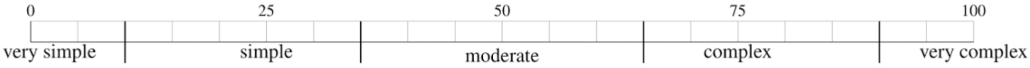
2 How much mental workload the task imposed on you?



3 In order to answer the following questions, please place an ‘x’ in one of the 20 boxes

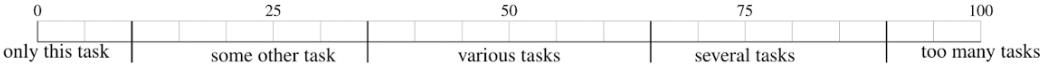
Mental demand

How much mental and perceptual activity was required by the VR task you have just executed? In other words, was the task easy or demanding, simple or complex?



Parallelism

Did you perform just this task (VR task) or were you engaged in other parallel tasks (mobile browsing/social networks, chatting, reading, conversations etc.)?



Temporal demand

How much time pressure did you feel due to the pace at which the task or its elements occurred? In other words, was the pace slow or rapid?

0	25	50	75	100
very slow	slow	moderate	rapid	very rapid

Manual activity

How much attention was required for manually responding to the elements/activities of the task (writing, coding, drawing, clicking, etc)?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Visual attention

How much attention was required for executing the task based on the information visually received?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Effort

How hard did you have to work (mentally and physically) to accomplish your level of performance during the task?

0	25	50	75	100
very low	slightly	moderately	considerably	extremely

Solving & deciding

How much attention was required by the task for activities like remembering, problem-solving, decision-making and perceiving (eg. detecting, recognising and identifying objects)?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Frustration

How irritated, stressed and annoyed versus content, relaxed, and complacent did you feel during thetask?

0	25	50	75	100
very relaxed	relaxed	normal	annoyed	very frustrated

Context bias

How often interruptions during the task occurred? In other words, were distractions (mobile, interruptions, noise, questions, other participants, etc.) not influential or did they negatively influence your attention towards the task?

0	25	50	75	100
not influential	some influence	moderate influence	considerable influence	extremely influenti

Task and space

How much attention was required by the task for spatial processing (spatially pay attention around you)?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Motivation

Were you motivated by the task?

0	25	50	75	100
not at all	slightly	moderately	considerably	extremely

Verbal material

How much attention was required by the task for verbal material (eg. reading or processing linguistic material or listening to verbal conversations)?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Skill

Did your skills have no influence or did they help executing the task?

0	25	50	75	100
not influential	some influence	moderate influence	considerable influence	extremely influenti

Auditory attention

How much attention was required for executing the task and executing its activities based on the information auditorily received?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Physical demand

How much physical activity was required by the task? In other words, was the task physically easy or demanding, slack or strenuous?

0	25	50	75	100
very easy	easy	moderate	strenuous	extremely strenuous

Speech response

How much attention was required by the task for producing speech responses (eg. engaging in conversation or talk or answering questions)?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Past knowledge/expertise

How much experience/knowledge did you have with the content of the task

0	25	50	75	100
no knowledge	some knowledge	moderate knowledge	considerable knowledge	full knowledge

Alertness

Were you alerted during the task? In other words, were you sleepy/tired or fully activated/awake?

0	25	50	75	100
extremely sleepy	sleepy	normally awake	activated	fully activated

Performance

How successful were you in the task? In other words, how satisfied were you with your level of performance?

0	25	50	75	100
no at all	slightly	moderately	considerably	extremely

Response selection

How much attention was required by the task for selecting the proper response channel (manual or speech) and its execution?

0	25	50	75	100
no attention	some attention	moderate attention	considerable attention	total attention

Experiments Results Set

Task A															
User ID	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
T1_Interface	A	B	A	B	A	B	B	B	A	B	A	B	A	A	A
T1_Frustration_Pre	50	0	35	50	30	25	25	35	50	25	10	45	25	15	25
T1_Motivation_Pre	70	60	75	65	80	75	65	35	50	75	75	60	50	65	60
T1_Arousal	80	90	45	35	80	100	75	30	50	50	60	40	100	70	90
T1_State	60	10	20	0	20	0	0	25	0	50	5	65	0	55	40
T1_Overall_Effort	30	60	20	60	40	30	80	70	20	80	10	30	10	20	40
T1_Workload	30	50	35	35	30	50	65	15	5	50	10	40	25	25	15
T1_Mental_Demand	20	10	30	35	50	50	55	15	20	75	10	35	0	5	0
T1_Parellism	0	0	0	0	25	0	0	10	0	0	0	0	0	0	0
T1_Temporal_Demand	45	20	35	0	75	35	50	15	45	25	20	50	25	40	70
T1_Manual_Activity	65	40	15	60	25	55	70	20	20	25	40	75	25	20	70
T1_Visual_Attention	60	100	65	65	75	35	70	30	70	60	60	75	25	30	40
T1_Effort	30	70	25	45	30	50	70	35	15	60	5	35	0	10	25
T1_Solving_Deciding	10	50	20	50	35	35	80	40	15	30	20	80	25	0	35
T1_Frustration_Post	45	50	10	55	20	35	65	50	0	60	5	35	25	15	0
T1_Context_Bias	0	0	5	0	0	5	0	35	0	0	0	0	0	0	0
T1_Task_Space	20	20	55	40	65	40	30	60	20	30	10	80	0	20	20
T1_Motivation_Post	45	70	55	75	65	70	60	50	10	90	35	55	50	55	40
T1_Verbal_Material	30	0	15	0	10	30	70	35	15	5	45	76	50	20	5
T1_Skill	80	0	20	0	30	30	20	10	5	10	30	35	0	0	5
T1_Auditory_Attention	5	0	0	50	0	50	75	35	5	35	15	25	50	0	5
T1_Physical_Demand	65	0	10	25	20	15	5	0	5	10	15	5	0	10	5
T1_Speech_Response	5	0	5	65	0	10	40	10	5	0	5	35	25	0	0
T1_Prior_Knowledge	5	90	45	0	10	10	5	20	0	0	75	35	75	0	0
T1_Alertness	85	90	45	75	75	70	40	35	40	65	95	75	100	85	100
T1_Performance	75	70	80	35	100	75	25	30	70	65	90	55	100	80	80
T1_Response Selection	70	0	45	35	45	35	70	30	25	65	40	50	50	5	50
T1_SUSQ1	4	1	2	2	4	3	3	3	3	4	4	2	4	4	4
T1_SUSQ2	2	1	1	3	3	2	3	2	2	2	1	3	1	1	1
T1_SUSQ3	4	2	4	2	3	4	3	2	4	3	4	2	5	3	4
T1_SUSQ4	4	1	1	4	1	3	4	4	2	2	1	2	1	1	1
T1_SUSQ5	3	3	4	4	4	4	4	4	3	4	3	2	5	2	4
T1_SUSQ6	2	5	1	3	1	3	3	3	1	2	1	1	1	1	1
T1_SUSQ7	4	5	5	4	1	4	3	4	5	3	4	4	5	5	4
T1_SUSQ8	2	5	1	4	4	3	4	2	2	4	3	2	1	2	1
T1_SUSQ9	4	3	4	2	3	3	3	3	4	4	4	4	5	4	5
T1_SUSQ10	1	1	1	5	1	2	3	2	1	2	1	1	1	1	1
T1_Time	169.691861	230.94577	70.408218	437.986417	85.181435	207.388382	224.506088	162.118778	93.9290294	320.059692	126.329826	100.822685	104.250198	70.356033	112.103104
T1_Clicks	33	47	28	147	22	46	54	43	29	74	22	23	22	27	24
T1_Mistakes	17	27	14	97	8	24	28	17	10	63	10	10	10	12	13

Task B															
User ID	**0**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**
T2_Interface	B	A	B	A	B	A	A	B	B	A	B	B	A	B	A
T2_Frustration_Pre	45	15	20	35	25	30	35	40	50	25	10	45	0	30	20
T2_Motivation_Pre	85	70	65	75	65	65	50	55	25	75	70	60	75	80	70
T2_Arousal	85	90	50	75	80	90	75	30	50	75	70	55	100	80	100
T2_State	60	10	20	0	35	0	0	15	0	50	5	65	0	60	50
T2_Overall_Effort	70	30	40	30	70	40	40	80	40	20	40	50	30	40	40
T2_Workload	45	25	60	35	70	45	35	60	15	20	35	45	50	65	40
T2_Mental_Demand	50	20	65	40	70	45	30	65	20	10	35	45	50	40	35
T2_Parellism	10	0	5	0	25	5	0	5	0	0	0	0	0	0	0
T2_Temporal_Demand	50	70	30	0	70	30	55	40	40	15	35	50	0	45	65
T2_Manual_Activity	60	20	40	60	85	45	55	45	40	20	40	50	65	40	30
T2_Visual_Attention	80	35	65	50	85	25	55	15	45	45	45	70	50	50	40
T2_Effort	50	25	55	55	75	25	40	55	20	5	35	45	35	35	25
T2_Solving_Deciding	65	45	45	40	80	30	35	50	45	10	35	50	30	25	35
T2_Frustration_Post	50	30	30	25	65	45	70	75	25	5	5	65	0	55	0
T2_Context_Bias	50	0	0	0	0	35	10	50	0	0	0	0	0	0	0
T2_Task_Space	50	20	40	20	85	25	10	30	50	0	20	60	10	50	20
T2_Motivation_Post	65	70	60	75	70	55	70	30	25	100	45	60	40	65	65
T2_Verbal_Material	85	35	60	65	0	30	40	10	5	0	5	50	50	65	30
T2_Skill	50	0	40	0	25	5	35	15	5	5	35	40	50	0	10
T2_Auditory_Attention	10	0	0	35	0	45	55	55	5	25	10	40	0	55	25
T2_Physical_Demand	65	0	30	0	0	30	25	40	25	0	10	0	0	30	5
T2_Speech_Response	5	0	0	10	0	55	55	35	5	0	5	35	0	25	5
T2_Prior_Knowledge	10	55	35	45	0	5	10	35	5	5	20	70	40	0	5
T2_Alertness	75	90	55	75	50	90	70	30	40	100	55	70	100	80	100
T2_Performance	55	100	70	75	75	80	60	20	50	100	70	40	100	50	90
T2_Response Selection	75	0	55	65	80	55	55	70	50	70	50	20	35	35	45
T2_SUSQ1	3	2	3	4	1	3	3	2	2	4	3	1	1	4	5
T2_SUSQ2	3	4	2	1	2	2	3	4	4	1	2	4	1	2	1
T2_SUSQ3	4	5	4	4	1	2	4	2	2	5	3	2	5	2	4
T2_SUSQ4	4	1	1	4	2	3	2	4	4	1	2	2	1	2	1
T2_SUSQ5	4	4	4	4	3	4	4	1	2	4	3	2	4	2	5
T2_SUSQ6	2	2	1	1	1	2	3	2	4	1	2	3	1	2	2
T2_SUSQ7	3	5	4	5	1	4	4	4	2	4	3	2	5	2	4
T2_SUSQ8	2	3	2	4	3	3	2	4	4	2	3	4	1	3	1
T2_SUSQ9	4	5	4	3	3	4	4	3	2	5	3	3	5	3	5
T2_SUSQ10	1	1	1	4	2	3	3	3	2	1	2	1	1	3	1
T2_Time	196.531296	202.12901	151.650223	312.375793	79.888184	230.89386	304.708832	334.925201	285.400879	63.974121	475.02298	255.547531	96.422958	231.949326	182.427261
T2_Clicks	18	12	13	13	9	13	19	36	73	27	43	22	7	21	10
T2_Mistakes	6	0	2	7	0	1	0	14	22	13	26	16	2	8	5