A Quality of Experience Evaluation of an Immersive Multimedia Speech & Language Application

Conor Keighrey

Submitted for the award of

Doctor of Philosophy (PhD)

Supervised by: Dr. Niall Murray & Dr. Ronan Flynn

Submitted to Athlone Institute of Technology, August 2020

DECLARATION OF AUTHORSHIP

I hereby certify that this PhD thesis, which I submit for examination to the Athlone Institute of Technology is entirely my own work and has not been taken from the work of others and to the extent that such work has been cited and acknowledged within the text of my work.

I confirm that:

- This work was done wholly or mainly while in candidature for a doctoral degree at this Institute.
- Where any part of this PhD thesis has been previously been submitted for a degree or any other qualification at the Athlone Institute of Technology or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the others, the source is always given.
- With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.

Signed: ______________________ Date: ___________
Conor Keighrey

ABSTRACT

Recent advances in technology have supported multimedia experiences to become more interactive and immersive. Traditional multimedia devices aimed to capture user attention through content rich visuals presented on two-dimensional displays. However, in recent times, emerging head mounted display (HMD) technologies such as Virtual reality (VR) and augmented reality (AR) HMDs aim to captivate users through the delivery of 360° visuals, immersive audio, and environments within which the user can interact. In this research, an immersive multimedia speech and language disorder assessment application was developed and novel user perceptual quality evaluations across three different platforms were undertaken. The speech and language disorder assessment application gave context to this research and directed it in terms of how to design an ecologically valid health application. The user perceptual quality evaluations were carried out on *tablet*, *AR*, and *VR* platforms under the auspices of the quality of experience (QoE) framework.

In this context, the novel contribution of the PhD work presented in this thesis reflects efforts to design, develop, and understand user QoE of an immersive multimedia speech and language disorder assessment application. The research involved a comprehensive and rigorous comparison of three different platforms (AR, Tablet, and VR) by exploring the use of: explicit (subjective ratings); implicit (e.g. physiological, and psychophysiological); and objective measures of user performance and interaction. The comparison required a novel QoE assessment methodology and evaluation which facilitated not only comparison between the different platforms but also analysis of various captured modes of user responses (objective and implicit) for each platform. The findings from this first multi-platform QoE evaluation led this PhD work towards the need to understand physiological measures at a deeper level. More specifically, the next novel study reports the efforts undertaken to understand task-evoked physiological response within the immersive virtual speech and language disorder assessment application in VR. A correlation is discovered between implicit measures of Electrodermal Activity (EDA) and pupillary response. This finding is reported within the context of data derived from the monitoring of objective (interaction) metrics throughout the immersive experience. Finally, whilst a holistic approach to understanding user QoE is crucial, it also brings challenges with respect to the amount of data and different modes of explicit, implicit, and objective data captured. Processing this data through traditional techniques is very time consuming and challenging. In this context, the work places an emphasis on the automatic data processing and classification of user emotional states during an experiment as an insightful measure of QoE.

ACKNOWLEDGEMENTS

Acknowledgements can be difficult; but over the past four years I have been surrounded by an amazing group of people who have supported my academic journey. To begin, I would like to take the first opportunity to thank my supervisor Dr. Niall Murray. I have been lucky enough to have Niall support my research and career development not only for the duration my PhD, but for my undergraduate work too. The extraordinary level of encouragement, guidance and support he has provided over the last few years has been an inspiration. His passion for academic excellence has been encouraging as I pursued my own journey in academy, I was delighted to have his support as a supervisor and friend during my PhD.

Second to this, I would like to take the opportunity to thank my secondary supervisor, Dr. Ronan Flynn. His expertise guided my early attempts at research during my undergraduate degree and postgraduate research. Over the past four years Ronan has provided fantastic personal, academic, and career advice and I would like to express my gratitude for this.

I'd also like to express my gratitude to my external examiner's Dr. Andrew Hines, and Professor Andrew Perkis. I was lucky enough to meet both early on in my research career, and I was delighted to have two key members of the QoE research community examine my thesis. Their words of encouragement have inspired me to continue working within the research domain. In addition, I would also like to mention and thank Professor Patrick LeCallet who examined my PhD transfer thesis. His inspiring words pushed me to achieve what my PhD became.

Lastly, from an academic perspective, I'd like to extend my gratitude to the faculty members in AIT who have been of huge support over the last few years. Both the Head of Department Dr. Enda Fallon, and Dean of Engineering Dr. Sean Lyons provided an environment for my career to develop to where it is today, and I look forward to my future in AIT. In addition, I would like to thank the Irish Research Council for funding my research over the last four years.

I would also like to also take this opportunity to thank my family and friends who have supported me since the start of my academic journey. First, I would like to express my gratitude to my parents Jack, and Myra Keighrey, who supported me financially for my undergraduate degree. In parallel, I would like to extend this thanks to my four sisters, Dela, Raph, Jillian, and Jacqueline for sitting through family events as I discussed how immersive technologies will change the future.

Over the past four years I have been blessed to be surrounded by a supportive group of researchers. I would like to extend my gratitude to everyone who was part of my academic journey in the AIT. First, I would like to thank Emer Coughlan for her amazing support, the long “therapy” chats in the late evenings while writing papers, and for organising unforgettable nights out. Following this, I would like to thank Debora Pereira Salgado. From the day she joined the lab, Debora has helped bring a group of computer science geeks closer together, making an effort to celebrate academic milestones.

I have made many journeys over the four years attending conferences in Europe, North America, and Asia. Despite the fantastic locations, nothing will be more memorable than visiting the little cottage in Montbron, France. The location caused panic for us all, as we arrived at a village which forced us to disconnect from the digital world we were so accustomed to and experience the great outdoors in a relaxing atmosphere. For this, I would like to thank Katja Osterwald for organising what I can only describe as an unforgettable break.

Last, but not least I would like to thank Thiago Braga Rodrigues for his phenomenal support especially in the last year of my PhD. Thiago and I both began out PhD journeys together in 2016. We have shared some amazing memories together, driving the speed of light on the Autobahn from Frankfurt to attend Oktoberfest in Munich, roof top parties in La Rochelle, France, and one to many pina coladas in Amherst. Reflecting on these memories, I look forward to making more over the next few years.

A PhD is often only considered to only be difficult and challenging academic journey, but for me, it was far from it. My PhD was an opportunity for personal development, in addition it provided a platform to travel the world and connect with an international community who have a passion for science. Of most things a PhD provides opportunity to meet new people and share experiences. So, with that said, I would like to end the acknowledgements with a quote for everyone who has helped me over the past four years.

> *“Here is to the night’s we can’t remember; with the people we won’t forget...”*

Table of Contents

TABLE OF FIGURES

TABLE OF TABLES

Chapter 1: INTRODUCTION

In this Chapter, the research question is introduced, and in addition the key contributions are presented alongside an overview of publications.

1.1 Background

In recent years, there has been an emergence of new multimedia experiences. Touch screen devices such as tablets, mobile (cellular) phones, and smart devices now deliver interactive multimedia experiences on the go. As such, the development of non-linear content which enables control and interaction has become one of the key efforts of the multimedia industry, thus facilitating a greater level of engagement with a user.

In parallel to this, there is increasing interest in head mounted display (HMD) technology. Virtual reality (VR) [1], augmented reality (AR) [2], and more recently mixed reality (MR) [3] HMDs have gained traction as immersive mediums [4] for entertainment. While entertainment is a key driver for the development of multimedia content and technologies, recent works have highlighted the application of immersive multimedia technologies within other domains. Examples of these include, the use of tablet technology in learning environments such as schools [5], and HMDs as a delivery method for training exercises within the retail [6], and manufacturing [7]. In parallel, cultural and creative industries are utilizing these novel technologies to visualise art and cultural heritage [8] [9] in new and innovative ways, captivating users through immersive story telling [10]. Existing applications of immersive multimedia technologies in manufacturing industries have shown the benefit of working with such novel systems. Examples of this include companies such as Audi [7] who have developed safe training applications in an effort to train employees who may engage with hazardous working scenarios. As new approaches to manufacturing embrace automation, the future of immersive technologies may play a fundamental role in the fourth industrial revolution (industry 4.0) [11] and beyond. These applications have highlighted how multimedia experiences can be used not only to entertain, but to enhance educational and training methodologies across multiple domains. There are many other domains which may benefit through embracing immersive multimedia technologies.

One such domain within health, is the clinical field of speech and language therapy (SLT). The ability to communicate is a fundamental skill in life. Deficiency in communication or language skills can inhibit an individual's academic, social, and emotional development [12]. It is estimated that speech and language disorders effect up to 20% of people at some stage in their lives [13].

Professionals who evaluate, diagnose, and treat individuals who may have difficulty with speech and language are known as speech language pathologists (SLP) [14] [15]. The field of speech and language pathology is often misconstrued as having a focus on assisting those with pronunciation or fluency problems with respect to language production. However, the scope of speech language pathology is much broader with SLPs treating a wide array of disorders which include but are not limited to vocal; swallowing; verbal communication; and cognitive communication disorders [15].

Due to the complex nature of speech and language disorders, SLPs must employ a wide array of assessment strategies to successfully diagnose and treat disorders. There are many assessment strategies which exist, and as such are a fundamental part of speech and language evaluation process within the clinic. These assessment strategies are typically carried out using paper based medium. As a result, the measure of user performance is somewhat limited. They employ binary scoring mechanisms (correct vs incorrect) as a mechanism to reflect a person's speech and language capabilities. Some existing assessment strategies acknowledge the need for and attempt to employ precision-based measures through the capture of assessment performance within the time domain [16]. However, these metrics are often monitored by a stopwatch at the subjective discretion of the SLP. This type of approach facilitates a limited range of accuracy with respect to the overall performance.

Through digitalization and exploration of new and immersive multimedia technologies, there is significant scope to capture of a range of measures which can provide valuable information on user performance. Although training and learning environments can be enhanced using immersive multimedia technologies [17] [18], critical to the success of any application, is the understanding of user centric requirements which influence quality perception.

In recent years, quality of experience (QoE) has become an accepted framework in industry and academia to understand the user perception of multimedia. While traditionally concerned with the optimization of computer networks [19] [20] to satisfy a user, it has expanded to other areas of research. QoE evaluations have been carried out in a wide array of domains such as multimedia, gaming, interactive experiences, and medicine [21] [22] [23] [24] [25]. Research into QoE aims to answer key questions about products, systems, and services which are targeted towards the consumer market. As has been reported [26], user QoE can be influenced by many factors, including technical, social, and psychological influences. Traditionally, user experience has been captured using surveys (post-experience). However, more recent approaches explore the utility of objective measures and implicit measures of physiological response. Thus, providing invaluable insight into user emotion [27] throughout a multimedia experience.

1.2 Research Question and Contributions

1.2.1 *Research Question*

The primary research question is:

> *"Can immersive multimedia technologies (tablets or head mounted displays) be employed as part of a Speech and Language assessment and if so, what is the user Quality of Experience of such an application?"*

Through answering this question, a series of sub-questions are answered:

- **SubRQ1:** Can immersive multimedia systems invoke a similar level of performance when compared to conventional paper-based semantic memory assessments?
- **SubRQ2:** Do immersive multimedia experiences result in a higher degree of QoE when compared to that of a traditional 2D multimedia experience?
- **SubRQ3:** Can objective measures of user performance support a deeper understanding of user and physiological response in the context of a QoE evaluation of an interactive and immersive multimedia speech and language therapy assessment application?
- **SubRQ4:** Can precise measures of physiological response be used as a reliable indicator of emotion and as such provide a better understanding of user response within the context of a QoE evaluation of an immersive virtual reality experience?

1.2.2 *Aims and Objectives*

The aim of this research is multi-fold. First, the primary aim is to explore the usability and utility of immersive multimedia technologies as a delivery method for an assessment strategy within the domain of SLT. To accomplish this, an experimental evaluation must be carried out. As such, this work explores and applies the holistic concept of QoE within the context of a user's perceptual quality evaluation of an immersive SLT assessment.

Employing experimental methodologies which have gained traction within the QoE domain, traditional explicit measures will be gathered to understand, and quantify user perceived quality. Furthermore, implicit (physiological and psychophysiological) and objective measures of user interaction are captured and evaluated. As such, a key aim of this work is to explore, process and understand the benefit of these new metrics in an effort to understand human emotional response and user QoE throughout a multimedia experience. Such novel measures may provide additional insight into user QoE. In parallel, they may also be used in the future for SLT assessments highlighting the potential benefit of digitization of clinical assessments. SLT assessments which aim to evaluate receptive language capabilities provide an ideal use case as they employ visual stimuli to challenge cognitive processing of language. In this context, capturing of implicit responses (e.g. physiological and psychophysiological) and understanding of same, may provide the clinician with additional valid context based information on the user’s performance and behaviour. This can potentially become part of a powerful new assessment tool within the domain of SLT.

To achieve these aims, the following research objectives are presented:

- First, a state-of-the-art review of QoE concepts, QoE evaluation in the context of immersive multimedia applications, and the use of technology within the SLT domain is carried out.
- Identify immersive multimedia technologies which have the potential to serve as a catalyst for the delivery of a virtual speech and language assessment.
- Explore SLT assessment strategies and identify a test which can be digitalized for the purpose of evaluation.
- Develop an immersive multimedia SLT assessment.
- Construct a novel experimental protocol for the purpose of evaluating QoE in immersive multimedia applications.
- Carry out a QoE evaluation exploring the applicability of multiple immersive multimedia technologies towards SLT assessment.
- Design and develop a protocol to process and extract understanding from various implicit metrics as an informative insight into user QoE.
- Appraise the use of physiological and psychophysiological response as an insightful measure of QoE within the context of an immersive multimedia assessment for SLT.

- Explore the potential for delivering more insightful measures of user performance to the clinician without altering the widely accepted assessment strategies within the domain of SLT.
- Disseminate the experimental findings through publications and presentation at academic conferences.

1.2.3 *Research Contributions*

The primary contributions of this work are the development of an immersive SLT assessment and the novel QoE evaluation of same. The following contributions reflect the impact of this work:

(i) From a QoE methodology perspective, a novel phased-based approach to a QoE methodology was designed to capture, extract, and understand objective, explicit, and implicit measures. The holistic approach can be utilized to derive an understanding of physiological responses associated with user independent interactions. Therefore, delivering fundamental insight into user perception of QoE. Three immersive applications are evaluated (see (ii)) utilizing the approach.

(ii) A cross platform immersive speech and language assessment was developed and evaluated across three multimedia technologies, namely: augmented reality, tablet, and virtual reality. This novel and extensive QoE comparison of the different delivery platforms considered objective performance metrics, explicit, and implicit QoE metrics.

(iii) Extending the QoE methodology applied in (i), a novel oculomotor processing framework was developed for the rapid reporting of user visual attention within an immersive virtual reality experience. Thus, allowing for the extraction of deeper, task-related physiological responses associated with changes in emotional arousal, and thus QoE.

1.2.4 *Dissemination of Work*

The following publications have resulted directly from this work to date:

1) **Poster Presentation** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray, and Niall Murray. "A QoE evaluation of immersive multimedia technologies for speech & language assessment applications." at Daughters of Charity Technology and Research into Disability (DOCTRID) conference, Limerick, Ireland, 2017.
2) **Paper** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray, and Niall Murray. "A QoE evaluation of immersive augmented and virtual reality speech & language

assessment applications." In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pp. 1-6. IEEE, 2017 – **(Best Student Paper)**

3) **Paper** - ***Conor Keighrey***, Ronan Flynn, Sean Brennan, Siobhan Murray, and Niall Murray. "Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR Speech and Language Therapy Applications" Proceedings of the on Thematic Workshops of ACM Multimedia (ACMMM) 2017, pp. 485-492. ACM, 2017.
4) **Journal** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray and Niall Murray. " A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications" IEEE Transactions on Multimedia (IEEE T-MM), 2020.

Additional contributions to the field of QoE are as follows:

1) **Journal** - Niall Murray, Yuansong Qiao, ***Conor Keighrey***, Darragh Egan, Débora Pereira Salgado, Gabriel Miro Muntean, Christian Timmerer, Oluwakemi A Ademoye, Gheorghita Ghinea, Brian Lee. "Evaluating QoE of Immersive Multisensory Experiences". In *IEEE MMTC Communications Frontiers, Special Issue on "QoE Evaluation and Control in Immersive Multi-modal Multimedia* Applications*"* (vol. 13, no.1), January 2018.
2) **Paper** - Darragh Egan, ***Conor Keighrey***, Sean Brennan, John Barrett, Yuansong Qiao, Christian Timmerer and Niall Murray. "Subjective Evaluation of an Olfaction Enhanced immersive Virtual Reality Environment" In 25th ACM International Conference on Multimedia (ACM MM 2017), Alternate Realities Workshop (Short Paper), Oct 2017, in Mountain View, CA.
3) **Paper** - Débora Pereira Salgado, Felipe Roque Martins, Thiago Braga Rodrigues, ***Conor Keighrey***, Ronan Flynn, Eduardo Lázaro Martins Naves, and Niall Murray. "A QoE assessment method based on EDA, heart rate and EEG of a virtual reality assistive technology system." In Proceedings of the 9th ACM Multimedia Systems Conference, pp. 517-520. ACM, 2018.

1.3 Thesis Outline

To this point, Chapter 1 has introduced the research motivation, scope, outputs (to date) and contributions of this PhD work. Chapters 2, 3, and 4 have been formed as a partitioned literature review of the state-of-the-art, each of which have their own focus. The purpose of this approach is to introduce key components of each of the areas of work incrementally. Chapter 2 introduces the fundamental concepts of the research topics. This is then expanded upon later in Chapter 3 which explores the methodologies used for each area i.e.

QoE and SLT assessment. As such, it highlights the existing approaches to QoE evaluation in addition to describing different assessment strategies used within a speech language therapy context (focusing on those relevant to this body of work only). To compliment this, Chapter 4 then builds on the knowledge introduced in Chapter 2 and 3 by exploring the current body of work which exists in these domains. More specifically, in Chapter 4 the fundamental works which have informed this research are highlighted and critiqued.

An overview of the different experimental efforts undertaken during this PhD are presented in Chapters 5 and 6. Chapter 5 includes a detailed description of a new QoE methodology and results of the comparison between the AR, VR and Tablet. It highlights the novel data capture, extraction, and interpretation of results. Following this, Chapter 6 introduces the second experimental study which is a deeper evaluation of the findings from the VR group, exploring the use of task-evoked physiological measures as an informative overview of QoE. Lastly,

Chapter 7 concludes the thesis, discussing the key findings related to the aforementioned research questions, acknowledges limitations of the work, and also proposes multiple avenues of future work.

Chapter 2: BACKGROUND RESEARCH

In this Chapter, the author presents a contextual overview of the fundamental topics of this research. Section 2.1 introduces speech and language pathology by outlining the complex nature of speech and language disorders. In section 2.2, a state-of-the-art review of the current generation of multimedia devices is presented. Lastly, in section 2.3, the author explores the evolution of QoE, discussing the widely accepted definitions and associated factors of influence with respect to the perception of quality of multimedia experiences.

2.1 Speech Language Pathology

2.1.1 *Speech Disorders and Language Disorders*

A communication disorder which disrupts the ability to communicate can be defined as a “speech disorder” or “speech impediment” [28]. Those who suffer from *speech disorders* have difficulty in articulation, fluency, or sound production. Articulatory disorders include the omission, distortion, or substitution of words. For example, early learners often substitute words for those which may sounds similar (e.g. children often say “wabbit” when they mean “rabbit”). Fluency disorders can be characterized by repetition, or hesitation in speech. A common example of a fluency disorder is stuttering. Stuttering is the abnormal production of speech through repetition (e.g. th-th-this). Similarly, stuttering is also associated with the prolongation of words (e.g. DDDDEEESSscribe the assignment…).

Speech and language impairments which result in difficulties understanding or expressing thoughts through language fall under the sub-category of *language disorders*. Language disorders can be developmental [29] or acquired [30]. Developmental language disorders are those which originate in childhood. Alternative to this are language disorders which are acquired later in life. Acquired language disorders can affect people of all ages, they often occur as the result of traumatic brain injury (e.g. stroke or car accident) [31].

2.1.2 *Expressive Language and Receptive Language*

Maintained and developed by the World Health Organization (WHO), the International Statistical Classification of Diseases and Related Health Problems (ICD-10) [32] document classifies language disorders which inhibit speech and language development into two independent subgroups: expressive language disorders and receptive language disorders.

According to [32], *expressive language disorders* are defined as:

> *"A specific developmental disorder in which the child's ability to use expressive spoken language is markedly below the appropriate level for its mental age, but in which language comprehension is within normal limits. There may or may not be abnormalities in articulation."*

While the focus within this definition is on spoken word, expressive language can be defined in a broader scope. Speech is not limited only to articulation, alternative production of language such as writing may also be recognised. Some of the common symptoms associated with expressive language difficulties in children [33] are: putting words into sentences, naming objects, and learning songs or rhymes. In short, the term expressive language disorder encompasses the inabilities with respect to the output of language.

The ability to express ourselves through the output of communication is a key component in our speech and language capabilities. However, to express ourselves we must first develop the ability to comprehend language. Receptive language describes a subgroup of language disorders which focus on evaluating the ability to understand, store, and retrieve language. The WHO [32] defines receptive language disorders within the ICD-10 as:

> *"A specific developmental disorder in which the child's understanding of language is below the appropriate level for its mental age. In virtually all cases expressive language will also be markedly affected and abnormalities in word-sound production are common."*

Symptoms of receptive language disorders [33] can include an impairment in comprehension. This can be seen in those who have trouble understanding language in written or spoken form; experience difficulty answering questions; and understanding gestures. Understanding is a key component in developing a semantic network of language within the brain [34]. Often those who experience receptive language difficulties encounter word retrieval problems. This can be described as an individual impairment with respect to the access or retrieval of associative language concepts.

Early intervention for speech language impairments is critical. Studies have shown that problems associated with speech and language impairment can often contribute to behaviour difficulties or escalated levels of anxiety later in life. In [35], a study which explored behavioural difficulties at a clinical level in children who attend special speech and language impairment classes (also known as language units) is described. The study revealed that children who exhibit symptoms of a mixed language disorder (expressive and receptive language) show a greater increase in behavioural scores. These scores are utilized within psychology to understand behavioural difficulties, reduced social skills, and emotional function. In addition, those who experience complex receptive language disorders were likely to score beyond the clinical threshold for a behaviour difficulty assessment. These types of studies highlight the long-term effects of speech and language disorders and requirement for early intervention in children.

2.1.3 *Aphasia*

It’s important to recognise that the majority of those who experience a speech and language difficulty can encounter symptoms of both expressive and receptive disorders. While focused disorders do exist, often both language subgroups have influence on a speech and language disorder. In addition, there are other influential factors which must be considered. For example, those who suffer from aphasia can experience complex difficulties with expressive and receptive language because of damage to the cognitive cortical or subcortical structures within the brain [30]. As a result, SLPs often explore and evaluate cognitive processing capability with an emphasis placed upon expressive and receptive language processing.

Aphasia can occur as a result of stroke or traumatic brain injury and can affect quality of life [36] for all ages [37] [38]. Symptoms of aphasia are often described under dichotomies such as receptive and expressive, fluent and non-fluent, or anterior and posterior [37]. However, aphasia is a complex disorder which is varied by nature, as such symptoms often do not fit within the limits of these paired descriptors. To compensate for this limitation, a series of aphasia classification systems have been developed to assist in the correct diagnosis. Each of these respective classification systems place varied weight on the influence of expressive or receptive language capabilities. Although there are numerous attempts to model aphasia, only two systems have gained clinical acceptance, namely the Bostonian classification system, and Lurian classification system [37].

Alexander Luria developed the Lurian classification system in Moscow in the 1940’s. Initial publications had been in Russian and an English translation was not available until the 1960’s [39]. As such, the Lurian systems dominance has remained central to eastern European countries. Alternative to this, and more widely accepted within western Europe and the United States is the Bostonian classification system [39] of aphasia. Originating in the Boston Veterans Affairs (V.A.) Hospital in Massachusetts, the Bostonian system classifies 8 accepted forms of aphasia [37]. Each of the respective classifications, are categorized through the identification of mild to severe difficulties associated with speech repetition, comprehension, and fluency. Figure 2.1 (adapted from [40]) presents an overview of the 8 subcategories of aphasia as per the Bostonian classification system. At a glance, it exposes the complex nature of diagnosis. Each of these classifications requires their own independent treatment strategies. Thus, demonstrating the high level of importance with respect to accuracy of diagnosis within this domain.

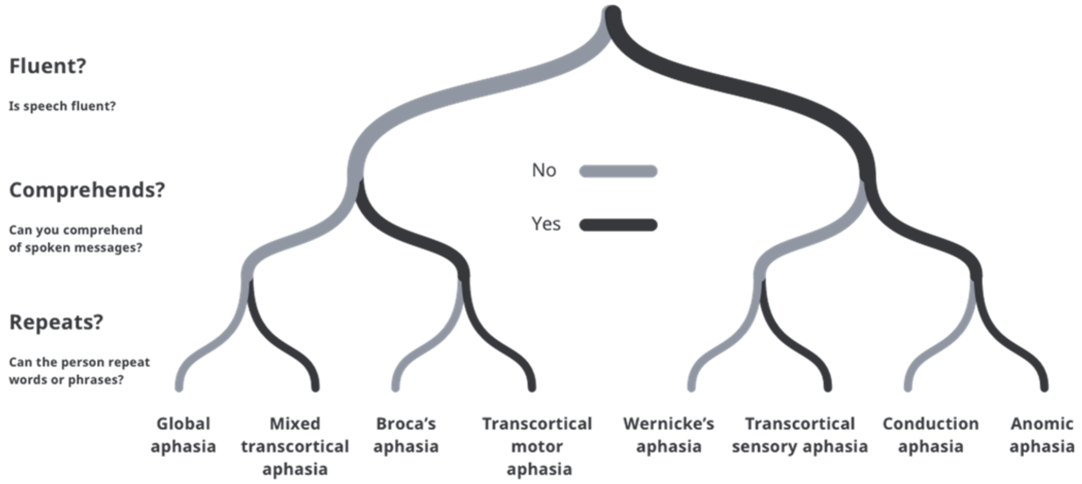


Figure 2.1: Bostonian Classification of Aphasia

2.2 Multimedia

2.2.1 *Media*

The field of multimedia is concerned with the presentation of multiple media sources. Media sources such as text, images, audio, and video are often combined to provide a multimedia experience to a user. Advances in computer technologies in terms of computing, storage, and networking has allowed multimedia to become an interactive and social experience. Compact devices such as smart phones allow users to capture and share media instantly. In recent times, technology has advanced to make multimedia more immersive and which alter how we interact with multimedia content.

2.2.2 *Immersive Multimedia*

Visual content is no longer limited to presentation on 2D displays. Head mounted displays (HMD) allow users to experience 360° visuals. HMDs consist of a display technology positioned within the user's field of view (FOV). An array of sensors (both internal and external) located around the device monitor the wearers head movement. Novel tracking algorithms in combination with sensor processing provide opportunities for users to move within virtual environments. However, the degree of freedom (DOF) (Figure 2.2) in which a user can move is limited by the complexity of the system.

Entry level HMDs are often restricted to 3-DOF (Figure 2.2 (left)). This allows users to explore virtual environments from a fixed position within a virtual space. As such, only head movement is tracked by internal sensors and translated into the virtual environment. Alternative to this, are systems which employ a more complex configuration which usually consists of both internal and external sensors. This allows users to experience 6-DOF (Figure 2.2 (right)) in which head movement and body position is tracked within a real world space and translated into the virtual world. These more advanced setups have now evolved to a stage upon which tracking is not only limited to HMD, but to accessories such as controllers. The evolution of these technologies has paved the way for three unique multimedia experiences, namely: virtual reality (VR), mixed reality (MR), and augmented reality (AR).

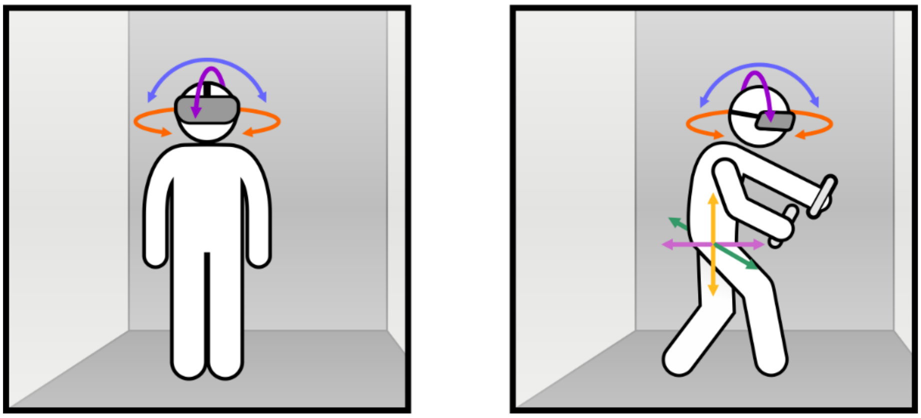


Figure 2.2: Head Mounted Display – Degrees of Freedom

(Left) 3 Degrees of freedom, (right) 6 Degrees of freedom

Typically, a VR system consists of an LCD, LED, or OLED display within the wearers field of view (FOV). This removes their ability to see their real-world surroundings. Mobile ready VR experiences such as the Samsung Gear VR [41] allow consumers to user their mobile as a hardware component to experience virtual worlds. However, these types of experiences are restricted within the boundaries of 3-DOF. More advanced systems are available for consumers, but they come at a cost. The HTC Vive [1], Oculus Rift [42], and PlayStation VR [43] deliver 6-DOF by using external hardware not only for tracking, but for computational and graphical processing too.

VR content centres on the presentation of fully simulated virtual environments. These environments are developed to engage the user, transporting them to an artificially developed fantasy or emulated real-world experience. With external controllers [1] [44], it is possible to interact with virtual content within an immersive world. This is typically restricted to systems which are capable of 6-DOF. Beyond this, technology does exist [45] which allows for a natural level of interaction without the implementation of controllers. These novel sensor technologies bring users hand movement directly into the virtual environments (Figure 2.3, adapted from [46]), thus advancing immersive multimedia experiences further.

Figure 2.3: Leap Motion Controller

More recent advances in HMD tracking technology has brought about the debut of what has been marketed as a mixed reality (MR) experience. Launched under this premise, is the HP Mixed Reality [3], and Acer Mixed Reality [47] HMDs. Despite consumer expectations, these novel headsets do not deliver a “mixed” reality experience; instead, they present a user experience more in line with that of a VR HMD. However, a key difference in MR HMDs is that they implement a novel tracking mechanism known as inside-out tracking (Figure 2.4 (left)), whereas VR implements outside-in tracking (Figure 2.4 (right)). Inside-out tracking utilizes wide angle camera sensors within the HMD to monitor and track distinctive characteristics within the environment. This, in combination with onboard sensors delivers a 6-DOF experience to users without the need for external sensors. Although these advances are key to the future of HMD technology, given the current use of MR HMDs and available applications, the author believes the technology should be defined under the context of VR until the applications and HMDs mature further providing a true MR experience.

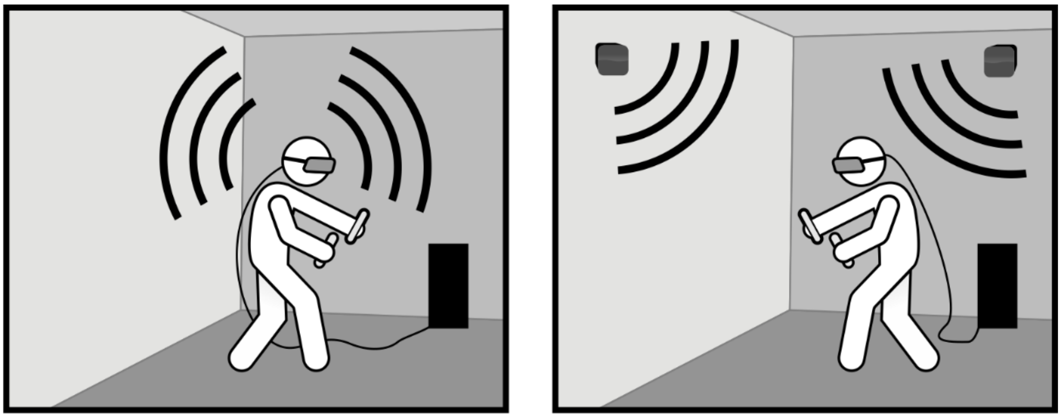


Figure 2.4: Virtual Reality Tracking Methodologies

(left) Inside-Out Tracking, (right) Outside-In Tracking

While VR and MR (in its current state) aim to captivate the user within an immersive virtual environment, alternative technologies exist which enhance and augment the user surroundings. Augmented Reality (AR) HMDs allow the wearer to perceive virtual content which is overlaid on the real-world environment. Entry level devices such as the Epson Moverio [48] or Vuzix Blade [49] deliver an experience in which content is augmented within the users FOV at a fixed position. Typically, applications designed for these entry level devices place an emphasis on delivering live data as communicated from a server or 3rd party device. For instance, a mobile phone can be connected to deliver a calendar update, or social media notification stream within the wearers FOV. Alternative to this are AR systems which represent the current state of the art within this domain. An example of such can be seen in the Microsoft HoloLens [2] and Magic Leap One [50]. As the technology has evolved, AR HMDs are now using inside-out tracking (Figure 2.4 (right)) to recognise the environment [2]. This novel approach allows virtual content to coexist alongside real world environments. Thus, creating a novel immersive multimedia enhanced world.

2.2.3 *Multiple Sensorial Media*

Current consumer level devices focus predominantly on presentation of aural and visual stimulus. Research and development in academia and industry are changing this as we see an ever-growing body of work [51] [52] [53] [54] in the area of sensory experiences or multiple-sensorial media (MulseMedia) [55] [56] [57]. Environments both digital (VR/AR) and real world (e.g. Theme park rides, or 4D Cinema) which stimulate more than aural and visual senses can be considered multisensory. Multisensory technologies now provide opportunity to stimulate haptic (sense of touch), gustatory (sense of taste), and olfactory (sense of smell) senses of a user [58]. The inclusion of multisensory components within a multimedia environment has the potential to create rich user experiences. These enhancements provide opportunity for developers to create immersive environments which draw similarities to real world experiences.

Applications of mulsemedia outside of entertainment can be seen with the implementation of multi-sensory components in education [51] [52]. Throughout early development, teachers and mentors alike engage multiple senses to assist in the learning and development process within children. Multisensory engagement often assists in the long-term retention of curriculum [59]. These approaches highlight the highly complex and individualized learning requirements for students. As we grow older, it’s often noted that people understand and retain information in individual ways. Those who prefer direct interaction and live demonstration are often considered more “hands-on”. This type of educational experience largely influences learning through the stimulation of our visual, aural, and tactile (haptic) senses.

It is well accepted, that the human brain has evolved to process the complex data gathered from our senses simultaneously. Although widely debated, Gardner’s theory of multiple intelligence [60] states that humans learn naturally if a learning experience is spread across multiple senses. Research, past and present has demonstrated this theorem through experimental evaluations [52] [57]. Within the context of immersive multimedia, it is expected that the addition of multisensory components has the potential to deliver virtual experiences which draw similarities to real world environments. Thus, there is potential to develop environments which deliver a similar level of learning and engagement which mimic educational settings.

2.3 Quality of Experience

2.3.1 *The Complex Paradigm*

The concept of QoE is a complex paradigm. To understand QoE, we must first develop an understanding of what “quality” is. According to Jekosch et al., quality can be best described as: “the judgment of the perceived composition of an entity with respect to its desired composition” [61]. This definition of quality has been widely accepted. It highlights the importance of understanding the process of perceptual reflection of an experience. As such, quality, in terms of QoE is described as a user centred measurement.

In 2010, a COST Action (IC1003) [62] was funded to develop a European (and International) network of QoE researchers known as Qualinet [63]. The scientific objective of the community was to explore the underlying concepts of QoE, in addition to developing methodologies for subjective and objective measures of quality. Collaborative efforts within Qualinet resulted in the publication of the Qualinet White Paper [64]. The document aims support those who wish to explore and understand the concept of QoE. In addition, authors worked to clarify and deliver a working definition of QoE. The Qualinet White Paper [64] defines QoE as:

> *"The degree of delight or annoyance of the user of an application or service. It results from the fulfilment of his or her expectations with respect to the utility and / or enjoyment of the application or service in the light of the user's personality and current state."*

Although widely accepted, more recent works have highlighted restrictions within this definition which prevents the concept of QoE being applied to a broader scope of multimedia experiences. Such limitations include the explicit description of "application or service", in addition an emphasis being placed on the term "user". In [65], Raake et al. proposed an updated definition which broadens the scope of QoE:

> *"The degree of delight or annoyance of a person whose experiencing involves an application, service, or system. It results from the person's evaluation of the fulfilment of his or her expectations and needs with respect to the utility and/or enjoyment in the light of the person's context, personality and current state."*

Through the inclusion of the word "system", the concept of QoE can be extended to allow the evaluation of new immersive multimedia technology as an experience (e.g. television, public address systems, or HMD). Additional recommendations highlight that not all media experiences have "users". For example, those who experience an enhanced audio experience within an auditorium are not users. Thus, the term "user" was replaced with "person". Finally, although the original definition places a focus on user emotions such as "delight" and "annoyance", there is a lack of fundamental acknowledgment of the "evaluation" process. As the concept of quality can be described as a perceptual experience, personal evaluation is a key component within the framework.

2.3.2 *Factors Influencing Quality of Experience*

The perception of quality within a multimedia experience is subjective. Factors which influence perception can occur due to the variant and invariant characteristics of a multimedia experience. However, perceptual influences go beyond the application, system, or service which is evaluated. User characteristics such as mood, socio-economical background, and environmental conditions are also to be considered. Identifying, and understanding these factors that influence the perception of quality is key to the success of a new multimedia experience. According to the Qualinet White Paper [64], an influencing factor is defined as:

> *"Any characteristic of a user, system, service, application, or context whose actual state or setting may have influence on the Quality of Experience for the user"*

Since QoE has emerged from the telecommunications industry, a significant amount of research works exist which place an emphasis on characteristics within the context of technical configuration (i.e. multimedia codecs, network transmission, etc.) [65]. However, QoE has matured and become a multidisciplinary field in which evaluations occur not only in the context of multimedia content delivery, but in novel multimedia technologies, platforms, and immersive experiences. As such, a broader scope of influencing factors must be considered.

In [23] and [66], the authors present a model of QoE (Figure 2.5) which extends the consideration of QoE beyond technical factors by considering the influence of human and context factors within the perception of a multimedia experience. This model, places the user at the centre of three primary influences of a multimedia experience, therefore providing core insight into user perception of quality. More specifically, within the context of interactive and immersive multimedia applications which place the user at the centre of the experience.

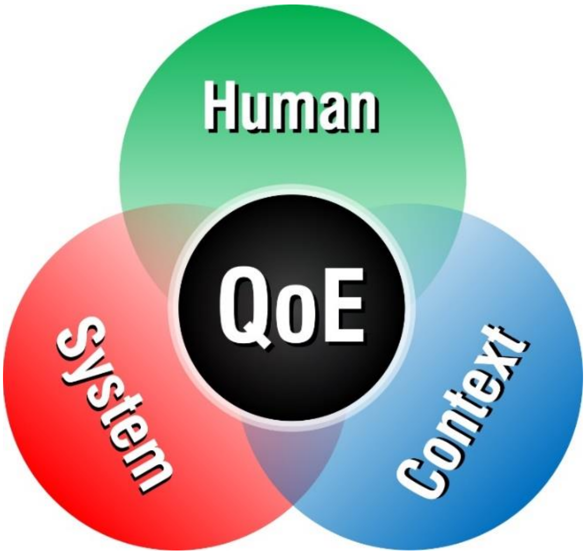


Figure 2.5: Influential Factors of Quality of Experience

2.3.2.1 *Human Influence*

The perception of quality is a subjective experience, as such the importance of human factors within a multimedia evaluation has never been greater. Human Influencing Factors (HIFs) describe the intrinsic properties and characteristics which are influenced directly from the users. HIFs are categorically described under processing levels (low-level processing, and high-level processing [66] [67]).

Low-level processing relates to the physical, mental, and emotional state of the user. As such, human influence can stem from an assessor’s visual [67] or auditory acuity, gender, or age. More in line with mental and emotional states, is the perception of an experience due to temporal components. Examples include, an assessor’s mood motivation, or attention throughout a multimedia experience. *High-level processing* (also known as top-down processing) is reflected by the human understanding at a cognitive level. There are several factors which cause human influence at a high level, they include: level of education; life stage; or socioeconomic circumstances. An emphasis is often placed upon understanding the exposure to knowledge prior to the multimedia experience.

2.3.2.2 *System Influence*

System related IF’s are perhaps the most controllable within a QoE evaluation. Properties or characteristics which describe system related IF’s can be subdivided into four groups: content, media, network, and device factors. Whilst some systems will contain a network element, others may not. Content and media related factors are interrelated and have traditionally described the controlled delivery of audio and visual stimulus. However, the key difference can be seen in that content places an emphasis on video (e.g. contrast, detail, and motion) and audio (e.g. dynamic range, bitrate, and compression). Whereas media focuses on the configuration of media types (e.g. codec, resolution, compression, and media synchronisation).

Network related factors are associated with the transmission of media over a communications network [65]. As multimedia has become a mobile experience, the impact of fundamental network characteristics has been shown to be very important. These include but are not limited to bandwidth, delay, jitter, loss and errors, and throughput [65]. Such measures have traditionally been the centre focus in QoS, however they are not to be ignored within a QoE perspective as media sharing and consumption has become a multi-user experience. Interconnected with content, media, and network related factors are the device related factors. A device, or end system, is limited by CPU, network, display and sound capabilities. Each of these respective characteristics can influence the perception of quality. CPU and network factors control the level of connectivity, while display resolution and sound capabilities have impact on content and media perception.

2.3.2.3 Context Influence

There are five factors which have contextual influence within an experience: physical, temporal, task, social, and economical [67]. *Physical* IF's are concerned with the characterises of location, and space. This can include the transition of environments (i.e. mobile multimedia experiences), sensory attributes (i.e. noise, or temperature), or location (i.e. personal, professional or public spaces).

A *temporal* factor is related with a time-based influence on quality perception. Examples of such include time of day or year, frequency of use, duration, and synchronisation (i.e. multisensory stimulus). From a *task* perspective, there are three situations which may impact the perception of quality: multitasking, interruptions, and task type. However, the influence of task can often be reduced through the delivery of training within an experimental protocol. Such approaches aim to alleviate any potential misunderstanding with respect to a task which may influence the perception of multi-modal interaction. Interpersonal relations which may or may not exist within an experience have the potential to impact user perception. Hence it is important to explore the influence of *social* factors within an evaluation.

Although related more with subscription services, economic factors are not to be overlooked. Evaluation of such influence may however be difficult to predict. In [68], authors explore the influence economic factors on the perception of video quality. Test participants could enhance the visual quality of their multimedia viewing experience through spending. The objective of the study was to explore the willingness to pay for quality, and content selection within a video quality perception test.

Chapter 3: METHODOLOGY

In Chapter 3, a review of the various assessment and evaluation methodologies used within the domains of SLT, multimedia and QoE is presented. Due to the nature of this work, an emphasis is placed on the presentation of speech and language assessment strategies which aim to evaluate a semantic memory assessment. These examples are selected to emphasise how the delivery of clinical assessments using multimedia technologies can enhance and further develop already existing speech and language evaluation strategies. In addition, in terms of multimedia quality evaluations, the use of various explicit, implicit, and objective data capture, processing, and analysis methodologies for QoE are presented.

3.1 Speech Language Therapy

3.1.1 *Paper Based Assessment*

SLPs employ a wide variety of evaluation methods to diagnose adults and children who experience speech and language difficulties. Historically, assessment stimuli have been presented on a paper based medium. These standardized assessments focus on the evaluation and diagnosis speech and language disorders through a collection of verbal and non-verbal testing strategies. Aggregated scores are used to reflect a person's abilities with respect to a specific region of speech and language.

An example of a paper-based assessment strategy is the Comprehensive Aphasia Test (CAT) [69]. The CAT (Figure 3.1 (adapted from [69])) is a speech and language test battery which aims to evaluate speech and language capabilities in adults and children who are experiencing symptoms of aphasia (section 2.1.3). To assess these difficulties, the CAT contains multiple assessments which have been developed to evaluate language cognition and production. 34 subtests are split into three key sections: (a) the cognitive screen has been constructed to identify deficits associated with cognitive processing of speech or language; (b) a language battery which consists of a series of assessments to evaluate receptive and expressive comprehension; and lastly, (c) a disability questionnaire which aims to identify any underlying disabilities. Results gathered through the completion of the CAT, assist clinicians in identification and understanding of patient's symptoms, thus providing an overall accurate diagnosis.

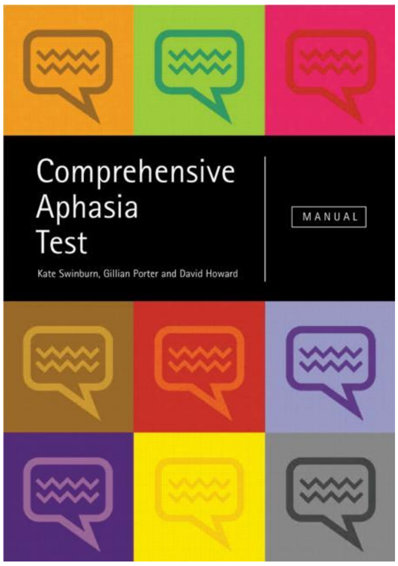


Figure 3.1: Comprehensive Aphasia Test – Paper Based Assessment

3.1.1.1 Semantic Links

Speech and language assessment banks are typically developed to diagnose a specific disorder. However, assessment methodologies often draw similarities in how they are presented and carried out. An example of this can be seen in assessment strategies which evaluate the ability to recognise or access pre-acquired knowledge. These are known as semantic link assessments [69] [70]. The fundamental concept of a semantic link assessment is to formulate an association with a central concept (presented as a word or image), with alternative stimuli. Depending on the assessment strategy, this is generally carried out using a non-verbal response (e.g. pointing action) from the user. As such, the purpose of the assessment primarily focuses on the evaluation of receptive language. Used in conjunction with other assessments, semantic links can be tailored to reveal a deficit associated with cognitive performance [69]. To illustrate similarities of semantic links, and the evaluation outcome, we contrast two semantic link assessments. The first of which is the Pyramids and Palm Trees (PPT) assessment [71].

The PPT was developed in the early 90's and focuses on a specific evaluation of receptive language capabilities. It is a comprehensive evaluation of semantic memory abilities. The focus is on an individual's strengths to formulate semantic (categorical) links in language. Assessment procedures outlined within the PPT require clients to identify links between presented stimuli (an example can be seen in Figure 3.2 (adapted from [71])). Stimuli consist of one category item located at the top, and two semantic relators located at the bottom. One of the semantic relators is considered the target while the other is known as a semantic distractor.

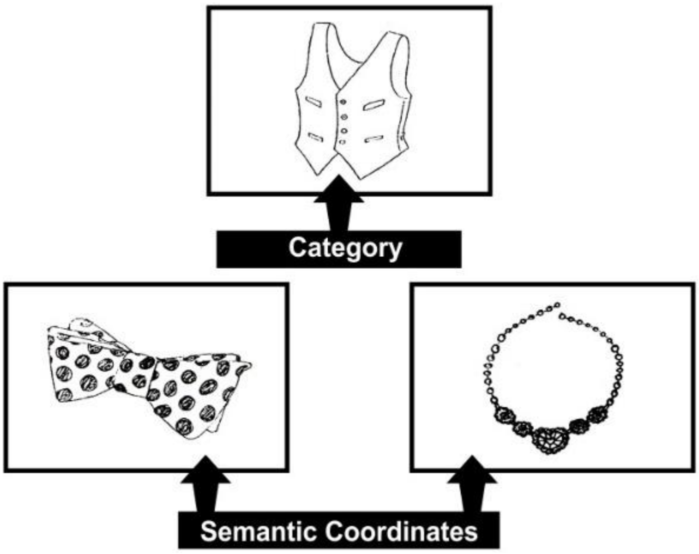


Figure 3.2: Pyramids and Palm Trees Assessment

An alternative to the PTT is the Peabody Picture Vocabulary Test (PPVT) [70]. The PPVT aims to measure receptive language capabilities in children and adults (2 – 90+ Years). As illustrated in Figure 3.3 (adapted from [70]), assessment stimuli consist of four images. A clinician reads an extract from the assessment procedure and a patient is asked to identify the image associated. For example, the clinician may read a statement as follows “the child is fast asleep”, in which the patient responds by pointing to option 3 in Figure 3.3. This differs from the PPT as clinicians use a call and response approach. As such, semantic processing is evaluated in parallel to receptive comprehension, whereas the PPT places more of an emphasis on the cognitive component of receptive language.

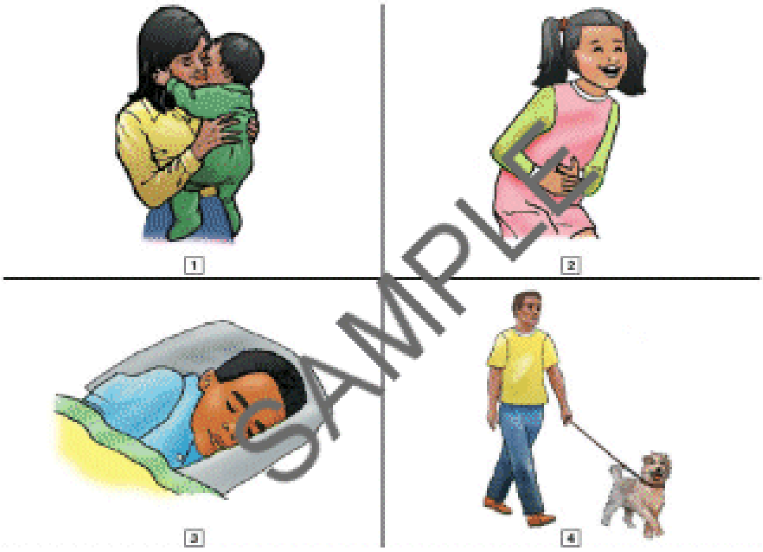


Figure 3.3: Peabody Picture Vocabulary Test – Example Stimuli

3.1.1.2 Visual Organisers

Vocabulary evaluation and development can be encouraged through interactive visual organisers (Figure 3.4) known as wordmaps (also referred to as mind maps, concept maps, or semantic maps). Wordmaps create, further develop, reinforce, and exercise semantic word knowledge (semantic links) within language. They are a multidimensional assessment in which both receptive and expressive language capabilities are evaluated. Semantic word knowledge can be described as the range of words which a person understands and actively communicates with. In short, semantic word knowledge can be thought of as, an individual’s mental dictionary which allows for more precise communication. SLPs place significant importance on semantic word knowledge as it assists in the articulation of spoken and written language.

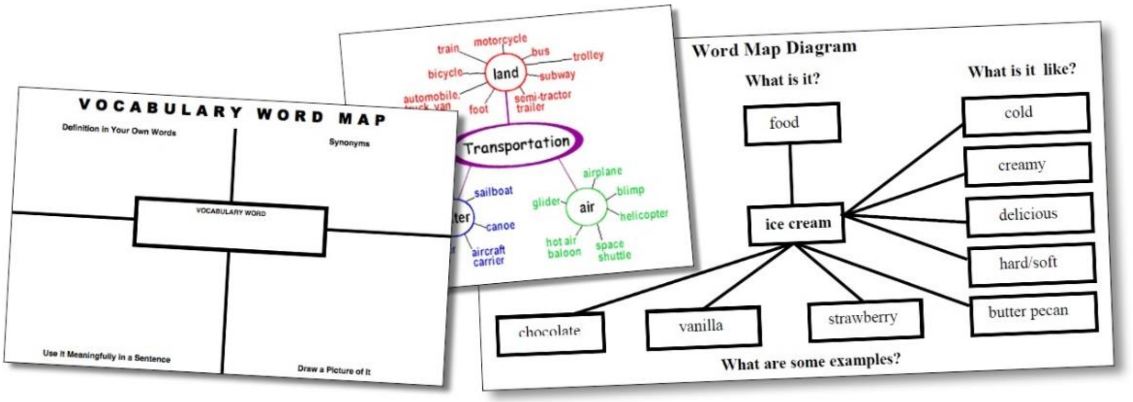


Figure 3.4: Visual Organisers

As illustrated in Figure 3.4, visual organisers can take many forms. However, wordmaps can generally be completed through expressive language communication means (written). Others often factor an alternative approach which requires a mix of words, sentences, or illustrations. Irrespective of the methodology, wordmaps all share one common goal. They assist in the understanding of new topics or complex ideas through visualisation of semantic or categorical relationships, thus building on previously acquired knowledge. According to [72], learning environments which stimulate previous knowledge in order to develop an understanding of a new topic are most efficient as they deepen knowledge and widen the span of already known semantic networks.

While wordmaps are predominantly used as learning activity, they can also be used as an evaluation mechanism in a clinical environment. SLPs can monitor and evaluate the long-term retention of knowledge through the assessment of completed exercises. These types of evaluations provide insight into the active lexicon available to person. Due to the complex nature and vast range of speech and language impairments, individualized assessments are critical to the success of treatment [15]. A single evaluation or treatment methodology will not suit all disorders and as such SLPs are trained to adapt treatment to suit the individual requirements of a client. While this adaptive freedom does exist a duty of care must remain, as variations to any treatment must always remain validated. Thus, a system which monitors and evaluates an individual user requirement has the potential to change these methodologies.

3.1.2 *Digital Assessment*

Building on the previous section and more closely aligned with the work presented in this thesis are digital applications which aim to assist in the clinical assessment of speech and language. Recent works highlight ongoing research and development of assessment strategies which aim to reformat already existing paper-based approaches and bring them into the digital age. Premium services such as Q-Interactive [73] by Pearson Clinical [74] have recently tailored their platform to deliver widely accepted speech and language assessments to clinicians through digital means. The service uses a combination of tablet and PC, as a multimedia platform for the delivery of digital assessment’s which were traditionally paper based. According to marketing claims, the use of multimedia technologies increases levels of engagement, more specifically in those who are born as digital natives (i.e. the digital 1st generation) [73]. However, online reports [75] (scientifically reviewed) claim to support the use of such technologies within the elderly population. In this report, expert clinician Dr. Loretta Bolyard reports an increased level of engagement from elderly clients. According to the report [75], faster interaction measures associated with multimedia form factors are a key influence in this outcome. As such, demonstrating that the application of multimedia technologies benefits all age ranges. Evolving pen and paper assessment strategies are beneficial in more ways than one. Digitalization of assessments pave the way for automated system reporting within the clinical field. Thus, alleviating any errors associated with the recording of data, and facilitating the delivery of informative performance tracking.

Attempts to create digital scoring mechanisms for alternative assessments exist. For example, [76] describes a novel scoring mechanism for the CAT. However, the scope of the “software” is limited to PC users with Microsoft Excel [77]. Although the software marketing promises the delivery of simplified scoring, automated calculation, and result interpretation it appears to be limited to the bounds of Microsoft Excel macros. Such tools are no doubt useful within a clinical setting, however they are restricted by the inclusion of only the clinician and fail to consider the needs of the user. In addition, they do not take advantage of many opportunities which can be delivered to users by digitalization of paper-based assessments.

Similarly, a manual [78] exists online which describes an application known as eCAT (Electronic Comprehensive Aphasia Test). According to the document, the application has been developed to run on a Windows PC and focuses on the delivery of a digitalized multimedia enhanced CAT. Assessment procedures and data capture remain untouched from the traditional paper based approach. Despite being accessible from Google, further searches of academic databases (Google Scholar, Research Gate etc.) yield no further documentation of the eCAT system. In addition, the manual states that the software is only available on PLORAS (a research group) laptops, and accessible by logging into a private computer network. As such it’s believed that this software has been developed for the purpose of internal testing. According to [79], the PLORAS project aims develop a prediction system which delivers a recovery score to those who are undergoing treatment for aphasia.

3.2 Quality of Experience

Key to the success of an application, system, or service is the understanding of factors which impact the user perception of quality. To identify factors which influence quality, we must explore and apply methodologies which gather insight into the perceptual recollection of a multimedia experience. In recent years, research within the multimedia community has predominantly focused on the capture of QoE through qualitative methods. Subjective (explicit) measures, in the form of surveys, or questionnaires are employed to gather insight into user perception of quality during an experience. Alternative to this, more recent works [21] [22] [80] employ a series of quantitative (implicit) measures. Thus, exploring the experimental capture of QoE through the monitoring of application performance, and capture of physiological measures. In the following sections, explicit and implicit data capture methodologies used in multimedia evaluations are presented.

3.2.1 *Explicit Measures*

Efforts to capture user perceived QoE have traditionally focused on the recording of user ratings or surveys post experience. Typically, user ratings are captured through self-reported questionnaires. Defined by the ITU-T, a series of recommendations exist [81] [82] which serve as a guide to subjective evaluation for experimental testing within the multimedia research community. Each respective methodology differs in terms of the presentation of multimedia stimuli (e.g. single stimulus, double stimulus, multi-stimulus, or pair comparison). In addition, the ITU-T acknowledge that there is no “one size fits all” approach to multimedia evaluation. As such, [82] outlines a series of acceptable changes which serve as a guide to modify subjective capture methods for the purpose of experimentation. Common changes often include minor adjustments to subjective rating labels on a per evaluation basis [83].

3.2.1.1 Absolute Category Rating

The Absolute Category Rating (ACR) is a well-established testing methodology used to capture subjective data from users [82]. Testing circumstances which meet two criteria can use ACR. The requirements are: (a) Subjects are restricted to the observation of a single stimulus prior to the evaluation, (b) Subjective questions must be answered through a categorical evaluation. Traditionally, the ACR method supports a five-level rating scale as outlined in [82]. Participants are often unaware of the numerical values associated with each rating, as this information is traditionally only useful to the researcher for performing statistical analysis. The single evaluation method provide opportunity for researchers to capture a large number of ratings within a short period of time [82].

Table 3.1: Absolute Category Rating – Five-Level Rating Scale

Rating:	**1**	**2**	**3**	**4**	**5**
Label:	Bad	Poor	Fair	Good	Excellent

3.2.1.2 Degradation Category Rating

An alternative method to ACR which focuses on the comparative analysis of two presented stimulus is the degradation category rating (DCR) system [82]. While stimulus is presented in pairs they are introduced via a predefined protocol. That is to say, a reference sample is presented first, following this the experimental stimulus is presented. This type of evaluation methodology is often used in the field of multimedia research in which new and innovative encoding methods for visual stimulus are being evaluated. Similar to the ACR system, participants are requested to rate the stimulus using a five-level scale, an example of which can be seen in Table 3.2. Given the focus is on comparative analysis, DCR provides the opportunity to evaluate multimedia content in contrast to a control stimulus.

Table 3.2: Degradation Category Rating – Five-Level Rating Scale

Rating:	**1**	**2**	**3**	**4**	**5**
Label:	Very annoying	Annoying	Slightly annoying	Perceptible but not annoying	Imperceptible

3.2.1.3 Comparison Category Rating

Similar to the DCR, Comparison Category Rating (CCR) [82] is a double stimulus method in which two stimuli are presented simultaneously [82]. However, the key difference in comparison to DCR is that participants are unaware of which stimulus (reference or processed) they are introduced to first. Unlike the ACR and DCR, CCR employs a seven-level rating system which ranges in values from -3 to 3 as outlined in Table 3.3. As participants are unaware of which stimuli is the reference or has been processed, CCR ratings encounter only minimal influence from a subject's opinion of content. Other influential factors which must be monitored is the randomization of stimuli, researchers must ensure that the randomization is fair (i.e. 50% of reference stimuli are presented followed by 50% of processed stimuli).

Table 3.3: Comparison Category Rating – Seven-Level Rating Scale

Rating:	**-3**	**-2**	**-1**	**0**	**1**	**2**	**3**
Label:	Much Worse	Worse	Slightly Worse	The Same	Slightly Better	Better	Much Better

3.2.1.4 Mean Opinion Score

Inspired from the field of telecommunications, Mean Opinion Score's (MOS) are used to quantify a single value which can represent the overall quality associated with a system, product, or service. The field of QoE adapts these ratings and extends its use to provide insight into user perception. User ratings are captured using the absolute category rating scale in which users report their experience on a scale of 1-5. Traditionally quality evaluations factor in both of these guidelines. However, they are not restricted to these absolute values. In recent studies we see MOS ratings evaluated on a scale of 1-7 or 1-9 [82], similarly the labels can be interchanged to represent a positive to negative scale pending the questionnaire content.

3.2.2 *Implicit Measures*

Whilst explicit measures remain a cornerstone within the evaluation of experimental multimedia experiences, they are not without their flaws [84]. Participants can often suffer from difficulty in communicating their reflection of a multimedia experience due to a limited range of responses. In addition, the capture of explicit measures can be time consuming, and difficult to interpret as a result of the non-uniform distribution of results [84]. As such, their perceptual evaluation may not be accurately reflected through questionnaires or survey responses.

Recent works highlight innovative approaches to gaining a deeper understanding of QoE through the recording and analysis of implicit measures. Examples of this include the capture of physiological measures such as heart rate [80], electrodermal activity [21], oculomotor response [85], or electroencephalography [86]. These novel implicit measures deliver insight into levels of emotional arousal [27], thus quantifying levels of enjoyment or stress throughout a multimedia experience. The experimental analysis and understanding of implicit measures have the potential to deliver a deeper understanding of human factors which affect the perception of quality within a multimedia experience. Beyond this, objective metrics are being explored as a supporting measure of QoE, examples of such are included in [22] which details the use of task completion time as a method to quantify user performance within an AR experience.

3.2.2.1 Physiological Measures

The human body consists of a complex network of systems which communicate using electrical signals [87]. Collections of systems work together to keep the body functioning. The digestive, respiratory, and circular systems have highly important roles in that they focus on keeping the body alive. Alternative to this are perceptual systems which work together allowing us to perceive the world around us. Each of these respective systems communicate and work in collaboration with each other through the sending and receiving of signals. For example, signals received from the perceptual systems can signal the digestive system to induce hunger [88]. Capture of these signals, or physiological measures as they are known, can provide insight into various human experiences [21] [27] [89].

QoE research has focused predominantly on the capture of physiological measures [27] from the central nervous system (CNS) and a subset of the peripheral nervous system (PNS) known as the autonomic nervous system (ANS) (Figure 3.5, adapted from [27]). Additionally, research has explored the monitoring of eye movement (Eye gaze, blinking, pupillometry [27]). These physiological signals captured from the ANS, CNS, and eyes can provide insight into the perceptual, cognitive, and regulatory functionality within the body. Analysis of this data can assist in the understanding of user perception in multimedia environments.

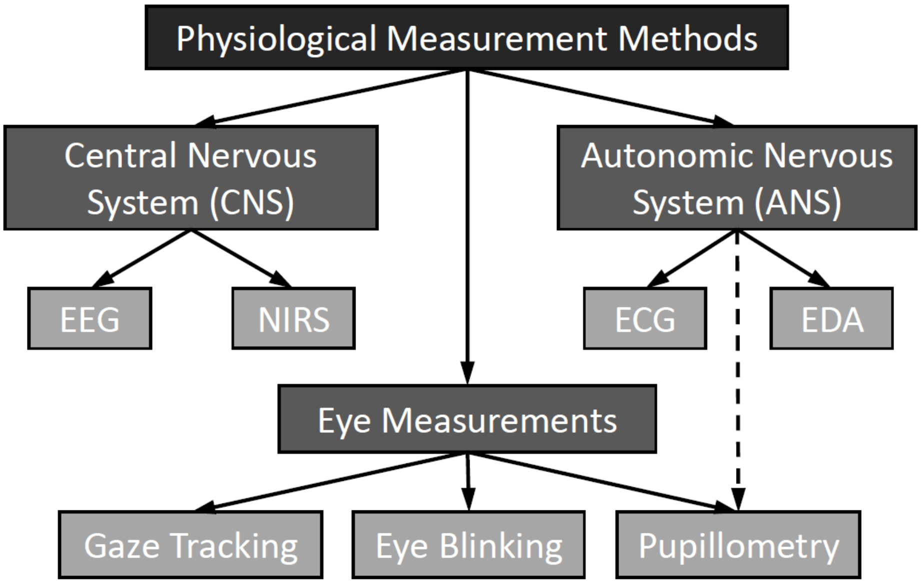


Figure 3.5: Physiological Measurement Methods

3.2.2.2 Central Nervous System

Consisting of the brain and spinal cord, the complex system which coordinates and controls, thoughts, and movements within the human body is known as the central nervous system (CNS) [27]. The system of pathways receives and interprets information from perceptual sources. These sources transmit electrical signals which can be captured to provide insight into the activity of the brain.

1) Electroencephalography

The brain is a highly complex organ in which an intricate network of electrical signals are transmitted and interpreted. These electrical signals can be observed using electroencephalography (EEG) [27]. Monitoring of EEG in a clinical environment typically occurs using an array of electrodes which are evenly distributed around a cap as per Figure 3.6 (left) (adapted from [90]). Less invasive devices are available and are more often used in multimedia research. An example of such a device is the Emotiv [90] EEG headset as presented in Figure 3.6 (right) (adapted from [90]). Abnormalities in EEG activity can provide insight into ill health, sleep disorders and epilepsy. Traditionally EEG has been used for research in the fields of cognitive science and psychology research [91]. However, recent works in the QoE [27] [86] [92] and multimedia communities have explored the application of EEG in quality evaluations.

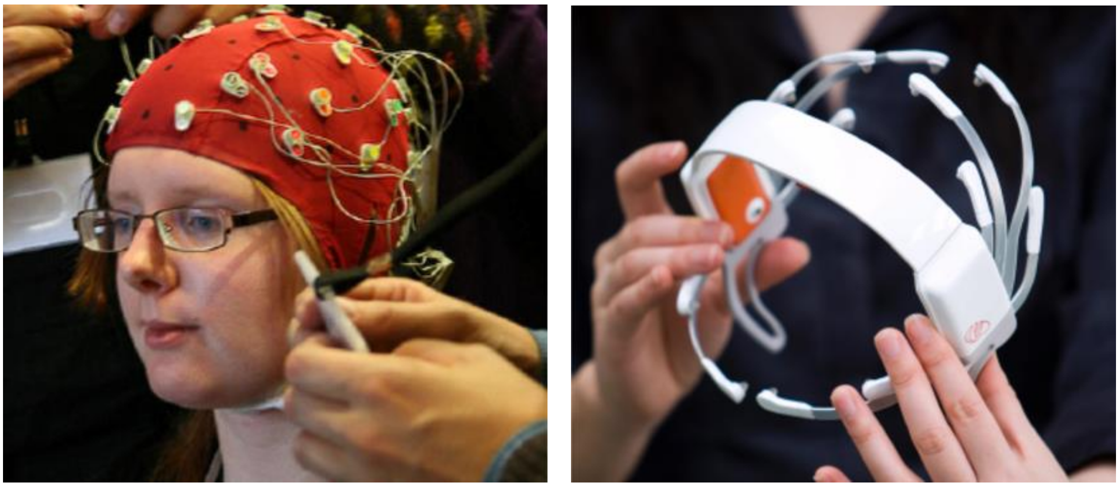


Figure 3.6: Electroencephalograph Systems

(left*) Complex EEG system which utilizes a cap – (right) Less invasive EEG headset*

3.2.2.3 Peripheral Nervous System

Compromising of multiple nerves and ganglia, the peripheral nervous system (PNS) serves as a communication relay between the CNS and the rest of the body [87]. The PNS is divided into two independent systems, the somatic nervous system and autonomic nervous system. The somatic nervous system transmits signals from the brain in order to control organs such as muscles, thus facilitating voluntary responses such as movement. There are numerous subdivisions within the somatic nervous system which each have their own individual functionality. One such division is the Sensory Nervous System (SenSys) [88]. The human experience is multisensorial. We experience the world around us through the stimulation of our visual, aural, olfactory, haptic, and gustatory senses. The SenSys in conjunction with the neural pathways of the brain is responsible for processing the stream of multisensory data. In short, the SenSys is a multisensory system which translates environmental signals captured though multiple senses allowing us to perceive the environment around us.

The second subdivision of the PNS is known as the autonomic nervous system (ANS) [87]. The ANS controls our automatic physiological responses. The primary mechanisms associated with this autonomic response are separated into two divisions. The parasympathetic nervous system (PSNS) division is often described as the “rest-and-digest” division. Its primary focus is to slow heart rate or stimulate activity within the stomach. More in line with research within the area of QoE is the sympathetic nervous system (SNS) of the PNS. The SNS is associated with the “fight-or-flight” responses within human physiology. Measures captured from the SNS include but are not limited to heart rate; blood volume pulse; skin conductance; and electromyography. Studies show [80] [93]

[89] that monitoring the autonomic signals generated from these areas of the body provide insight into human emotion. Given the lack of voluntary control, focus has been placed upon capturing these responses in research as they respond naturally to presented stimulus.

1) Heart Rate

Research, both past [21] and present [94] has explored the monitoring of heart rate (HR) as an implicit measure to gather insight into user emotional states. More specifically, the variability of heart rate when a user is experiencing a system or service can provide insight into the emotional state. For example, a rapid change in HR can be induced as a user experiences multimedia content which may excite or frighten them. These changes in HR occur as a result of the cardiovascular systems being controlled by the ANS. The ANS regulates HR by increasing activity as a result of a sympathetic response and decreasing HR due to a parasympathetic activity.

There are numerous methodologies which can be used to measure HR. Electrocardiography (ECG) is most commonly used in clinical settings and have been adapted to research studies due to its accuracy. The process of data capture involves the placement of electrodes on the user’s skin. Miniscule electrical charges which arise as a result of the heart muscle are captured for each individual heartbeat. Due to the invasive nature of the electrodes, the QoE community often avoid such technologies with a motivation to secure ecologically valid data on the experience.

Alternative, non-intrusive technologies exist to capture HR. In clinical settings, finger placed HR monitors are used to monitor a patient’s resting heart rate. Interestingly, this type of technology has been adapted by Fitbit [95], an activity (fitness) band manufacturer to capture, and monitor HR during intensive training routines. Fitbit use an optical sensor (known as PurePulse) to measure the wearers heart rate. The optical sensor uses plethysmography (PPG) as a method the capture HR [96]. PPG measures blood volume using a series of green LED’s which are located on the underside of an activity tracker or smart watch (Figure 3.7). Light emitted from the LED penetrates the skin and is absorbed by the blood. When a larger flow of blood is pumped through the vessels more light is absorbed and thus a change can be detected. PPG sensors monitor this light absorption rates to calculate the wearers HR.

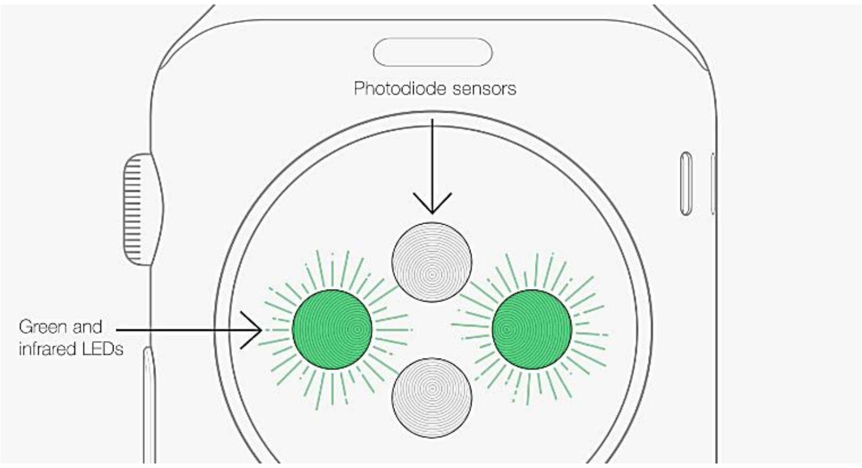


Figure 3.7: Plethysmography – Light Sensors Illustrated on a Smart Watch

Researchers in QoE often favour this method of capture, as wearing an activity band is considered much less invasive and supports ecologically valid data. In [21], and [80], authors describe the HR sensors in conjunction with other measures of the ANS to provide insight into immersive virtual experiences. In [80] authors reveal no statistically significant differences found within the results, a correlation is observed between elevated levels of HR and physiological arousal. Similarly, [21] discovers a correlation between feelings of frustration and elevated HR whilst low HR being indicative of enjoyment through low level challenges in a gaming experience.

2) Electrodermal Activity

Electrodermal activity (EDA) (also known as galvanic skin response) is the measure of physiological changes in skin conductivity. Skin conductivity can provide an indication of psychological arousal (i.e. cognitive activity) and physiological arousal (i.e. stress or excitement). Traditionally, EDA is monitored via the placement of two electrodes attached to the user’s fingers. A small electrical current is applied to the skin, after noise is filtered the skin conductance is captured and examined for changes to determine if any form of arousal (psychological or physiological) has occurred.

Measures of EDA can be subdivided into two distinct categories: tonic change and phasic change [97] [98]. Tonic change corresponds to a steady or slow change in skin conductance as outlined by the yellow signal in Figure 3.8 (adapted from [99]). Typically, tonic change is referred to as skin conductance level (SCL). Studies show that SCL can be used to provide insight into an increase in cognitive activity [97] [100] [101] or states of emotional arousal [27] [102].

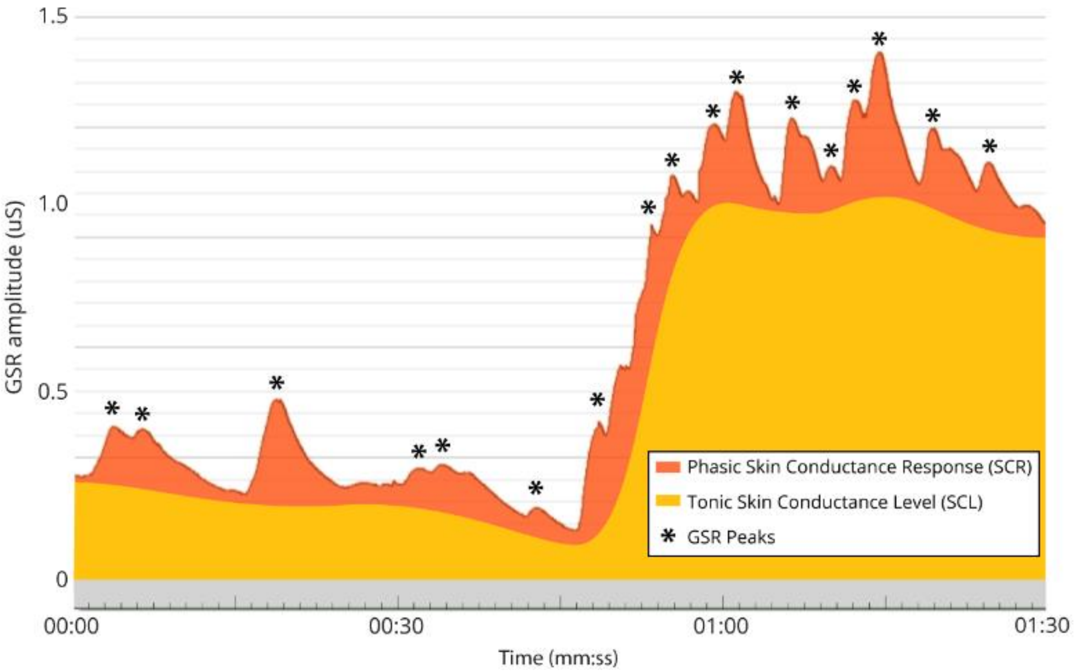


Figure 3.8: Electrodermal Activity – Tonic and Phasic Change

Signals from the autonomic nervous system in terms of arousal are reflected through phasic change. This type of change known as the skin conductivity response (SCR). Phasic events correspond to short-term peaks in the skin conductance that are accompanied by varied rates of decline. SCR signals can be triggered through the presence of environmental stimuli such as sound, smell, or sight. Thus, research has predominantly focused on the evaluation of SCR as it is more informative than SCL.

Like HR, EDA has been used by the research community to provide insight into emotional states. As discussed previously, alterations detected through the monitoring of EDA indicate a change in emotional arousal. While changes in arousal can be monitored and evaluated, it’s not clear what the state of arousal may indicate by simply viewing the signal. Researchers must correlate these changes with content which may induces excitement, or stress during an activity. For example, in [97] the author evaluates the use of EDA as a methodology to provide insight into stress for those who experience increase cognitive activity as they experience a test. Results indicate a change EDA, more specifically for a tonic signal when there is an increase in cognitive activity.

3) Eye Measurement

Interpreting the world around us, the human eye is a primary organ in the human sensory experience. The eye captures and delivers a stream of visual data from our environment to be interpreted by the brain. Although not literal, the eye is often described as a window to the soul. Research has long explored the voluntary and involuntary responses of the eye in an effort to gain insight into human emotion [27] and levels of cognitive activity [103].

There are numerous methodologies which exist to monitor the physiological responses of the eye. Within the context of multimedia evaluations, eye measurements can be gathered at an entry level using software driven solutions which incorporate consumer available web cameras. Examples of these software packages include PyGaze [104], or OGAMA [105]. Each of these respective options are however not without their limitations. Typically, software solutions are only capable of monitoring eye gaze and do not capture more complex oculomotor events such as pupillary response. More advanced software solutions are available but require specialist hardware. Companies such as Tobii [106], develop a wide range of devices (both desktop and HMD) which provide opportunity to track advanced measures of eye gaze, pupil dilation, and blink count. Typically, these products work by illuminating the iris with an infrared light (Figure 3.9 (adapted from [107])), high speed cameras work in parallel to track the pupil at a rate of 120 Hz. Oculomotor Events can be monitored using a proprietary software solution or interpreted manually through the Tobii SDK. Thus, providing the flexibility for research.

Figure 3.9: Eye Tracking Technology

(left) Close Up View of an Iris - (right) Eye Tracking Software Identifying a Pupil Illuminated by Infrared Light

a) Classifying Oculomotor Events

The human eye is controlled by three pairs of muscles (superior and inferior rectus, lateral and medial rectus, and lastly the superior and inferior oblique) [108]. These are responsible for altering the pitch, yaw, and roll of the eye therefore allowing for 3-DOF. Oculomotor movement can be classified into seven different events (Table 3.4, adapted from [109]). These events are classified by monitoring three distinct properties of eye movement, namely: movement duration; amplitude; and velocity. Movement *duration* describes the time it takes to complete an eye movement. *Amplitude* refers to the distance of an eye movement, and *velocity* measures the speed of eye movement.

Table 3.4: Approximations of Oculomotor Events
ms=milliseconds, s=seconds

Classification	Duration (ms)	Amplitude	Velocity (s)
Fixation	200-400	–	–
Saccade	30-80	4-20°	30-500°
Post-Saccadic Oscillation	10-40	0.5-2°	20-140°
Smooth Pursuit	–	-	10-30°
Micro-saccade	10-30	10-40’	15-50°
Tremor	–	< 1’	20’
Drift	200-1000	1-60’	6-25’

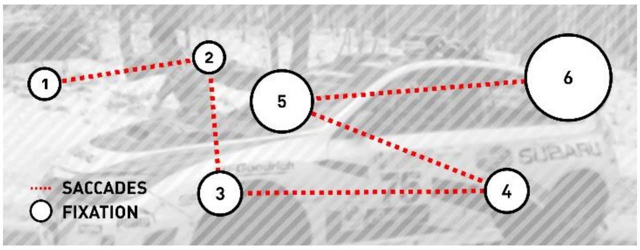


Figure 3.10: Fixations and Saccades

Although seven oculomotor classifications exist, these can be subcategorised into primary and secondary classifications. Primary classifications such as fixation and saccades (Figure 3.10) describe the core movements of the eyes, while secondary classifications are events which occur during or as part of a primary oculomotor movement. The following presents an overview of fixation and saccadic events, and their associated secondary classifications:

- **Fixation** – Events in which there is minimal movement from the eye can be classified as a fixation event. Typically, this occurs when the eye fixates on a region of interest for a duration of 200-400 ms (on average). These events are often used to understand levels of attention. Visual fixations are not completely static, they consist of micromovements which can be classified into three distinct measures:
  - ***Tremors*** – Although fixations report oculomotor responses in which the eye lacks movement, the eye is not completely still. Tremor events arise within fixations, these can be described as minor vibrations which occur as the eye fixates on the world.
  - ***Drifts*** – Ocular drift events are described as an event in which the eye remains fixated on a region but begins to slowly drift from the central focal point over time. Autonomic responses compensate for this drift, returning the eye to the original point of focus.

  - ***Micro-Saccades*** – Compensating for drift, micro-saccades are the tiny movements which assist in drift compensation, therefore allowing the eye to retain a fixation for a longer period.
- **Saccades** – The eye moves at a rapid pace between visual regions which capture human attention. These rapid transitions are described as saccades and are one of the fastest movements which the body can produce.
  - ***Post-Saccadic Oscillation*** – Often described as "dynamic over-shoot" or "glissade", a post-saccadic oscillation is the term used to describe instabilities in eye movement directly after a saccade occurs. It has often been debated if this micromovement is a result of hardware error/compensation or an actual measure of the eye.
  - ***Smooth Pursuits*** – Smooth pursuits are a variant of saccades in which the eye does not operate in a rapid motion instead the eye slowly tracks a visual movement within the FOV. An example of this would be the slow and steady movement as the eye locks focus on a moving object (e.g. a bird flying in the sky).

Despite the variety of oculomotor events, research within the multimedia community has primarily focused on the reporting of fixation and saccade data. This trend has been driven by limitations with respect to the hardware capabilities of entry to mid-range eye tracking devices. In parallel, there is more known about these primary events in contrast to secondary events.

b) Pupillary Response

Located on the centre of the iris is a black circular area known as the pupil. The iris dynamically changes the size of the pupil in response to visual stimulus. This involuntary change in pupil diameter is known as the pupillary response. Controlled by the sympathetic fibres which run throughout the body, the pupil expands and contracts as a result of signals originating in the subsystems of the ANS. Due to the autonomic nature, pupillary response has been widely used as a reliable measure of emotional and cognitive states [27] [110].

Dilation of the pupil has been linked to signals originating from the SNS which is an indicator of the fight-or-flight response. Existing works [27] [110] [111] have demonstrated a strong correlation between these responses and the influence of emotion or increases in mental workload. Physiological signals which contract the pupil originate in the PNS subsystem of the ANS, therefore demonstrating a correlation with the rest-and-digest state of the human body.

Although pupillary response serves as an accurate measure of physiological and psychophysiological response, from an experimental approach it is only technically valid under controlled circumstances. As the eye encounters a dimly lit space the pupil begins to dilate, similarly, in bright environments the iris will produce an autonomous response which results in the pupil contracting. Due to this, external factors such as rapid changes in luminance need to be regulated as they can greatly impact the interpretation of pupillary response data [110] [112] [113].

c) Interpretation and Presentation of Oculomotor Events

With such a wide array of reportable measures, oculomotor events are reliant on the interpretation and presentation of complex data sets to determine effect. There are numerous well-established manual and automatic approaches to the reporting of oculomotor events. Examples of such include the capture and reporting of gaze density maps [114], areas of interest [115] and scanpaths. [116]

(i) Gaze Density Maps

The presentation of visuospatial co-ordinates takes many forms. Gaze density maps is a term used to describe numerous approaches utilised to visualise complex eye data sets. There are three primary forms of gaze density maps they are, heatmaps, luminance maps, and gaussian landscape maps [114]. The following provides a summative overview of the core differences between each approach.

- **Heatmaps** – In short, heatmaps are used to visualise attention using a colour gradient which is often superimposed on the visuospatial scene.
- **Luminance Maps** – Similar to heatmaps, luminance maps transpose visuospatial data into brightness levels, as colours range from black to white. Subtractive image processing techniques are used to reveal visual attention.
- **Gaussian Landscape Maps** – Delivering a three dimensional model of visual attention, gaussian landscapes provide a more advanced mathematical approach to the visualisation of eye measures.

Gaze density maps in general are typically driven by two core parameters, visual position and time. The visual position controls where the map is drawn, while the time parameters alter the intensity of the visual area. Based on the approach, time parameters are typically used to alter the transparency and colour of the heatmap. Although not absolute, heatmap colours typically range from colder (i.e. blue) to warmer colours (red) in an effort to visualise intensity (Figure 3.11). To deliver a holistic view of visual attention, heatmaps are often overlaid on the original multimedia content (Figure 3.12).

Figure 3.11: Heatmap Colour Variation

Figure 3.12: Heatmaps – Capturing Insight into Oculomotor Attention

Considering the scales presented in figure 16, areas which are highlighted in red gain significant oculomotor attention.

The presentation of visuospatial data using heatmap based approaches is well recognised [114]. This is due to how often the approach has been embraced by multiple disciplines across the scientific community. Variations of heatmaps do exist, e.g. luminance mapping. Much like heatmaps, a luminance map transposes numerical visuospatial data into a gradient based image [114]. Generated images visualise areas which gained no significant attention as black, however as colours fade from this colour to white, they demonstrate an area of visual attention. These monochromatic images are then overlaid onto the original multimedia source. Subtractive image processing techniques are applied to convert bright areas to varied levels of alpha, therefore only reporting the areas which gain significate attention (as per Figure 3.13).

Figure 3.13: Luminance Mapping

Brighter areas visualise areas of attention, while darker areas reveal areas which were not considered.

Although gaze density maps such as heatmaps deliver insight into oculomotor movement and visual attention, there is a loss of detail in this form of reporting [114]. Unlike other approaches, the primary reporting mechanism is fixation. Due to the rapid change associated with saccadic events, visual trails between fixations are often not heavily visualised. This is evident in both Figure 3.12 and Figure 3.13 in which areas of attention are noted in the top-left, top-centre, and mid-right of the image. Despite these clear pre-defined areas, the visualisation of saccadic direction has been obscured.

(ii) Area of Interest

Understanding regions which gain significant attention throughout a multimedia experience is core to gaining insight into the human, system, and context factors which influence the user perception of an application [115]. Typically, these approaches rely on researchers pre-defining and capturing eye movement in specific areas of an application (Figure 3.14). In its simplest form, region data can be captured and reported as a percentage of time viewed during the multimedia evaluation (Figure 3.15). This methodological approach has been described under many different monikers throughout the years, examples of such include regions of interest, interest areas, or zones. In this research, we refer to the approach as "Area of Interest" (AOI) as this term has become more established in the research community. The primary objective of an AOI is to report what is considered "interesting".

Figure 3.14: Eye Tracking – Area of Interest

(left) A website overlay, (right) a series of pre-defined regions

Figure 3.15: Area of Interest – Percent Based Reporting

(Left) AOIs layered over a website, (Right) Output file summarising data captured

Unlike other efforts employed to understand oculomotor events, defining what constitutes an AOI can be challenging. There are numerous parameters which must be considered. Examples of such include visual entities, position, size, and other considerations such as how to handle overlapping stimulus. An example of this can be seen in Figure 3.14 which presents an overview of a website. Both regions 1 and 2 are located within a similar area. From a web development perspective, this is described as the navigation bar. This however can be further subcategorized, as the left and right area serve their own independent purposes. Region 1 visualises the logo and can be utilized to evaluate user attention from a brand recognition perspective. However, region 2 presents the website menu navigation. As a result, it is expected that oculomotor responses in this area differ in contrast to region 1. There may for example, be more events which lookback to region 2 overtime as a user attempts to navigate the website.

Typically speaking, areas of visual interest can be defined manually. Figure 3.14 presents the use of AOIs in a web context, regions are clearly separated using the website structure, alterations in colour and by considering areas of interaction. Other visual stimulus, such as photography often do not consist of clearly defined visual regions. Photographers often convey a clear definition in midground, foreground, and background (Figure 3.16). Although considerations within this context would work for landscapes and macro based photography, it does not work for all visual stimulus. As a result, other mathematical or algorithmic based approaches to the definition of AOIs should be considered.

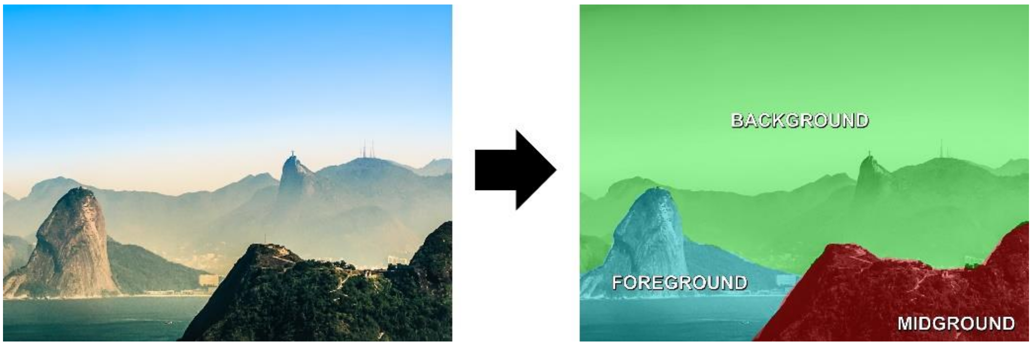


Figure 3.16: Photography – Midground, Foreground and Background.

One common method employs a mathematical visualisation model known as a Voronoi tessellation diagram [115]. This approach is visualised in Figure 3.17 (adapted from [117]), first central markers are placed over key visual objects and a series lines connect the regions. Perpendicular lines are then drawn, and sections are subdivided equally. Once this process is complete, the lines connecting the central markers are removed and a series of cells are revealed. The cell structure is often defined through a unique variation in colour or contrast changes.

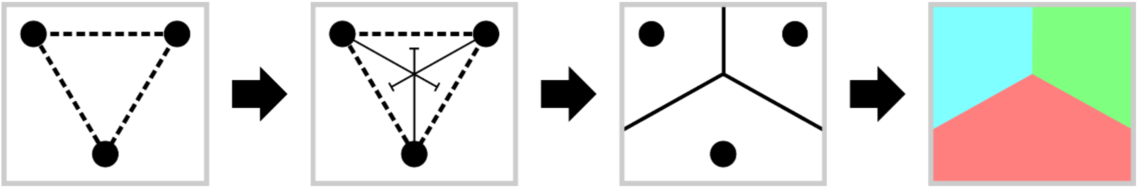


Figure 3.17: Voronoi Tessellation Operational Overview

Voronoi tessellation automates the division of AOIs to a large extent. However, they are also utilized in research which contains problematic data, quantifying and reporting the distributions of fixation of a specific area. Such approaches should be implemented with caution, due to the reduced level of accuracy. Pending the complexity of the imagery, Voronoi tessellation diagrams can encompass regions not directly associated with the region identified using the central markers. Areas in which the visual stimulus does not cover, are known as whitespaces. Typically, whitespaces occur when the Voronoi tessellation overfits areas to reach the image boundaries. This process is visualised in Figure 3.18 (adapted from [117]) which highlights the overfitting as the white shaded area.

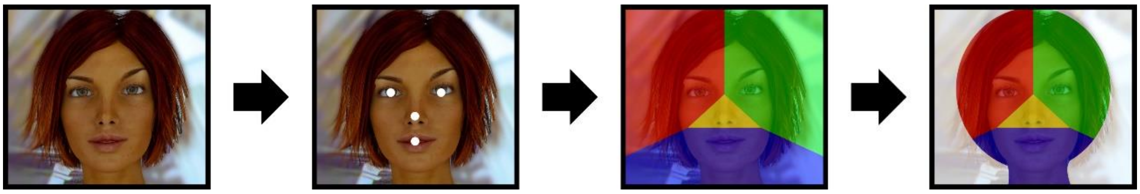


Figure 3.18: Voronoi Tessellation – Whitespace (Overfitting)

The definition of an AOI in static multimedia can be challenging. Voronoi tessellation is only one of the many methods which can be used to define visual regions in image content. There are numerous other approaches such as gridding (discussed later) and the limited-radius Voronoi tessellation method [117]. Each of which have their own advantages and disadvantages.

Going beyond static images, multimedia experiences which present video content (i.e. animation or video) consist of another challenge. Unlike traditional imagery which content remains in the same position. Moving images may alter the position of the AOI and in some cases only present it for a small amount of time before the scene changes. To circumvent these challenges, researchers must take an alternative approach and define what are known as dynamic AOIs. Keyframe based approaches are utilised to mark position, shape and size changes of objects throughout multimedia playback. Much like the approach to animation, intermediate frames are calculated using software generating smooth transitions between multiple keyframes.

Such approaches alleviate the analysis of recorded video content. These methodologies not only solve challenges with pre-recorded video content. But they can be also implemented as a method to understand visual attention for eye tracking in the real world. In recent times, devices such as the Tobii Pro Glasses 2 [118] now allow the capture of oculomotor movements, whilst also allowing their users to freely traverse the real world. These smart glasses utilize a series of eye tracking cameras positioned under the eye, and a front facing camera which captures the real world. Data is then combined post-experience, thus facilitating a holistic view of a user's oculomotor movement as they independently traverse a real world environment with 6-DOF.

Due to the uniqueness of each user's movement, there is a requirement (from a time context) to classify fundamental AOIs within the video content. As such, approaches within this context are focused and should minimise the research focuses within the environment. Often marketing giants employ such approaches to determine if the visualisations of their products stand-out amongst others within a supermarket shelf. Regardless of which approach is taken, decisions with respect to AOI parameters should be driven by the research hypothesis.

d) Scan Path

Historically described as "scan pattern", "fixation tracks" and "gaze sequence", the term scanpath originated in the 1970's. According to [116], a scanpath is defined as the *"route of oculomotor events through space within a certain timespan"*. At its simplest, a scan path is used to understand the viewing pattern of an individual, therefore reporting how the eye traverse's visual stimulus within their FOV. Scanpaths have been visualised via numerous methodologies in recent years. However, the most well-known approach is String-based. Scanpaths rely on an understanding of AOIs and how they are predefined. There are two approaches to string-based scanpaths, they are semantic AOI and gridded AOI. As their names suggest, both of these approaches are inspired by how AOIs are defined. Unlike normal AOIs, it's important to remember the reporting mechanism places an emphasis on visual spatial tracking as opposed to temporal based reporting of oculomotor response.

Figure 3.19: Scanpath – Semantic Reporting Methodology

(Left) AOIs are predefined in an effort to subcategorise the website structure
(Right) Sematic labels are utilized to deliver an overview of visual scan-path

Semantic AOIs report visual attention as a string of semantic descriptions. As described previously, AOIs often consist of multiple, pre-defined interest regions (Figure 3.14). Semantic AOIs requires the alteration of these regions, into descriptive labels which describe the underlaying content. An example of this can be seen in Figure 3.19, first the visual stimulus is segmented into core visual areas. Following this, semantic descriptors are utilized to deliver a descriptive overview of each area. As the eye traverses the image, saccadic events are captured as the eye moves between area to area. This is often reported as a string of characters, thus the umbrella term of "string-based" methodology is used. In Figure 3.19, the reported scanpath would take the form of Slider, Heading, Article 1, Article 2, Article 2 and Article 1.

Sematic AOIs rely on pre-defined regions of interest. The variable nature of stimulus content often results in an unbalanced collection of regions, each of which have their own respective size. As a result, this results in a loss of detail. To circumvent this, smaller sub-regions can be defined. However, this can alter the perception of results, due to this it is often better to explore alternative approaches such as gridded AOIs.

Unlike semantic AOIs, gridded AOIs are implemented utilizing an overlay of equal size regions across the stimulus, this overlay takes the form of a grid. To begin, the visual region is separated into a series of rows and columns (Figure 3.20 (Left)). The grid size can be altered pending the complexity of the visual stimulus. Values on the X and Y axis are each assigned a series of letters or numbers to differentiate each region. As the eye fixates on various regions, a semantic description of the location is generated utilizing an X and Y value. For example, in Figure 3.20 (Right) the generated report would be A1, D3, D5, B6, 7D and F4.

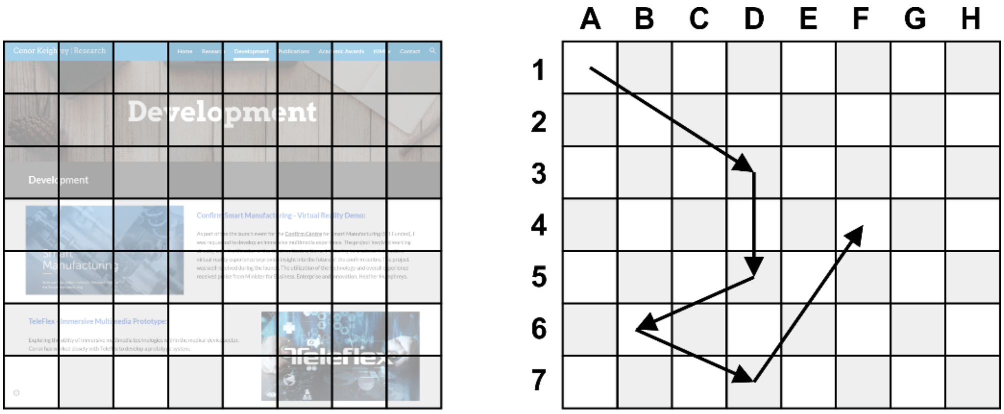


Figure 3.20: Scanpath – Gridded Reporting Methodology

(Left) A gridded area is predefined over the website structure
(Right) Scanpaths are reported by referencing each cell on the X and Y axis

Both gridded AOIs and semantic AOIs report spatial change through the visualisation of saccadic data. They can however be adapted to also include temporal information captured from fixations. Traditionally, scanpaths are visualised moving from fixation to fixation (Figure 3.21 (Left)). This approach, regardless of reporting methods results in a loss of detail. By altering the visual representation, translating temporal data into a change in fixation size a hybrid overview of the scanpath can be revealed (Figure 3.21 (Right)). This minor change delivers a more holistic overview of the movements as they traverse the visual stimulus.

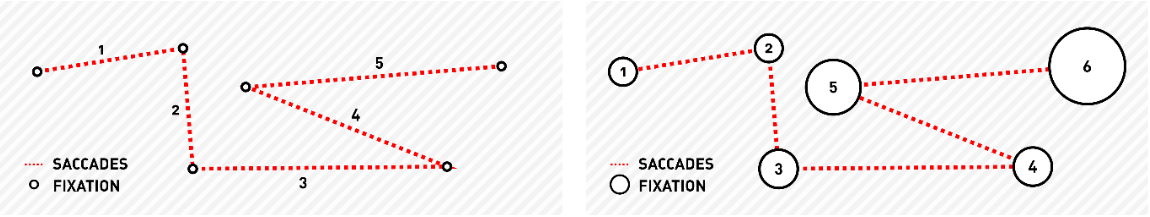


Figure 3.21: Spatial and Tempoaral Scanpath Reporting

(Left) Squential based reporting methodology, (Right) Squential and Fixation based reporting methodology

3.3 Quality Measures of the Future

In recent times, the QoE community has embraced the measure and analysis of physiological metrics inspired by works like [27]. Recent works have taken an experimental yet cautious approach to the measure of QoE through the delivery of mixed methods evaluations. These approaches aim to understand the synergy between explicit, implicit, and objective measures, with qualitative findings supporting the experimentally captured of quantitative data.

However, the physiological measure of user experience is not without challenges. Although key insight into user emotion can be formulated through data analysis, more work is needed to develop a better understanding of user responses. As it stands, reliability lies within the reproducibility of research, as such the delivery of datasets to the multimedia community are a fundamental cornerstone in the evolution in QoE evaluations.

Explicit metrics have remained a core component within the QoE community due to continued support and development of widely accepted evaluation strategies for the purpose of experimentation. Similarly, the future development and understanding of implicit (physiological) measures relies on the formulation of an accepted standard for capture, and analysis of such novel metrics. This of course is complex, as is the ANS in which these signals originate. Thus, the future understanding of physiological responses relies on interdisciplinary works across multiple domains within health, computing, and multimedia.

Of interest, current trends in technology highlight how users embrace sensor technologies not only on their devices, but those linked to their bodies also. A prime example of this can be seen in the adaptation of smart watches which can now capture and report measures of fitness and wellbeing to the user each day. This trend is welcome to researchers in the field of QoE research. As sensor devices become more mainstream, there is greater potential for the capture and analysis of physiological responses within the future of multimedia experiences.

From an analysis perspective, these complex systems create many challenges. As devices evolve, data streams must converge into a single point of reference. Alternative to this, computing may face an increased demand in computational requirements in order to interpret siloed complex data sets. As such, alternative approaches to data analytics such as machine learning (ML) may alleviate some of these challenges. As more researchers embrace these processing methodologies, the future of multimedia applications could potentially pertain to smart multimedia experiences. Which not only adapt to the objective measure of user performance, but the autonomic physiological responses of the human body. Therefore, facilitating unimaginable continuity between immersive computer systems and humans. Thus, paving the way for adaptive systems which consider the continuous and uninterrupted evaluation of QoE.

Chapter 4: RELATED WORK

In Chapter 4, related works are presented. An emphasis is placed on presenting and critiquing the key influential research works which have guided the topic of this PhD. Crucial areas include: QoE, Multimedia, and SLT assessment methodologies. In addition, more precise works in the context of cognitive load measurement using physiological measures and interpretation of emotional arousal are explored. The objective of this is to provide more context with respect to the findings outlined later (Chapters 5 – 7) within this PhD thesis.

4.1 The Complexities of Explicit Measures

A fundamental overview of subjective capture methodologies was presented in section 3.2.1. However, these methodologies merely serve as a guide for researchers to implement a framework for the purpose of subjective evaluation. Beyond this, the creation process associated with questionnaire or survey content for a subjective evaluation can be a complex task. Often overlooked, there are numerous influence factors of a subjective evaluation, such items include wording [119], participant attention span [119], and questionnaire length [119] .

The importance of wording is a well-accepted paradigm within the development of questionnaires for the purpose of QoE evaluations. According to [119], improper wording, or phrasing can affect the validity of the experiment and its results. As such, domain specific terminology should be used sparingly. Instead an emphasis should be placed on presenting clear and precise language which does not cause confusion for the participant. Reports highlight how humans can often have a short attention span [119], as such the length of a question can play a fundamental role within an evaluation. Other issues which are not to be ignored include questionnaire length and repetition. In [119], the researchers highlighted how longer questionnaires can induce a degree of urgency within a participant to finish. As a result of this, the data captured can often represent an inaccurate measure of subjective experience [119]. To circumvent these challenges, researchers often employ reliability evaluations [120] in which aggregated user feedback is utilized to ensure clarity within the questionnaire.

Running full-scale evaluations is often considered time-consuming, and in some cases can be costly. As such, the importance of pilot testing should not be overlooked. Pilot tests consist of small scale preliminary studies, they are used to reveal any ambiguity in questionnaires and to identify problems with the overall test methodology [119]. Typically, pilot tests are carried out with those familiar with the works (i.e. colleagues). Feedback should be extrapolated from participants and integrated into the overall test protocol before beginning the full-scale evaluation. It’s important to remember, that although feedback is invaluable, no questionnaire will not be criticized to some extent [119].

Traditionally, explicit capture of user emotion has been accomplished using user ratings such as the Self-Assessment Manikin (SAM) scale. The SAM was developed in 1980 [121], it is a pictorial rating system in which users’ rate three measures of emotion: arousal, valence, and dominance. Users provide positive or negative ratings associated with each of the three dimensions. Responses can be correlated with the widely accepted valance, arousal, and dominance (VAD) [122] emotional state model (Figure 4.1 (adapted from [122])). As such revealing a glimpse into the explicit measure of user emotion during the time of testing.

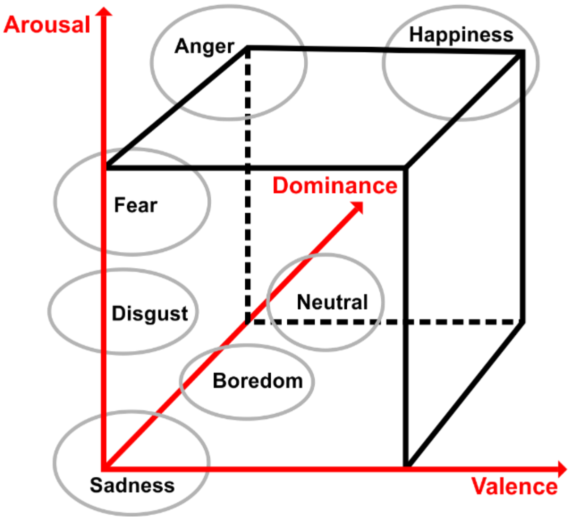


Figure 4.1: Valance, Arousal, and Dominance (VAD) Model

Due to the nature of QoE evaluations, researchers consider a multitude of influence factors associated with their specific application, system, or service. Researchers not only have to consider human influence during the development of a post experience questionnaire, they must also reflect on alternative influence factors. One such example is the influence of system expectations. New and emerging multimedia technologies are often associated with marketing content which promises users a more immersive experience. But what exactly is immersion? According to [123], from a gaming perspective, immersion (also known as presence) is the ideology that a player feels truly “in the game”, this feeling dwells beyond the holding of attention, it can often captivate the user to the point that they are emotional involved within a virtual environment. To fully understand this concept, we must factor how users become immersed in content, according to [123] there are three core components: physical, emotional, and narrative immersion. The physical component measures the feeling of a person being transported within the virtual environment. At an emotional level, factors such as the ability for a game to elicit an emotional response are considered. Lastly, there is a level of immersion associated with narrative, as such character interaction, or an engaging story can often provoke this type of a response.

The ideology of immersion fits well within the scope of user perception and as such the importance of evaluating this topic is arguably key to the successful development of the next generation of multimedia experiences. As highlighted in section 2.3.2, the influence of quality goes beyond human factors. System and content factors play a fundamental role within the judgment process of quality within an immersive experience. These are important factors which should not be ignored. But how exactly is immersion measured? To answer this, we must look to the games industry, who have been motivated over the years to deliver engaging experiences for gamers as an attempt to develop a player base.

A series of questionnaires exist which have been designed to assist in the evaluation of this very topic. In [124], authors present a comparative analysis of three widely accepted questionnaires, namely: Immersive Experience Questionnaires (IEQ) [125], Game Engagement Questionnaire (GEQ) [126], and Player Experience of Needs Satisfaction (PENS) [123]. Table 4.1 (adapted from [124]) provides a comparative overview of the key components within each of these respective questionnaires. While the IEQ and PENS take a more direct approach to the measure of immersion, the GEQ factors immersion through the measure of engagement.

Table 4.1: Player Experience Questionnaires and Components

Questionnaire	Components
Immersive Experience Questionnaire (IEQ)	Cognitive Involvement Emotional Involvement Real World Dissociation Challenge Control
Game Engagement Questionnaire (GEQ)	Absorption Flow Presence Immersion
Player Experience of Needs Satisfaction (PENS)	Competence Autonomy Relatedness Controls Presence/Immersion

Despite core differences associated with each of the corresponding components, there are similarities associated with the content of the questionnaires. To demonstrate this, authors in [124] carried out an experimental evaluation online consisting of 270 participants. The purpose of this was to gain insight into the level of convergence associated with questionnaire content. Results of the study reveal a large correlation with core components of the questionnaires. Thus, demonstrating the validity of each approach to the evaluation of user experience within virtual environment.

Targeted specifically towards QoE evaluation, more recent works [127] highlight attempts to measure and quantify immersion as two separate entities, spatial immersion, and emotional immersion. To measure these individual aspects of immersion, authors developed a spatial and emotional immersion questionnaire which was inspired by already validated questionnaires on immersion and presence. A subjective evaluation was carried out in which authors compared the results of two groups (TV vs Tablet). Using video content sourced online, participants rated two video scenes. The findings of the study are multi-fold, authors highlight how current evaluations often place an emphasis on delivering an immersive experience by focusing on the spatial content. But the importance of narrative elements which play a key role within the perception of emotional immersion are often overlooked.

The future of interaction in immersive multimedia environments is multidimensional. Users are no longer limited by the confinements of a 2D interface. The advent of room scale tracking provides opportunities for users to experience their virtual space using 6-DOF. As such, the ability to perceive spatial content is an important aspect which must be considered within the future evaluation of immersive multimedia experiences.

Although questionnaires have remained mainly unchanged, recent works highlight novel approaches to capture user ratings using interactive methods e.g. controllers, wearables etc. In [128], authors adapt a handheld controller as a subjective rating mechanism throughout a multimedia experience. There are many advantages associated with such an approach. The presentation of surveys post experience can often suffer from a lack of reflection on an entire experience. Alternative approaches have seen the introduction of surveys throughout an experience; however, this can interrupt a multimedia experience. As such, controller based systems which capture user perception throughout the experience has the potential to alleviate these influences.

Beyond the scope of subjective capture methods is the understanding of user ratings. MOS ratings serve as a cornerstone within the QoE community; however, they only provide a small insight into the overall user experience. It is only through deep analysis that the true measure of subjective response can be captured. Hoßfeld et al. explores the importance of evaluating user ratings beyond the restrictive nature of a MOS value [84]. More specifically, the authors highlighted how averaging user ratings often remove diversity in response. As such, authors describe an approach in which the standard deviation of Opinion Scores (SOS) should remain central to experimental analysis in QoE evaluations. Thus, delivering a more insight evaluation of results to the research community. The authors also demonstrated: how factoring SOS reveals if user studies can be directly compared; how SOS assists in revealing the reliability of a QoE evaluation; and further evaluates the underlying systems.

4.2 Understanding Implicit Measures

Moving outside the recording of subjective metrics is the capture of implicit measures through the monitoring of physiological responses in multimedia evaluations. A great body of work exists within the QoE research community which implement sensor technology to capture insight into user emotion [21] [80]. Section 3.2.2 provides an in-depth overview of the multitude of measures which can be captured an analysed within the experimental process. For the purpose of this PhD, a focus is placed on presenting works which directly relate to the experimental results discussed in Chapters 5 – 6.

In [101], Shi et al. presented the preliminary findings of a study in which EDA was evaluated as an index of cognitive load. Participants interacted with three traffic management scenarios which consisted of varied interaction methods (hand gesture only, speech only, and hand gesture and speech only). Four difficulty levels were developed to challenge the user as they progressed through the activity. To induce cognitive load, application parameters such as visual complexity, number of actions to complete, and time limits were modified over time. Results demonstrate an increase in EDA (as a mean) associated with cognitive load as users progressed from low to high difficulty scenarios. Although a trend exists with respect to captured EDA signal, the small sample size proved to be not statistically significant, as such more work needs to be explored within this application domain.

Similarly, Chen et al. [129] explored the influence of stress in the measure of cognitive load using EDA. Utilizing a within group design, 8 subjects took a quiz in which mathematical questions were answered via multiple choice. As participants progressed throughout the activity, the difficulty increased through the introduction of additional mathematical operators. The results of the study indicate that the mean EDA value increased over time. This coincided with the increase in difficulty. However, a second condition which aimed to induce stress could not emulate this finding. Stress was induced through the introduction of time limits, display of feedback (i.e. out of time, wrong/correct), and monitored user interaction. Authors hypothesized that when stress levels are fluctuating, as per the second condition, EDA cannot be used to effectively index cognitive load. The authors also explored the use of feature extraction techniques to deliver a more informative insight into EDA responses which may indicate adjacent levels of stress associate with cognitive load.

Works outside of the multimedia domain have also evaluated the use of EDA as a measure of cognitive load. In [130], authors explore the capture of multiple physiological measures from pilots during flight. Ten general aviation pilots took part in the evaluation. Throughout a series of flights, objective measures in the form of HR, EEG, EDA, and eye measures were recorded. In addition, pilots completed surveys at key points during flight to provide insight to their mental workloads. Of interest here, are the results associated with an increase in EDA during take-off, touch and go exercises, and landing procedures. It is at these key stages in which there is a higher mental demand placed on the pilots. As such, the results further validate the use of EDA as a measure of cognitive load.

Although research described in [101], [129], and [130] describe similar responses with respect to the use of EDA as measure of cognitive load. They suffer from small sample sizes, and as such do not deliver a statistical significance associated with the findings. Of interest though, is that all three of these studies demonstrate a similar change as discovered in early research into EDA as a measure of cognitive load. In [97], Kilpatrick et al. explored the capture of EDA in high and low stress scenarios. The document reports that cognitive activity can be observed through a slow increase within the tonic signal component of EDA. Arguably, the findings in [101] and [129] follow a similar pattern as both studies discover a slow and steady increase as cognitive load associated with a task is increased.

4.3 QoE in Immersive Multimedia Experiences

Research into the use of immersive multimedia technologies as a learning mechanism within education [17] and industry [18] has received significant attention in recent times. The results of such studies often indicate a high success rate in terms of the transfer of knowledge. Developed to enrich the learning of chemistry within a classroom setting, [17] presents the findings of a field study which evaluated a mobile AR experience. Results of the study demonstrate a greater level of learning in AR when compared to that of a traditional classroom environment.

A survey of existing quality assessment methodologies which focused on AR visualizations is presented in [22]. Concerns and limitations associated with AR visualizations, such as the perceptual issues and restricted field of view relating to display technology, were identified. Of particular interest were applications associated with neuronavigation. Each participant was asked to perform a precision-based task, which was similar in terms of ergonomics, interaction, and human behaviour to a neuronavigation system. As a result of the experiment, the author proposed a quality assessment method that focused on a mixed-methods approach to assess quality in AR neuronavigation fields.

Similarly, in terms of VR, training scenarios have been developed to target a diverse range of learning environments. One such example is within the agricultural sector. In [18], authors evaluate an immersive VR tractor driving simulator to assist in the training of drivers in high risk scenarios. Such systems are ideal for training drivers in real world scenarios which may present risk to the driver. An example of such is tractor roll over which is one of the leading causes of fatalities in the agricultural sector. The results demonstrated an increased perception of risk and safety amongst participants, in addition to fewer errors in deploying relevant tractor safety mechanisms.

In [21] the correlation between physiological measures (EDA and heart rate) and subjective data as users experienced a virtual environment in a video game was investigated. The subjects were exposed to three first-person shooters for a twenty-minute time period and asked to complete an in-game experience questionnaire (iGEQ) every five minutes. The results reported a statistically significant correlation between heart rate and the subjective data gathered from the iGEQ across seven dimensions of gameplay.

The work presented in [80] described the evaluation of user QoE comparing immersive VR and non-VR environments. Using a within group design, a total of 33 participants evaluated the multimedia experience presented on an Oculus DK2 and a computer monitor. The mixed methods evaluation captured subjective measures via post-test questionnaire and objective measures in the form of heart rate and electrodermal activity. The results of the study reveal a slight elevation in heart rate from participants during the immersive VR experience.

The future of immersive multimedia is no longer limited by the stimulation of two senses. There now exists a series of components which allow developers and researchers alike to enhance immersive multimedia through the stimulation of haptic [131] and olfactory [132] senses. Such advances in technology aim to explore the addition of natural sensory processing to enhance a multimedia experience. Examples of works within the domain can be seen in [9], an extension of the work carried out in [80]. [9] describes the initial results of a study in which participants experienced a VR experience which aimed to evaluate the addition of multisensory components within a virtual tourism environment. Results published to date have focused on the presentation of the subjective ratings of two groups: a VR only group, and a haptic enhanced VR group. The subjective analysis revealed only one statistically significant result in which the olfactory group reported higher levels of sensory stimulation when compared to the VR only group. Although only a small finding, the utility of adding multisensory components to immersive multimedia environments is revealed.

4.4 Speech and Language Therapy Applications

The use of technology as a way to treat those who experience symptoms of aphasia, or speech and language difficulties is not a new concept. However, a predominant number of publications often place emphasis on the development of novel treatment methodologies, whereas this work places a focus on the use of multimedia technology to enhance the assessment process.

The application of HMDs as an assistive technology for those who suffer from aphasia was explored in [133]. The authors carried out two studies using Google Glass. The first, investigated the benefits and challenges associated with the use of such technology. An interview-based study was carried out with 8 participants who had aphasia. Participants experienced two storyboard scenarios in which vocabulary prompts were displayed on the HMD. Despite some issues with device specific interaction methods, early findings were positive. To investigate this further, a second study developed and evaluated a prototype application which required participants to complete specific conversational tasks. Results indicated that participants were able to maintain focus on the conversation without the reliance on external tools. However, in some scenarios, multitasking between the system and conversation proved challenging.

Developed as a language rehabilitation tool, an immersive multimedia experience for storytelling was evaluated in [134]. The research aimed to assist those who experience symptoms of expressive aphasia in the redevelopment of language skills. Participants experienced the immersive application using an AR HMD. Using a series of specifically developed markers, interactive props appeared to the user to stimulate their imagination.

Users were asked to use the interactive props to assist in a story telling exercise which aimed to support the redevelopment of expressive language skills. A study was carried out between an AR and non-AR assisted story telling activity. Results reported that the AR application provided an immediate improvement in terms of language skills. In addition, the novel application of a multimedia technology demonstrated an overall improvement in terms of linguistic flow.

Using a HMD as a vocabulary support mechanism for people with aphasia, authors carried out two studies in [133]. The first of which aimed to explore the usability of such novel technologies, following this a lab based study was carried out which evaluated the delivery of a novel application on the multimedia platform. The target application aimed to replicate the use of symbol based dictionaries which are often used by those who suffer from aphasia as a means to recall words. Although only exploratory, results of the study demonstrated that the use of HMD technology had the potential to allow participants to better focus on conversational topic. The removal of paper based medium as a conversational assistant proved favourable to the participants. Subject reports described how the glanceable nature of a HMD allowed them to remain within the conversation, and as such, more aware of their conversation partner. Overall, the formfactor of a HMD proved positive over the use of pen and a paper system, as content was within their own respective line of sight, thus providing an additional level of privacy to the wearer. Although the results of this study proved promising, there was no comparative analysis carried out between the HMD experience. To address this limitation, authors carried out a follow up study [135] which seen HMD compared to a smartphone in the delivery of a multimedia enhanced symbol based dictionary. The results of the study demonstrate similar findings to [133].

4.5 Novelty of This Work

Considering the extensive background research, methodology, and literature review presented in Chapters 2, 3, and 4, this section highlights the key novelties of this PhD works. It is clear that the capture and processing of implicit measures within QoE evaluations is recently emerging as a valid approach to understanding QoE. Recent works [21] [80] [136] highlight the adaption of such novel measures across a multitude of multimedia applications. Despite this, the existing approaches primarily focus understanding physiological measures across the duration of an *entire experience* and generally have not compared immersive experiences *across multiple platforms*.

Emerging immersive technologies, such as those on the mixed reality spectrum facilitate user independent interaction and control over a multimedia experience. Thus, enabling users to experience these immersive virtual worlds in an unconstrained manner. Future QoE evaluations must adapt to such novel experiences, understanding physiological response not from a limited passive perspective but within an active user-centric multimedia context. In this context, there is a need to design, develop, and understand user perception of quality utilizing a holistic novel methodology which considers explicit, implicit, and objective measures throughout an experience. Such an approach provides opportunity to derive an understanding of precise physiological response at *core stages of interaction* throughout an experience. Beyond this, the capture and processing of *task-evoked physiological response* is proposed, in an effort to understand implicit measures which occur as a result of user specific states of interaction within an immersive multimedia experience.

Delving deeper into the experience, visual attention is a fundamental component in the understanding of user perception within a multimedia experience. As such, the development of a *QoE fused oculomotor interpretation framework* is proposed. This novel component aims to automate the reporting of AOIs and generate a gaze density map of visual attention through a QoE focused methodology.

The next Chapter presents the results of the first experimental study which evaluated and compared the user QoE of an immersive SLT assessment application across three multimedia platforms (AR, tablet, and VR).

Chapter 5: EVALUATING AR, TABLET, AND VR

The findings of this Chapter have been published in the following conferences and academic journals:

1) **Poster Presentation** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray, and Niall Murray. "A QoE evaluation of immersive multimedia technologies for speech & language assessment applications." at Daughters of Charity Technology and Research into Disability (DOCTRID) conference, Limerick, Ireland, 2017.
2) **Paper** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray, and Niall Murray. "A QoE evaluation of immersive augmented and virtual reality speech & language assessment applications." In Quality of Multimedia Experience (QoMEX), 2017 Ninth International Conference on, pp. 1-6. IEEE, 2017 – **(Best Student Paper)**
3) **Paper** - ***Conor Keighrey***, Ronan Flynn, Sean Brennan, Siobhan Murray, and Niall Murray. "Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR Speech and Language Therapy Applications" Proceedings of the on Thematic Workshops of ACM Multimedia (ACMMM) 2017, pp. 485-492. ACM, 2017.
4) **Journal** - ***Conor Keighrey***, Ronan Flynn, Siobhan Murray and Niall Murray. " A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications” IEEE Transactions on Multimedia, 2020.

5.1 Aims & Objectives

This section presents the results of experiment 1, which was a QoE evaluation and comparison of the immersive SLT assessment application. Three technologies were identified due to their suitability as potential platforms for the delivery of such a novel experience: AR, tablet, and VR. Each of the respective technologies were evaluated in terms of explicit, implicit and interaction measures. Employing a between-group design, traditional explicit QoE measures remain central within the immersive multimedia evaluation process. Complimenting this approach, implicit and interaction measures are captured and explored as an insightful measure of QoE within the context of interactive and immersive multimedia experiences. These novel measures also have the potential to assist the SLP within a clinical context, providing a much more informed overview of assessment performance.

The objectives of this Chapter are summarised as follows:

- Evaluate three immersive multimedia technologies as potential platforms for the future delivery of an immersive speech and language assessments.
- Develop and evaluate a novel methodology for the capture of explicit, implicit, and objective measures within an interactive and immersive multimedia experience.
- Explore the utility of capturing objective measures of user performance as a supportive measure of user perceived QoE.
- Capture and report physiological measures of heart rate and EDA in an effort to derive emotional response, as such gaining a significant understanding of user perceived QoE.

As such, this Chapter addresses SubRQ1, SubRQ2 and SubRQ3 as outlined in section 1.2.1.

5.2 Experimental Setup and Methodology

5.2.1 *Immersive Multimedia Systems*

The following subsections describe the technologies used in the experiment.

5.2.1.1 Augmented Reality

The immersive AR experience was delivered using the Microsoft HoloLens [2] (Figure 5.1). The self-contained wireless device operates as a standalone computer. The device consists of a Central Processing Unit (CPU), Graphics Processing Unit (GPU), Holographic Processing Unit (HPU), holographic display, and batteries. Designed with users in mind, the operational components are evenly distributed across the headset to ensure a comfortable fit [137]. The display technology consists of three-layered holographic lenses (red, green, and blue). Two HD 16x9 light engines located above the holographic lenses project images to the user’s FOV, resulting in an AR experience. It is estimated that the FOV of the HoloLens is 30° by 17.5° [138].

Environmental data is captured from a series of cameras and sensors located at the front of the HMD. An inertial measurement unit (IMU) captures accelerometer, gyroscope, and magnetometer data. Obstacles such as walls and furniture are recognised using four monochromatic cameras located at the front of the headset. The state of the art HPU interprets the stream of data from the cameras and environmental sensors.

The primary mechanism for interaction involves the use of hand gestures. A 120°×120° depth camera is situated at the front of the device. It is used to recognise a set of preconfigured hand gestures from the user. Alternatively, there are two other methods of interaction: a handheld controller can be used to interact with content; and second, a series of built in microphones which permit voice operation through the Microsoft personal assistant Cortana [139]. An enhanced audio experience is delivered through two external speakers located above each ear. Working in sync with data from the IMU, users experience a 3D spatial audio environment.

Figure 5.1: Microsoft HoloLens – User Interacting with the AR Experience

5.2.1.2 Tablet

A Lenovo Tab 2 A10-70 [140] was used to present the content for the tablet group. The system runs on the Android 5.0 (Lollipop) operating system. Applications are processed using a 1.7 GHz quad core Cortex-A53 CPU, graphical content is computed in a dedicated Mali-T760 GPU. The 10.1-inch IPS screen has a full HD resolution of 1920 x 1200. A capacitive touch screen captures user input as they interact with the digital media content.

The dual speaker sound bar (Dolby Atmos) located at the rear of the device, delivers an immersive audio experience. Traditionally, tablet computers are operated as the user holds the device. However, to provide consistency between all three groups, a smart cover was fitted to the tablet thus allowing it to stand freely in a predetermined location in front of the user (Figure 5.2).

Figure 5.2: Lenovo Tab 2 A10-70 – User Interacting with the Tablet Experience

5.2.1.3 Virtual Reality

An Oculus Rift Development Kit 2 (DK2) [42] was used to evaluate the delivery of the virtual SLT assessment in a VR environment. An internal OLED screen delivers immersive visuals, providing each eye with a resolution of 960 x 1080 pixels. The DK2 has an FOV of 100°, which is similar to a human eye (approximately 120°). Using a combination of IMU sensors and constellation tracking, the DK2 tracks a user's head movements which are then translated into the virtual environment. This tracking is achieved via an infrared camera mounted in front of the user. It monitors a series of discreetly placed LED's around the HMD. Each LED flashes at unique intervals enabling the tracking of a specific position on the headset. The IR camera monitors the position of the LED's which are then processed to mirror the head position in the virtual environment.

The DK2 is a tethered device. It requires external hardware to perform graphical processing. In this experiment, a PC containing an Intel i7 6600K, 16GB of ram, and an NVIDIA 1080 GTX graphics card rendered the virtual content. This configuration of components ensured a stable framerate of 60 frames per second.

Current generation of VR hardware lack native support for gesture recognition. Although interaction mechanisms exist, they require handheld controllers. To facilitate gesture-based interaction, a Leap Motion controller was included in the system design. The Leap Motion controller is mounted to the front of the DK2 HMD as per Figure 5.3. A series of hand gestures was pre-programmed to imitate a similar interaction mechanism to the AR and tablet groups, thus creating a like for like experience.

Figure 5.3: Oculus Rift DK2 – User Interacting with the VR Experience

5.2.2 *Virtual Speech & Language Assessment*

The virtual assessment was developed using the Unity3D game engine [44]. Unity3D supports application development and compilation for a multitude of operating systems. This approach provides opportunity to develop for all three immersive multimedia platforms in parallel. It also provides equality in terms of signal capture, and processing across all platforms.

The virtual SLT assessment was modelled on the semantic memory assessment contained within the CAT (section 3.1.1). It aims to identify the level of ease in which associative knowledge is retrieved. Poor performance on this type of test is indicative of an impairment to object recognition (i.e. understanding of semantic relationships).

The assessment stimuli (Figure 5.4) consist of one centralized image which is surrounded by 4 outer images. The 4 outer images contain a semantic target, unrelated semantic distracter, distant semantic distracter, and close semantic distracter. The objective of the assessment is to formulate a link between the central image and the target. The test consists of 10 evaluation slides, and one practice slide at the beginning (11 total). A full overview of the assessment content can be found in Table 5.1 and Appendix 1.

Table 5.1: Comprehensive Aphasia Test – Semantic Memory Assessment Contents

Slide	Key	Target	Close	Distant	Unrelated
Slide P	Monkey	Banana	Pear	Chocolate	Envelope
Slide 1	Glasses	Eye	Ear	Mouth	Elephant
Slide 2	Hand	Mitten	Sock	Jersey	Lighthouse
Slide 3	Matches	Candle	Lightbulb	Radio	Star
Slide 4	Pillow	Bed	Chair	Stool	Flag
Slide 5	Snow	Igloo	Hut	House	Sunshade
Slide 6	Watch	Arm	Leg	Neck	Tortoise
Slide 7	Nun	Church	School	Factory	Skate
Slide 8	Tent	Fire	Torch	Rocket	Picture
Slide 9	Mask	Clown	Ballerina	Priest	Sheep
Slide 10	Flower	Watering Can	Bucket	Shower	Anchor

As the focus of a semantic memory assessment is receptive language processing, a non-verbal response is used throughout the assessment. Participants use hand gestures to identify the correct response and progress through the assessment. As part of the clinical assessment, participants also receive positive re-enforcement for providing a correct response. As demonstrated within the documentation of the CAT, it is the SLT who must provide these verbal responses. Examples of such include “that’s right, it’s the monkey, because monkeys eat bananas” [69]. To emulate the presence of a SLT within the virtual SLT assessment, an audio sample was triggered in response to progression.

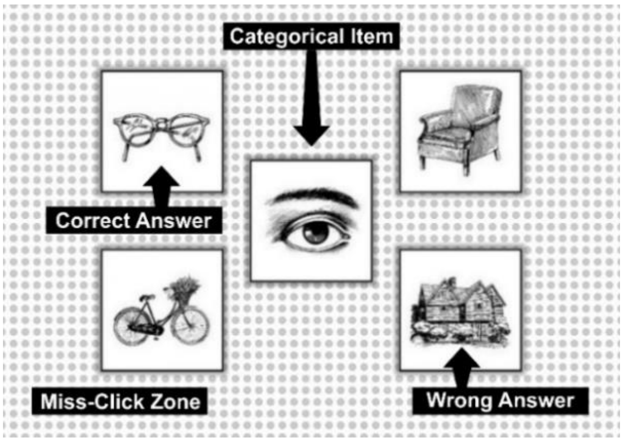


Figure 5.4: Comprehensive Aphasia Test – Semantic Memory Assessment

5.2.3 *Quality of Experience Assessment Strategies*

Derived from the QoE methodologies explored in Chapter 3, a QoE assessment was developed to capture explicit, implicit and interaction measures throughout the experiment.

5.2.3.1 Explicit Measures

Inspired by the QoE model categories (Figure 2.5), fourteen questions were designed to evaluate user perceived quality throughout the immersive multimedia experience. User response was captured using the absolute category rating (ACR) system as outlined in ITU-T P.913 [82]. Implementing a modified MOS five-point Likert scale, participants rated their level of agreement or disagreement with a question (Table 5.2). A reliability assessment was performed on the questionnaire to ensure clarity within the questionnaire. It was validated by 10 staff and students at AIT who took part in a pilot test.

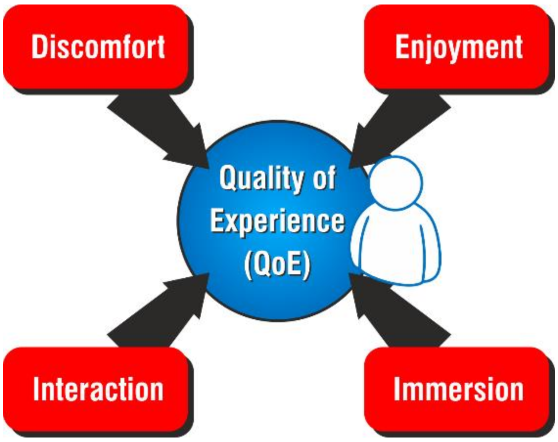


Figure 5.5: Immersive Multimedia Influence Factors

The questionnaire content focused the evaluation of four key components: discomfort, enjoyment, interaction, and immersion (Figure 5.5). It was inspired by [80] [141]. User *discomfort* when using a new technology can be an influencing factor on the user perceived QoE. As such, four questions were designed to evaluate feelings of nausea, user annoyance, and restriction in movement. The application of a new technology can often invoke a degree of excitement in users. Thus, three questions were created to analyse *enjoyment*, participant expectations, and interest in continued use of the respective systems. Three questions aimed to assess user perception of *interaction*. An emphasis was placed upon evaluating learning curve associated with each respective system and ease of use. Lastly, a series of four questions were created to evaluate user *immersion* through the multimedia experience. These queried the levels of engagement, activity immersion, environment realism, and sense of presence. The full content of the questionnaire can be found in Appendix 2.

Table 5.2: Modified 5 Point Likert Scale

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

5.2.3.2 Implicit Measures

1) Physiological Measures

In this research experiment, heart rate and electrodermal activity were recorded. Both of these physiological measures originate within the ANS [87] (Figure 3.5). The autonomic nature of these signals alleviates the influence of conscious decision making in the multimedia evaluation process. As such, it is believed that these measures often gather a more accurate measure of user emotion throughout an experience [27] [87] [142].

HR was monitored throughout the experiment using a Fitbit Charge HR [143]. The wireless activity tracker attaches to a user's wrist. Utilizing an optical sensor, the device monitors blood volume pulse. Internal Fitbit algorithms interpret this data into BPM at a rate of 1 Hz [143]. Recorded data is stored locally on the device and then synchronised to the Fitbit server using Bluetooth connectivity (via PC). The captured data can then be downloaded from the Fitbit website and processed offline.

A Personal Input Pod (PIP) [144] was employed to monitor changes in EDA. Held between the thumb and index finger the wireless biosensor samples EDA at a rate of 8 Hz [145]. In addition, the PIP uses internal signal processing to categorize emotion. Recognised events are correlated to states of stress, relaxation, or constant (neither stressed or relaxed). Data captured is streamed to the PIP Data Recorder [145] software package using Bluetooth and stored locally. Output files contain timestamped measures of EDA in microsiemens (µS or uS), in addition to the pre-calculated events (stressed or relaxed) and information of signal trend.

Although fluctuations in EDA can be correlated to changes in emotional arousal [146] [142], the measure alone does not specify which emotion is being elicited [142]. As such, the capture of EDA is often paired with other measures. In this work, the author pairs the implicit capture of physiological signals and novel capture of interaction measures (time to response, assessment performance, assessment errors, and miss-clicks) as a method to support the hypothesized emotional responses.

2) Interaction Measures

In a clinical setting, interaction with the semantic memory assessment is monitored by a clinician [69]. Observation focuses primarily on the progressive element of the assessment (i.e. did a user respond correctly or not). Outside of the restrictive nature of a paper-based medium, there are supplementary measures which may be captured to enhance the assessment procedure.

In this work, 4 different interaction metrics were captured: virtual assessment performance, progression time, assessment errors, and interaction error. *Virtual assessment performance* was calculated as per the CAT assessment strategies [69]. Each target which was correctly identified scored 1 mark. There are 11 slides in total however, the first slide is a practice slide, and thus a maximum score of 10 is possible. Stimulus interaction errors which occur are noted but do not alter the test score. *Progression time* was calculated based on the amount of time it took a user to respond after a stimulus was presented (as per Figure 5.6). User progression time, in a semantic memory assessment, in combination with a successful answer can highlight a delay in semantic processing. Thus, it extends the binary element of the assessment strategy to include other relevant data considerations. Therefore, highlighting a level of uncertainty within the presentation of stimulus, or the possibility of guesswork advancing the assessment. *Assessment errors* were noted when an interaction event was detected on the stimulus, but an incorrect semantic target was selected. An *interaction error* was recorded if an interaction gesture was made outside of the presented stimulus i.e. in the miss-click zone (refer to Figure 5.4).

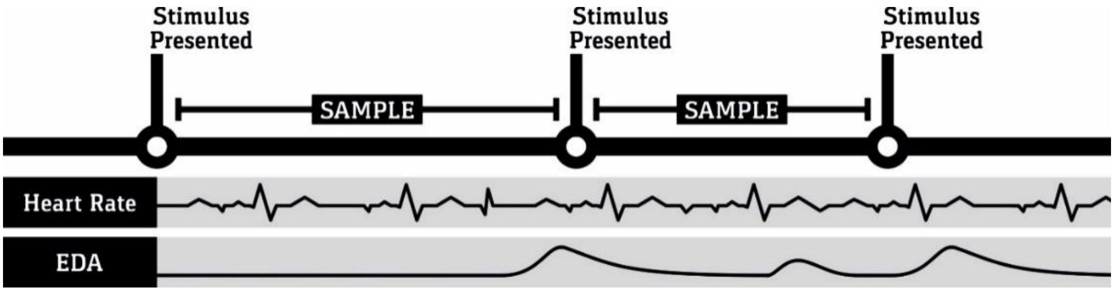


Figure 5.6: Implicit Measure Sampling Methodology

5.2.4 *QoE Assessment Protocol*

To ensure consistency throughout testing, a novel QoE assessment protocol was developed (Figure 5.7). Four key phases were included: (1) Information and Screening; (2) Resting; (3) Training; and (4) Testing. The phases were designed to capture baseline measures, introduce users to the new technologies, address “novelty” influences associated with the use of new technology through training, and finally perform the actual QoE evaluation. Completion of all phases required approximately 30-35 minutes. Typically, this included a 12-minute information and screening phase, 8 minute resting phase, 10 minute training phase, and a 5 minute testing phase.

Figure 5.7: Quality of Experience Assessment Protocol

5.2.4.1 Information & Screening Phase

Divided into two key stages, the Information and Screening phase aimed to inform and screen participants prior to undergoing the experimental process. During the information stage, an overview of the experimental process was provided by the researcher. Participants were given an information document (Appendix 3) and an opportunity was provided to ask any questions about the experimental process. To acknowledge an understanding of the experimental process, participants were required to provide written consent (Appendix 4) to continue.

Following this, participants underwent the screening process. The virtual SLT assessment aims to evaluate receptive language through the recognition of visual stimulus. As such, vision is a key component in the quality perception process. To prevent visual deficits impacting the perception of the experience, participants were asked to complete two vision assessments.

It is estimated that 8% of men, and 0.5% of women who have Northern European ancestry experience a restriction in their colour vision [147]. As the population sample was primarily from a European descent, colour perception was evaluated using an Ishihara Test [148]. Testing procedures required participants to identify a series of thirty-eight coloured plates [149]. Each plate contains a sequence of numbers or lines which are masked using multiple colours, an example of such can be seen in Figure 5.8. As per the evaluation guidelines, no more than four errors were permitted to progress.

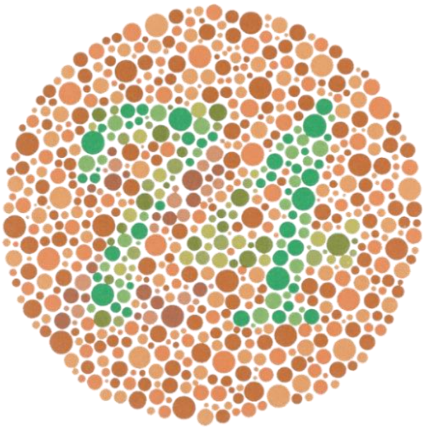


Figure 5.8: Ishihara Colour Perception Test – Example Stimulus

Following this, visual acuity was evaluated using a Snellen Test [150]. The evaluation process requires participants to stand 10 feet away from the Snellen chart (Figure 5.9, adapted from [151]). The content of the chart must be read aloud from left to right, progressing from the larger letters located at the top, to the smaller letters at the bottom. Evaluation occurs on a per eye basis, one eye is covered while they use the other to read the chart. Once this is complete, participants switch eyes and read the chart again. As each of the respective HMD devices accommodate glasses and wearing of prescription lenses was allowed during testing. A score of 20/20 was required to pass the test.

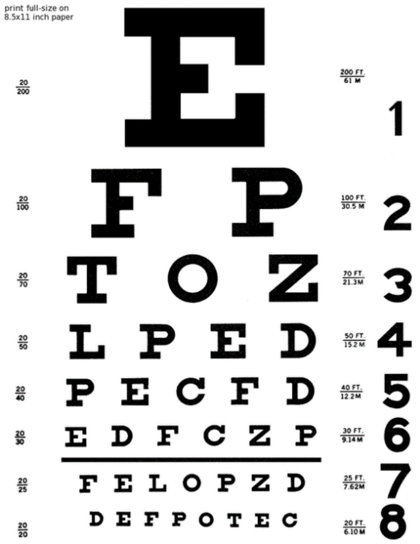


Figure 5.9: Snellen Test

5.2.4.2 Resting Phase

In this experiment, HR and EDA were monitored using consumer available devices. HR was monitored using the Fitbit Charge HR [143] wireless activity tracker. Improper wrist placement or inappropriate fitting has the potential to impact the Fitbit's blood volume data readings. Thus, as per the device settings, participants were asked to wear the wireless activity tracker on their non-dominant hand. In addition to this, the device was placed a finger width above the wrist bone as per the manufacturer's recommendations [152]. The PIP biosensor [144] was used to capture measures of EDA. The non-invasive wireless Bluetooth device is held between the thumb and index finger. While not a requirement, it was requested that the PIP be held in the participant's non-dominant hand. The purpose of this was to free up one hand for interaction only.

Due to the independent variability associated with of each of the physiological responses. Baseline measures of heart rate and electrodermal activity were captured over a 5-minute period. During this period, participants were asked to keep movements to a minimum. The objective of this procedure was to provide participants the opportunity to rest prior to the training, and test phases. Additionally, these baseline measures were used for comparative analysis.

5.2.4.3 Training Phase

To introduce the virtual SLT assessment and develop an understanding of interaction with the technology, a series of training videos and a training scenario were developed. The content of the training videos included: (a) an introduction to the assessment arrangement; (b) assessment progression; (c) and system interaction. An opportunity for questions was provided to participants upon viewing of the three training videos. Following this, participants were introduced to the technology. To ensure a comfortable fit, the AR and VR groups were provided opportunity to adjust the HMD. For the tablet group, the device was placed central to the user's field of view. To ensure comfort for participant, they were allowed reposition the tablet at the start of the training phase. This adjustment period occurred prior to the training assessment being launched for all three groups.

The training assessment was developed to normalise testing conditions in terms of assessment interaction; progression; and the use of technology. The training assessment mirrored the layout of the real assessment, however instead of matching images, participants were simply asked to match colours (Figure 5.10). This approach validated the learning outcome of the training videos, in addition to ensuring users had developed and demonstrated a level of understanding with respect to interaction. Thus, ensuring test performance was not influenced by such factors. As part of the protocol, participants were provided opportunity to repeat the training assessment if they encountered an error rate of more than 50%. This degree of error was not experienced by any participant.

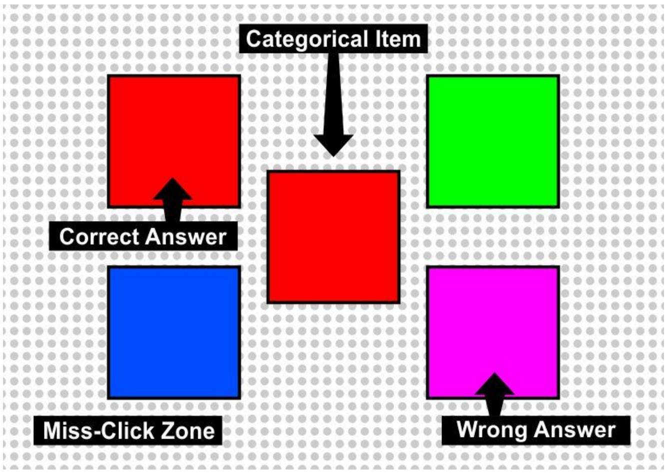


Figure 5.10: Example of Virtual Training Exercise

5.2.4.4 *Testing Phase*

Participants continued to wear the physiological measure devices throughout the testing phase. The phase consisted of two key components: assessment and the post experience questionnaire rating process. First, participants interacted with the virtual SLT assessment. Adhering to the clinical guidelines outlined within the CAT, participants interacted with the virtual assessment as discussed in section 5.2.2. Throughout the quality assessment, interaction measures (outlined in 5.2.3 Part 2 (b)) were captured for the purpose of post-processing. Once complete, the immersive multimedia device, and physiological capture devices were removed.

Lastly, participants were asked to complete the post-evaluation questionnaire (Full questionnaire available in Appendix 2). The 14 questions evaluated user QoE through analysis of discomfort, immersion, enjoyment, and interaction as per section 5.2.3 (Part 1).

5.2.5 *Test Environment*

The test environment was identical for each of the respective groups. Inspired by the guidelines outlined in ISO 8589:2007 [153], the test room (Figure 5.11) was designed with a focus on sensory analysis. Due to the nature of a VR experience, a high degree of immersion can often disengage users from the real world. Often users describe the phenomenon of "forgetting" where they were in the real world after prolonged exposure to a VR environment. To prevent such influences, the VR group experienced the virtual SLT assessment, within a 1:1 scale replica of the real lab environment. This ensured consistency in the experiment protocol across all three groups.

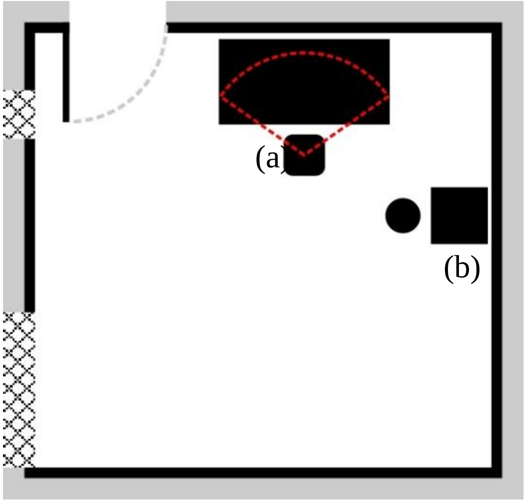


Figure 5.11: Test Environment Layout

(a) Location of the assessor during testing - (b) Location of P.I. during testing

5.2.6 *Sample Population*

Utilizing a convenience sampling approach, 67 participants took part in the experiment. Due to a combination of screening (outlined in section 5.2.4 (Part 1)) or technical errors encountered during testing, 7 participants were omitted from the sample. Therefore, the findings of a total of 60 subjects for experiment 1 are presented. The participants were split evenly between the three test groups (AR, Tablet, and VR). Table 5.3 presents an overview of the population sample, highlighting the average age of participants and gender balance. Age can often play influence in ability to adapt to new and emerging technologies. As such, it can impact confidence associated with interaction. Figure 5.12 presents an overview of the sample population, each of the groups have been broken down into three age brackets: 18-30, 30-40, and 40-50. The data reveals that the majority of participants were aged 18-30 for each respective group.

According to [65], previous experience with technologies has the potential to impact the perceptual evaluation process. As such, participants were asked if they had any prior experience with each respective technology. The AR group revealed some familiarity with mobile AR applications in the gaming industry due to the prevalence of Pokémon Go [154]. Although seven participants had experienced mobile AR, none had any prior experience with an AR HMD. Six participants within the VR group had prior experience with a VR HMD. However, once again it was noted that these were mobile platforms (i.e. non-interactive visual experiences). Lastly, all participants within the tablet group had a high level of familiarization with the functionality of touch-screen devices (due to the advent of touch screen mobile phones).

Table 5.3: Participant Sample

		Age		**Gender (%)**	
	Total	**Average**	**SD**	**Male**	**Female**
AR	20	25.6	7.059	75%	25%
Tablet	20	28.3	7.349	70%	30%
VR	20	29	10.105	70%	30%

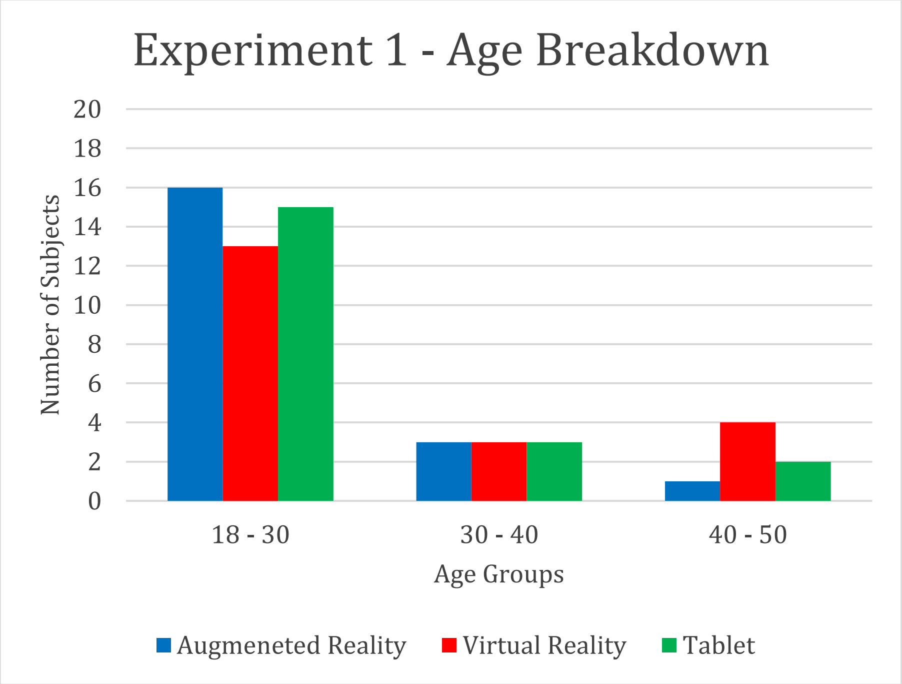


Figure 5.12: Experiment 1 – Age Groups

5.3 Results

In section 5.3.1, the results of the explicit, implicit, and interaction measures are presented. A statistical analysis was carried out using IBM SPSS [155]. A multivariate analysis of variance (MANOVA) was performed using a confidence level of 95%. The full output of this analysis is presented in Appendix 5, 6, 7, and 8. Sections 5.3.1 and 5.3.2 report the key findings and discussion of the study in a concise manner.

5.3.1 *Explicit Measures*

The results of the post-test questionnaire are presented in Table 5.4. The initial discussion places an emphasis on evaluating MOS within the context of the hypothesized influence factors (Figure 5.5) as mentioned in 5.2.3. This is then followed by the statistical analysis. To provide a clearer overview of the MOS ratings, the responses have been grouped into categories of: discomfort, immersion, enjoyment, and interaction.

Measuring levels of discomfort, the MOS ratings of questions 1, 4, and 8 are favourable towards the Tablet group i.e. tablet users report that (Q1) they experienced a greater level of comfort, (Q4) they found the device least annoying, and (Q8) were less restricted in their movements. In comparison to the results of the AR and VR groups, this was somewhat expected. More specifically, it is believed that the outcome is an influence of the tablet group not having to wear a multimedia display device on their head.

Table 5.4: MOS Ratings Captured Post-Experience

		AR		**Tablet**		**VR**			
		MOS	SD	MOS	SD	MOS	SD	F	Sig
Discomfort	Q1	3.900	1.021	4.450	1.099	3.550	1.468	1.475	0.238
	Q4	2.500	1.000	1.700	0.923	2.000	0.858	3.221	0.048
	Q8	2.550	0.887	1.900	1.119	2.400	0.995	1.739	0.185
	Q9	1.150	0.366	1.250	0.550	1.300	0.470	0.103	0.902
Immersion	Q2	4.550	0.510	4.650	0.587	4.600	0.503	0.016	0.984
	Q12	4.500	0.513	3.900	0.788	4.050	0.605	5.531	0.007
	Q14	2.300	0.865	1.900	1.021	1.850	0.813	1.648	0.202
Interaction	Q3	3.650	0.988	4.450	0.686	4.050	0.945	1.624	0.207
	Q7	3.350	1.663	3.050	1.761	3.350	1.599	0.074	0.929
	Q10	4.600	0.598	4.950	0.224	4.500	0.607	4.213	0.020
Enjoyment	Q5	4.350	0.587	4.150	0.813	4.350	0.489	1.223	0.302
	Q6	4.700	0.470	4.500	0.513	4.600	0.503	1.427	0.249
	Q11	3.450	1.146	3.900	1.119	3.900	0.641	0.067	0.935
	Q13	2.650	0.875	3.050	1.317	3.300	1.174	2.256	0.115

Levels of immersion were evaluated through Questions 5, 6, 11, and 13. Although only marginally, the results reveal that a higher level of immersion was experienced within the VR and AR groups. This trend is interesting, as it would agree with the marketing associated with immersive HMD technologies.

Enjoyment was evaluated through questions 2, 12, and 14. In question 2 users were asked to gauge the level of enjoyment throughout the overall experience. The AR group reported a MOS of 4.55, the VR group reported a MOS of 4.6, and the tablet group reporting a MOS of 4.65. With only a small deviation in the reported MOS, it can be said that the overall experience of an immersive SLT application was enjoyable to all groups.

Questions 3, 7, and 10 provided an overview of user interaction. The results reveal that two out of the three questions (Q3 and Q10) were favourable towards the tablet group. In contrast to the novel interaction measures associated with AR and VR experiences, touch screen interaction is not a new concept. As such, the result that tablet users report a higher level of QoE with these respective questions is not surprising. Despite AR and VR experiments implementing a novel method of interaction, MOS ratings for Q3 and Q10 do not differ much for the tablet group. This finding is of interest, as it reveals a level of ease associated with the use of hand gestures as method of human computer interaction.

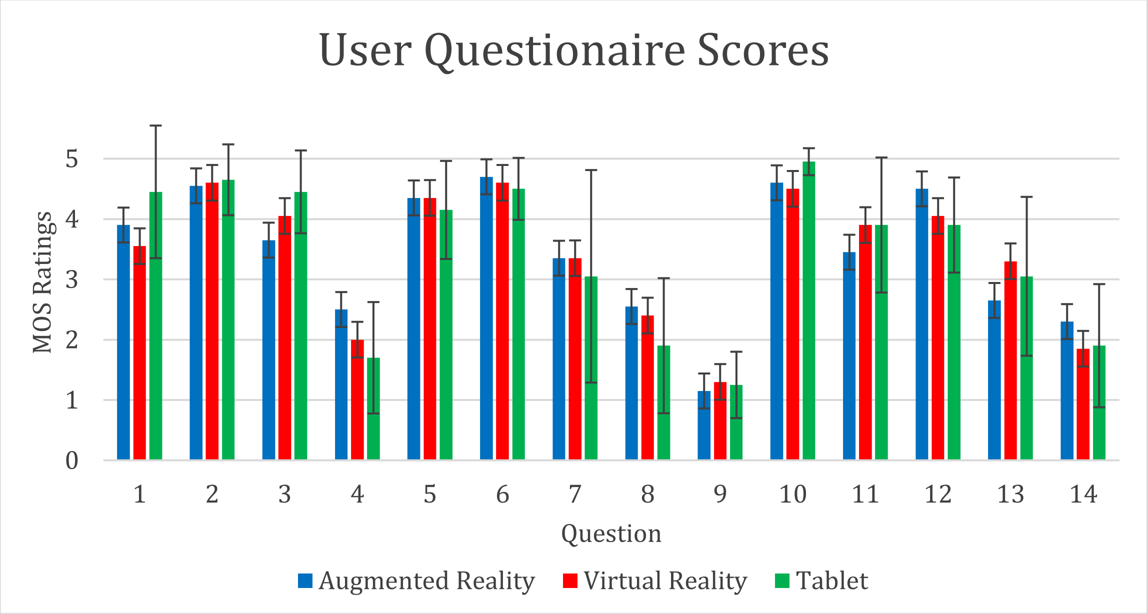


Figure 5.13: Post-Test Questionnaire – MOS Ratings

5.3.1.1 Augmented Reality vs Tablet

A statistical analysis of the results between the AR and tablet group reveals two statistically significant questions. In question 4, participants were asked to rate the levels of annoyance associated with each of the technologies (lower is better). The tablet group revealed a lower MOS of 1.7 when compared to the AR group reported MOS of 2.5. The result was statistically significant (p=0.022). Influential factors such as the wearing of a head mounted multimedia device, when compared to a tabletop device may have been a factor in the outcome.

Question 12 aimed to evaluate if participants would like to experience the environment again. The variation in the MOS between the groups was again statistically significant (p=0.003). The AR group provided a higher MOS of 4.5 when compared to the tablet group rating of 3.9. It is likely that the novel delivery of the AR experience, in terms of real-world projected content when compared to a traditional screen had an impact on the users response here.

5.3.1.2 Tablet vs Virtual Reality

Analysis of the tablet and VR groups reveal one statistically significant question. Considering the influencing factors associated with “interaction”, question 10 (p=0.003) asked users to rate how easy the system was to use. The tablet group provided a higher overall MOS of 4.95 when compared to that of the VR group (MOS 4.5). Through ad-hoc discussion post experience, participants reported a low level of familiarity with human-computer gesture-based interaction as experienced by the VR group. Therefore, the novel interaction mechanics may have been an influence of this outcome.

Overall, in terms of interaction with a virtual content, the VR group appeared to enjoy the novel experience of using hand gestures. Through ad-hoc discussion after testing had been complete, participants often highlighted the preference of using hand gestures as opposed to traditional interfaces such as controllers, or a keyboard and mouse. This is further highlighted within their reported MOS of 4.5 for question 10.

5.3.1.3 Virtual Reality vs Augmented Reality

Comparison of the AR and VR group revealed two statistically significant questions. Question 12 asked users if they would like to experience the environment again. The statistical analysis revealed a borderline significant result of p=0.05. This is reflected in the user rating which seen the AR group provide a MOS of 4.5 when compared to that of the VR group (MOS 4.05).

Question 13 asked participants if they did not feel a strong sense of presence whilst experiencing the system. Prior to the experiment it was hypothesized that the VR group would have a greater degree of presence due to being transported to a virtual environment. However, taking a further look at the MOS, it is revealed that the AR group felt more presence within the virtual assessment with a MOS rating of 2.65 when compared to that of the VR group (MOS 3.3). This difference was statistically significant (p=0.030), although the SD of 1.21 in the VR group reveals a level of disagreement between participants.

5.3.2 *Implicit Measures*

5.3.2.1 Electrodermal Activity

Figure 5.14 presents an overview of the average increase in EDA during the assessment expressed as a percent of change (from the baseline). 0% is representative of each group’s respective baseline value recorded during the resting phase. The graphed lines reflect the increase in EDA for each group during the test. Visual analysis of Figure 5.14 reveals similarities between the AR and VR groups. The AR group experienced an average increase of 24% and the VR group experienced an average increase of 23% from the baseline values. Overall, the AR group appears to experience a downward sloping trend in response. This is indicative of the AR group becoming accustomed to the use of the immersive multimedia application. Similar to this, the VR group experience minimal change in signal throughout the assessment. However, average signal values increase over the course of the activity and more specifically during interaction with content from slide 6-7 onwards. As described in 3.2.2.3 (Part 2) and 4.2, this slow and steady change is indicative of an SCL response in EDA.

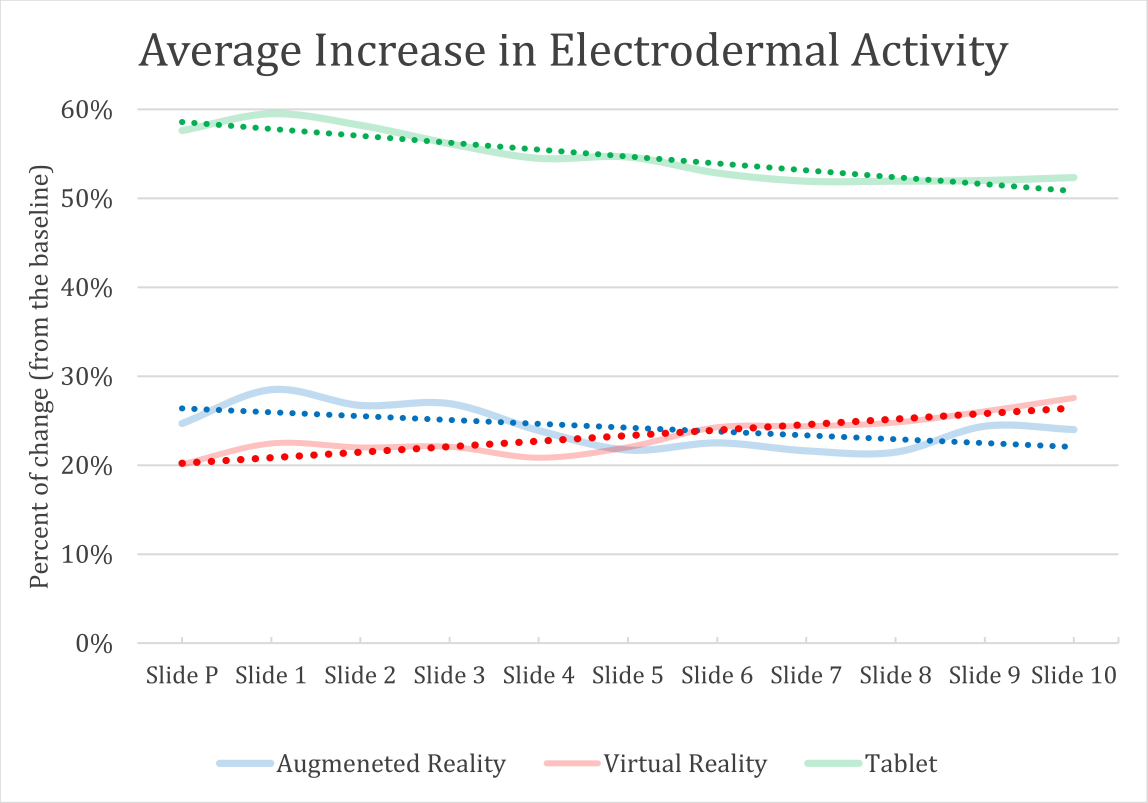


Figure 5.14: Average Increase in EDA Throughout the Assessment

Previous research has indicated that these types of changes can be correlated with an increase in cognitive activity [97] [101]. As outlined in 3.1.1.1, semantic assessments aim to challenge the brains ability to access semantic and conceptual information from images or words. Experiencing difficulty with such a task should elicit an increase in cognitive activity. It's believed that the small change noted here, highlights a difficulty in semantic categorization experienced by the VR group. Further evidence of this can be seen in 5.3.3 which explores progression time and assessment performance throughout the assessment.

Unlike the AR and VR group who experienced similarities in skin conductivity, the tablet group reported a much larger increase of 55%. As a result of this, an investigation was carried out to explore the possibility of 3rd party influences. The influence of environmental factors such as weather, and the impact of capacitive touch screen technology were considered. Exploring each of these avenues yielded similar results. Thus, it was validated that the measure of EDA within the tablet group was not influenced by external factors.

Evaluating the signal trend of the tablet group reveals similarities in the signal slope of the AR group. Both AR and tablet groups experience a downward slope in EDA. This is interesting as each of the respective groups also experience the virtual content presented within their real world surroundings. It's assumed that this is an indicator of the two groups experiencing an initial level of physiological arousal associated with interaction of a new experience. The increased levels of arousal are then followed by a slow and steady decline as they become acclimatized to the virtual content.

Table 5.5 presents a statistical analysis of EDA carried out between all three groups. All results at interaction points during the test are significant. Following on from this, a comparative analysis between each group was performed to explore the possible reasons. Comparing the AR and tablet groups, it is revealed that 63.63% of interaction points remain statistically significant. Similarly, a comparison between the tablet and VR groups reveals that 100% of interaction points are significant. In a way, this high rate of significance was expected. The key differences in display technology in terms of the presentation on a head mounted display, versus a table-top display would have been a key factor.

Lastly, a comparison between the AR and VR group reveals no statistically significant differences between the groups. This finding is again interesting as it reveals that, in terms of emotional arousal, both AR and VR groups experienced a similar physiological response throughout the evaluation.

Table 5.5: MANOVA of EDA at a 95% confidence level

	AR		**Tablet**		**VR**			
	EDA	**SD**	**EDA**	**SD**	**EDA**	**SD**	**F**	**Sig.**
Baseline	3.142	1.294	2.997	1.128	2.896	1.205	0.617	0.543
Slide P	3.584	1.054	4.539	1.448	3.278	1.148	6.252	0.004
Slide 1	3.691	1.095	4.579	1.432	3.297	1.105	6.750	0.002
Slide 2	3.656	1.143	4.551	1.418	3.293	1.115	6.524	0.003
Slide 3	3.650	1.151	4.498	1.434	3.301	1.120	5.928	0.005
Slide 4	3.616	1.181	4.465	1.444	3.291	1.134	5.795	0.005
Slide 5	3.570	1.192	4.482	1.481	3.300	1.124	5.754	0.005
Slide 6	3.589	1.202	4.424	1.425	3.339	1.133	4.850	0.012
Slide 7	3.575	1.241	4.389	1.405	3.371	1.165	4.282	0.019
Slide 8	3.561	1.246	4.408	1.437	3.360	1.157	4.714	0.013
Slide 9	3.621	1.186	4.402	1.432	3.398	1.183	4.499	0.016
Slide 10	3.618	1.177	4.414	1.449	3.422	1.173	4.374	0.017

5.3.2.2 Heart Rate

In Table 5.6, a comparative analysis of HR for each of the respective groups is presented. The analysis reveals no statistically significant differences between the AR, VR and tablet groups. Figure 5.15 presents a visual overview of the data. Although diverse, all three groups follow a similar trend. Initial signals appear relatively in line with their average baseline. This is followed by an increase which is shared across groups. Despite this trend, the percent of change which occurs is varied. Minimal change is seen within the tablet group as they experience a 1.8% in signal variability. The VR group experienced a 3.33% in signal variability. Lastly, the AR group experiences the highest signal variability with a rate of 5.7%.

Table 5.6: MANOVA of Heart Rate at a 95% confidence level

	AR		**Tablet**		**VR**			
	HR	SD	HR	SD	HR	SD	F	Sig.
Baseline	77.845	13.000	78.454	8.169	78.860	9.987	0.150	0.861
Slide P	75.526	13.775	79.663	10.506	78.225	9.866	1.193	0.311
Slide 1	75.764	13.600	80.621	10.978	78.380	9.859	1.353	0.267
Slide 2	76.195	13.414	80.770	10.455	78.960	9.843	1.105	0.338
Slide 3	76.880	13.397	80.439	10.493	78.945	9.627	0.627	0.538
Slide 4	77.641	13.465	80.325	10.637	79.446	9.624	0.359	0.700
Slide 5	77.714	13.471	80.201	10.663	79.348	9.378	0.359	0.700
Slide 6	78.177	13.714	79.942	10.654	79.387	9.323	0.301	0.741
Slide 7	78.779	14.212	79.902	10.633	79.454	9.050	0.234	0.792
Slide 8	78.899	14.210	80.116	10.485	80.033	9.360	0.305	0.738
Slide 9	79.900	13.768	81.152	13.200	80.615	9.726	0.228	0.797
Slide 10	79.899	13.854	80.512	11.970	80.694	9.840	0.191	0.827

Surprisingly, both the AR and VR group HR’s begin below their baseline thresholds. Often the use of a new technology can create high levels of anticipation. This increase has the potential to raise the user’s HR above the normal. In addition, we must consider the impact of positive reinforcement. As participants progressed through the activity, a virtual SLP responded to each correct answer with a positive comment. These comments from the virtual SLT may have invoked a level of enjoyment or satisfaction as a result of performing well in the assessment. Lastly, it’s important to note, that due to the gesture based interaction element of the assessment. Subtle movement has the potential to cause a deviation in heart rate as the blood circulates around the body. This may also explain the lack of statistical significance between the groups as all groups would have experienced similar interaction measures.

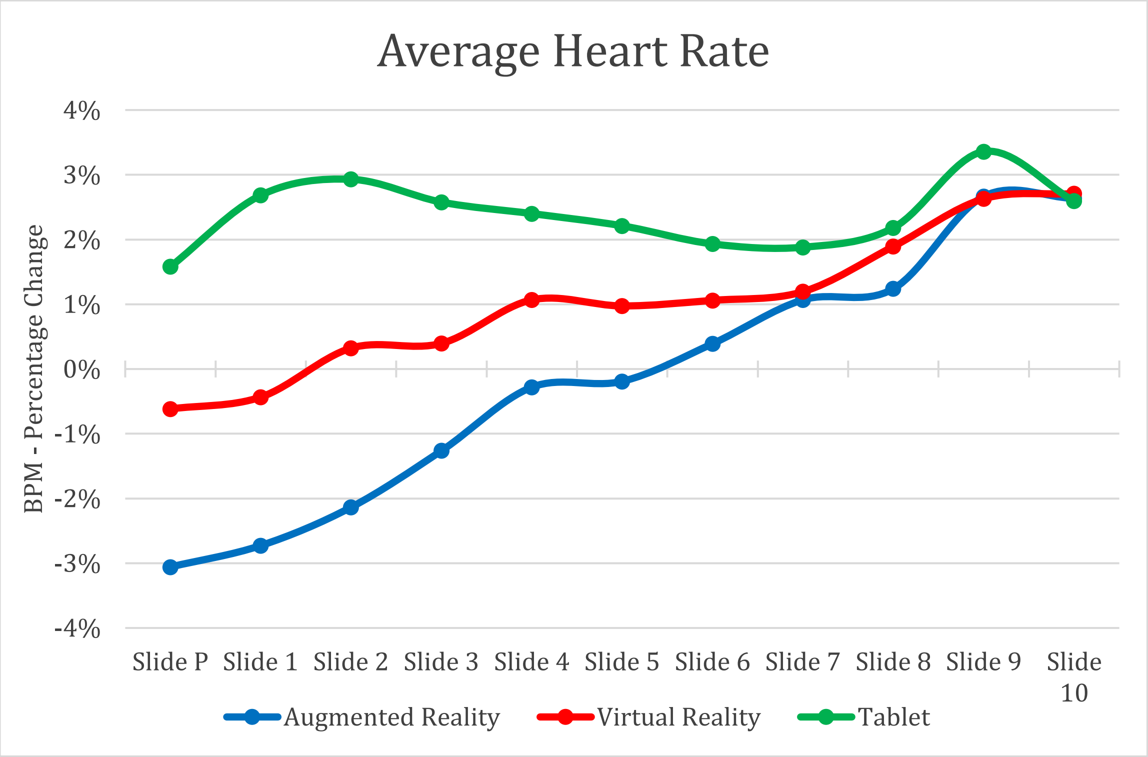


Figure 5.15: Average Increase in Heart Rate Throughout the Virtual Assessment

NB: 0% is representative of each respective group baseline value.

5.3.3 *Interaction Measures*

5.3.3.1 Assessment Error

In Table 5.7, a summary of the number of assessment errors encountered by each respective group is presented. The VR group experienced the highest level of errors (23 in total), followed by the AR group (6 in total), and the tablet group (6 in total). The table also reveals a trend in the number of assessment errors for slide 9. Each of the respective groups experienced the highest rate of error during this part of the assessment.

Table 5.7: Average Number of Assessment Errors

Slide	**P**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total Errors**
Augmented Reality	0	1	0	0	0	0	0	0	0	3	2	6
Tablet	0	0	0	0	0	0	0	1	0	5	0	6
Virtual Reality	3	1	4	4	1	1	0	1	1	7	0	23

To understand this trend, we must first acknowledge that the normative data as provided within the CAT, highlights that the typically developing population may experience a certain degree of error. Beyond this, alternative factors such as the content, and influence of the system design must be considered. More specifically, slide 9 (Figure 5.16) requested users to formulate a link between a humorous mask (central image) and a clown (target image), alternative selections (distractors) include a priest, sheep, and ballerina. The majority of errors encountered were as a result of an incorrect match with the ballerina. While not traditional, ballerinas can be depicted wearing face masks. However, the humorous (large nose) aspect of the stimulus would provide a more valid link with a clown.

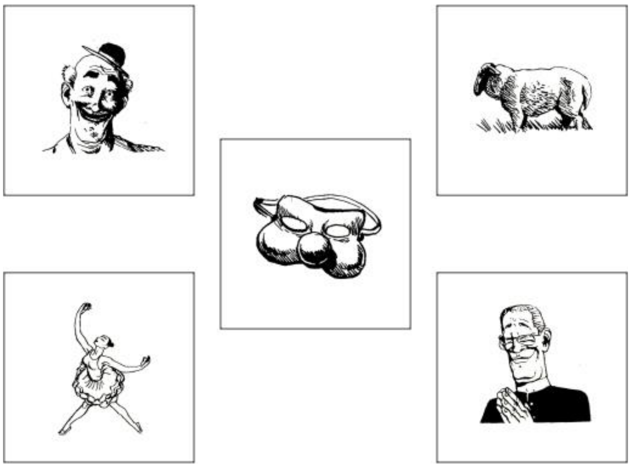


Figure 5.16: Comprehensive Aphasia Test – Slide 9

Open ended discussion post-test revealed some interesting feedback. Participants often described a difficulty associated with interpretating the finer details of the stimulus in slide 9. Traditionally, these images are presented on a paper-based format which is no larger than A4. Each of the respective technologies used aimed to create a similar experience in terms of the size of the presented stimulus. However, influencing factors in these errors may be a result of limitations with respect to each of the respective devices display technologies.

For example, VR experiences are often hindered by visual artifacts as a result of the display technology being presented close to the users vision, this is known as the screen door effect [156] [157]. To a lesser extent, the AR results may have been influenced by the multi-layer display technology used to present holographic content to user's field of view on the Microsoft HoloLens. While display technology can be an influence in HMDs, this does not explain the high rate of error experienced by the tablet group. Through process of elimination, the combination of stimulus of slide 9 is without a doubt a key factor in this result. In addition, the outcome coincides with the performance levels expected from the normative population as outlined within the CAT.

5.3.3.2 Comprehensive Aphasia Test Performance

Accompanying the CAT assessment manual [69], is a collection of normative data which serves as a guide for the clinician during assessment procedures. As per these guidelines, it is expected that the typically developing population score a mean value of 9.81 with a standard deviation of 0.4. As mentioned previously, this scoring mechanism does not factor in the practice slide (slide P). Therefore, a total score of 10 is attainable.

In Table 5.8, the mean performance for each of the respective groups is presented. The mean performance of the AR (9.7) and tablet (9.7) group fell within one standard deviation of the normative data. The VR group performs within two standard deviations of the normative data with a score of 9.0. These results are favourable towards the AR and tablet group. However, VR should not be ruled out as a restriction in the current display technology may have been a key influential factor within these results.

Table 5.8: Comprehensive Aphasia Test Performance Metrics

	Augmented Reality		**Tablet**		**Virtual Reality**			
	Average	***SD***	***Average***	***SD***	***Average***	***SD***	***F***	***Sig.***
Assessment Performance	9.7	0.923	9.7	0.470	9	1.29	4.983	0.010

5.3.3.3 Progression Time

Figure 5.17 provides an overview of each of the respective group’s progression time. Measures are provided on an interaction basis highlighting the adaptation rate as users progressed throughout the assessment. Whilst there are varied levels of interaction between the groups, there is a strong trend throughout the measures. From slides 7 through 9 there is a clear increase in progression times for all the groups. This increase can be attributed to the levels of error associated with interaction which are further discussed in 5.3.3 (Part 1).

Both the AR and tablet groups reported a trend in terms of average progression time. This similarity may coincide with the retained real-world connection associated with both respective technologies. Interestingly, the AR group experiences an increased delay during interaction with Slide 1. This, however, can be attributed to the interaction of an individual subject who took some time to formulate a link within the content.

Excluding slide 1, the VR group reported the slowest overall progression time. Progression time deviated up and down from slides P through 5. However, as we draw closer to slide 9 a steady increase is revealed. As per 5.3.3 (Part 1), when participants interacted with slide 9 they experienced a high volume of errors. Thus, it is believed that the difficulty users encountered with Slide 9 correlate to the prolonged delay in progression time.

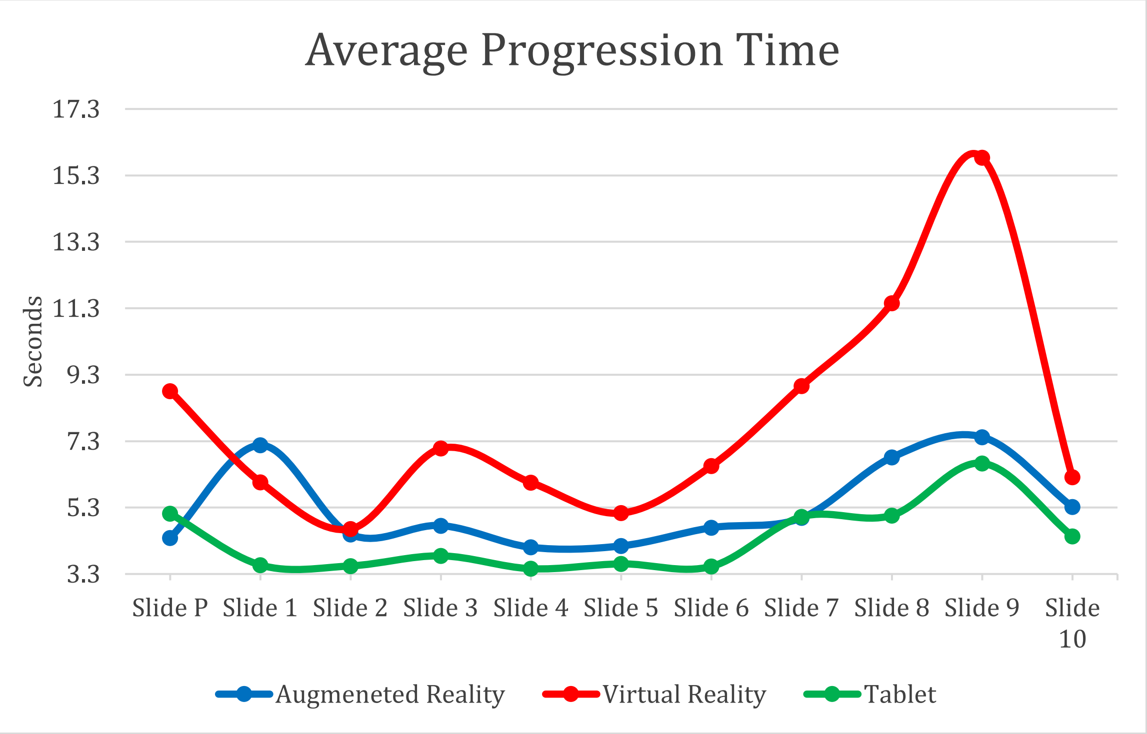


Figure 5.17: Average User Progression Time During the Test Phase

5.3.3.4 Interaction Error

Miss-clicks were recorded as the user interacted with non-assessment content (i.e. they did not correctly select one of the presented stimuli). Table 5.9 presents the total number of miss-clicks encountered by each group throughout the testing phase. Analysis of the data reveals that the VR group experienced a greater degree of interaction error when compared to the AR and tablet groups. A total of 14 errors were encountered by the VR group. Interestingly, most of these errors occurred later within the assessment with a peak of 4 errors encountered on slide 6.

In comparison, the AR group experienced a total of 4 errors which are evenly distributed across 4 of the stimuli. This result is interesting as it would have been expected that the AR and VR groups experience similar levels of error given that the interaction measures mirror each other. It again leads to the question of, if the display technologies in the VR HMD could have led to increased rates of error within the VR group.

The tablet group experienced no errors during the virtual SLT assessment. Given that all participants were accustomed to interaction measures for touch screen devices (i.e. mobile phones), this was expected.

Table 5.9: Average Number of Error Rates

Slide	**P**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**Total**
Augmented Reality	1	0	0	1	1	0	1	0	0	0	0	4
Tablet	0	0	0	0	0	0	0	0	0	0	0	0
Virtual Reality	0	1	0	1	0	1	4	2	2	2	1	14

5.4 Discussion & Conclusion

In this experiment, three immersive multimedia SLT applications were evaluated and compared. A QoE evaluation was performed in which explicit and implicit measures were captured. User ratings were captured via post experience questionnaire. Implicit measures in the form of HR and EDA were observed as an insightful measure of physiological and psychological arousal. In addition, the novel capture of interaction measures was explored as a method of quantifying user perceived quality throughout the immersive experiences. Analysis of the interaction measures delivered key insights into user performance throughout the assessment. More specifically, they quantify the adaptability of a user to a system, in addition to highlighting hesitation or delay in response throughout an immersive multimedia SLT assessment.

Results highlight the utility of monitoring interaction and physiological measures throughout an interactive and immersive SLT assessment. Multiple similarities are discovered within the AR and tablet groups in terms of: EDA, HR, progression time, and assessment performance. This finding is interesting as both groups experience multimedia content whilst remaining in the context of a real-world setting. While there are similarities within the VR group, they occurred less often. Instead, some unique observations are noted across the objective measures. More specifically, an elevation in EDA, increase in the number incorrect answers, and delay in progression time can be correlated to an increase in cognitive activity. It is believed the AR and tablet groups did not experience this outcome as they encountered a lower degree of errors throughout the assessment. Thus, there was a lack of physiological and psychological response.

Generally, the results reveal a great degree of suitability for AR and tablet as an alternative method for the delivery of a speech and language assessment. The novel capture of implicit metrics explored here also have the potential to inform the clinician beyond what is possible with paper based assessments. Thus, providing opportunity to make a more informed decision in the development of a treatment strategy.

Although the results reveal a greater degree of QoE experienced within the AR and tablet groups, VR should not be ruled out in the delivery of an SLT assessment strategy. The increase in progression time, and number of errors suggest that the limited capability of the current generation of VR HMD may be an influential factor. Limitations such as the display technology and lack of native gesture support may have influenced the results. It is possible that a newer VR HMD would alleviate some of these shortfalls, therefore bringing the QoE of the VR group more in line with the AR and tablet groups. Indeed, these questions inspired the next step of this PhD work as reported in Chapter 6.

Chapter 6: TASK-EVOKED PHYSIOLOGICAL RESPONSE

6.1 Aims and Objectives

The results outlined in section 5.3 highlight the utility of capturing and processing explicit, implicit, and objective metrics as a holistic approach to the understanding of user perceived QoE. In experiment 1, numerous similarities are reported across the AR and tablet groups. Both of which experience the immersive virtual assessment within the context of the real world. However, the VR results differ in terms of physiological response and objective performance raising a number of questions as outlined in 5.4.

In an attempt to address these questions, the work presented in this Chapters extends the already existing QoE evaluation presented in Chapter 5 with a second experimental study. The fundamental aim is to capture additional data in the form of eye tracking, in an effort to understand the physiological responses in EDA and differences in progression time for the VR group. It has been suggested that these anomalies have occurred due to the VR group experiencing the immersive virtual assessment on hardware which does not represent the current state-of-the-art. Thus, the results of a newer VR HMD (HTC Vive) are presented within this Chapter in an effort to understand the impact of system factors. In parallel, alternative measures of physiological arousal in the form of oculomotor response is explored as a supporting measure of EDA. The primary objectives are summarised as follows:

- Explore the utility of capturing oculomotor response as a supporting measure of EDA in an effort to further understand states of physiological arousal.
- Gain a fundamental understanding of user perception of the visual stimulus through the analysis of oculomotor movement.
- Identify challenging stimulus combinations in an effort to further understand user errors reported in the first experiment (Chapter 5).
- Validate the previously reported measures of physiological signals (HR & EDA) utilizing medical grade sensors.

The results of his Chapter address SubRQ3 and SubRQ4 as per section 1.2.1.

6.2 Methodology

The system design, inclusive of the novel immersive SLT application, stimuli, and interaction mechanics remain identical to that of experiment 1 (Chapter 5). The QoE methodology utilized in section 5.2 has been mostly preserved in experiment 2, apart from one primary change. The hardware used to capture the physiological responses in experiment 1 are marketed as consumer grade products. Thus, in this experiment, an Empatica E4 wristband was used to capture medical grade physiological responses of EDA, HR, and skin temperature. This decision was made in an effort to further validate previously found physiological responses in EDA associated with the VR group. All participants wore the Empatica E4 on their non dominant hand as per the device recommendations. This approach followed an identical protocol to the fitting of the previously used FitBit Charge HR. As a result, this change did not have any impact on the quality assessment protocol (Figure 6.1).

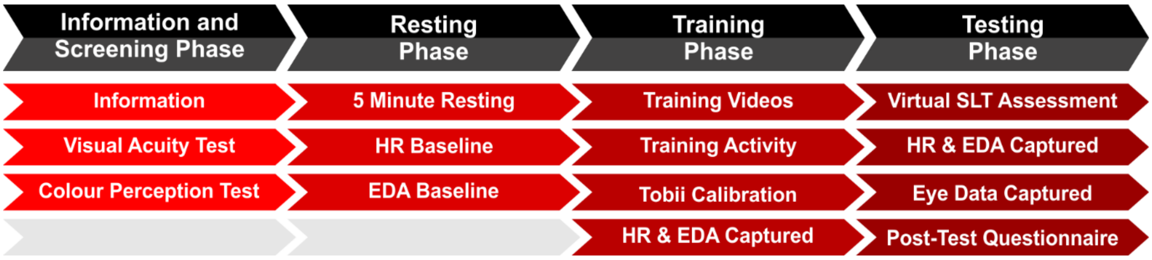


Figure 6.1: Quality Assessment Protocol – Tobii Calibration

Similarly, as the objective of this experiment is to explore eye tracking as a supportive measure for the interpretation of physiological response, the HMD was changed. A HTC Vive HMD modified by Tobii [106] was utilized in place of the Oculus DK2 HMD used in experiment 1. Beyond eye tracking, there are numerous benefits to the updated hardware (Table 6.1). From a display perspective, the HTC Vive provides an updated FOV, screen resolution, pixel density and refresh rate. Complementing this, the device also delivers a higher level of rotational and positional accuracy utilizing more recent approaches to VR HMD tracking. A comparison between the HTC Vive and previously used Oculus Rift DK2 (Chapter 5) is presented in Table 15.

Table 6.1: Oculus DK2 & HTC Vive (Tobii Pro) Comparison

	Oculus DK 2	**HTC Vive (Tobii Pro)**
Field of View	100°	110°
Resolution	1920 x 1080 (960 x 1080 per eye)	2160 x 1200 (1080 x 1200 per eye)
Pixel Density	386 PPI	455.63 PPI
Refresh Rate	60 Hz	90 Hz
Degrees of Freedom	6	6
Rotational Tracking	Gyroscope, Accelerometer, Magnetometer	Gyroscope, Accelerometer, Laser Position Sensor
Positional Tracking	Near Infrared CMOS Sensor	External Base Stations

Reflecting on Figure 6.1, there is one primary change added to the training phase. Similar to the capture of baseline metrics for physiological response, oculomotor response is relevant to the individual. To produce a data set with an accurate measure of oculomotor movement, the camera-based system needs be calibrated to each individual participant. This process was implemented using the standardised approach recommended by Tobii [158]. Calibration was carried out once the HMD was fitted as part of the training phase. The process requires users to focus on five red spheres which appear within the participants FOV. Only one of the spheres is visible at any given time. As a fixation event occurs, the two dimensional HMD co-ordinates (Figure 6.2 (Left), adapted from [159]) are captured and transposed to Unity three dimensional virtual world co-ordinates (Figure 6.2 (Right), adapted from [158]). This process allows the HMD to accurately report angular rotation of the eye within the virtual world. The process integrates in a seamless manner with the calibration occurring at the beginning of the training phase. Therefore, limiting the impact of luminance changes affecting pupillary size as both training and test environments contained identical lighting. On average, the calibration process added a mere 60 seconds to the QoE assessment protocol.

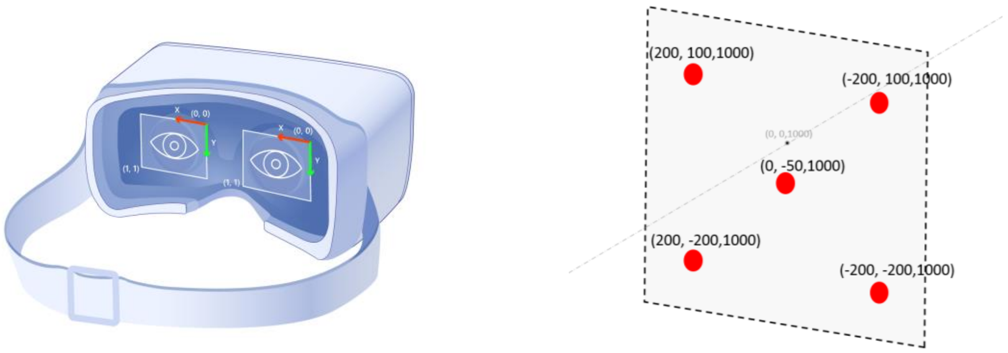


Figure 6.2: Tobii Pro HMD Calibration Points

(Left) Tobii HMD Eye Tracking Area, (Right) In-Game Calibration Process.

6.2.1 *Sample Population*

Similar to experiment 1, a convenience sampling approach was used to recruit new participants. 28 participants successfully completed the screening process and virtual assessment without experiencing difficulty or technical error. This included undergraduate and post-graduate students, in addition to members of the faculty and public. Table 6.2 presents an overview of the participant age and gender information. A total of 14 male and 14 female participants took part in the assessment. The primary age group for both male and female ranged between 18 – 30 with the average age of 26.14 (SD of 7.04 years). Nine participants (5 male and 4 female) reported previous VR experience, however once again these were non-interactive passive experiences. No participant had previous experience utilizing gesture-based interaction mechanics within a virtual space. This was to be expected. Such novel interaction mechanics are only recently emerging on consumer grade devices [160].

Table 6.2: Experiment 2 – Sample Population

		Age		**Age Groups**		
	Total	**Average**	**SD**	**18 – 30**	**31 – 40**	**40 – 60**
Male	14	24.43	6.42	12	1	1
Female	14	27.86	7.45	10	3	1
Total	28	26.14	7.04	22	4	2

6.2.2 *Oculomotor Data Capture*

The following section delivers an overview of the novel approach which was designed and developed to capture and process oculomotor data in the VR experience. This protocol (Figure 6.3) is divided into 4 stages. First, part one highlights the data capture process, reporting the objective and physiological measures used throughout. Following this, part two discusses the data pre-processing. Part 3 describes the novel approaches to data generation and reporting. Lastly, part 4 focuses on the presentation of the results which is described later in this Chapter (Section 6.3).

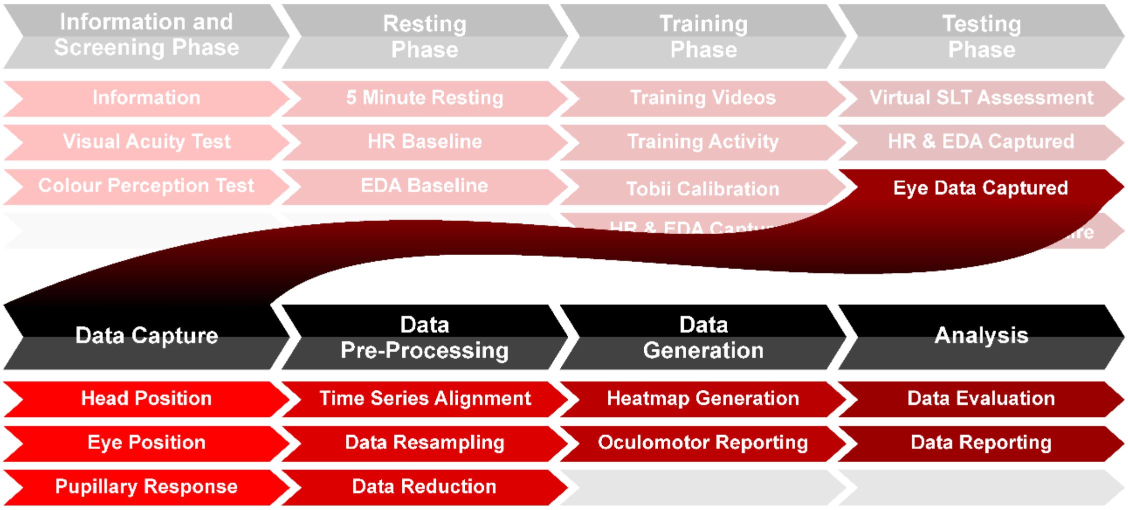


Figure 6.3: Oculomotor Data Capture & Processing Methodology

6.2.2.1 Data Capture

Tobii eye tracking hardware captures, processes, and reports eye measurements using a proprietary software (a premium option) [161] or through the Tobii Pro SDK [162]. A bespoke custom plugin was designed and developed using the Tobii Pro SDK in C# to capture and report head movement, eye position, and pupillary response throughout the experiment. All data available from the SDK was captured and recorded to 11 files, each of which contained on average 25 variables. The collection of files results in overlapping data, examples of such include files that report raw eye measurements from the HMD perspective (2 data points per eye). The same reported measures also exist as post-processed in-game variants of eye measurement (up to 9 data point per eye). Since only a small percentage of this data is relevant within the context of this experiment, and with clarity in mind, the focus and discussion is with respect to the key metrics only. A full breakdown of these primary data points is found in Table 6.3.

6.2.2.2 Data Pre-Processing

The data captured during step 2 of the methodology (Figure 6.3) using the Tobii Pro SDK produces two core files associated with (a) head movement and (b) oculomotor events. Each of these respective files can be synchronised using a proprietary approach by Tobii. During the development phase, the output of these files was updated to include human readable timestamps (as per Table 6.3). This was motivated by the requirement to synchronise all physiological and application interaction data at a later stage. Despite this effort, there was a minor variable drift (less than 0.02 seconds) noted post-recording. This is due to the fact that the eye tracking metrics are written to multiple files at once utilizing various data sources. For example, eye measurements are reliant on a hardware-based response from the HMD, whilst parameters associated with head movement are recorded within the Unity3D game engine. This hardware-software overlap has the potential to induce a minor delay due to how Unity3D operates utilizing a single core thread [163].

To resolve this, all eye tracking data was re-sampled to a rate of 60 Hz. This decision had multiple motivations: firstly, the minor challenges associated with time alignment were resolved; and secondly, the approach reduced the size of each participants dataset, aggregating only the primary parameters (as per Table 6.3) into a single readable file. This alleviated some minor computational limitations which would have been encountered during the data generation stage (see section 6.2.2.3).

Table 6.3: Eye Data Overview

Parameter	Purpose
Timestamp	A custom human readable timestamp was created to match an existing time format used throughout the experiment ("yyyy-MM-dd HH:mm:ss.FFFFF"), this allowed for data alignment within the application (i.e. interaction), and with other physiological sensors (Empatica E4).
Epoch/Unix	An Epoch/Unix timestamp is also included.
Pupil Diameter	A measure of pupil size (in millimetres), this parameter was captured for both the left and right eye.
Gaze Origin	Gaze origin reports the position of the left and right eye within 3D space. It should be noted that these values are relevant to the head pose position associated with the HMD in 3D space. Therefore, both are required for a complete dataset.
Gaze Direction	A 3D position representing the direction a user was looking at any given time, once again this parameter is complemented by the gaze origin as both work in synchronous.
Head Pose Position	A 3D co-ordinate representing the HMDs position in the virtual world.
Head Pose Rotation	The angular rotation (pitch, yaw, and roll) of the HMD within the virtual space.

6.2.2.3 Data Generation

1) Heatmap Generation

As part of the oculomotor analysis process, visual spatial maps were generated to deliver an overview of visual attention, more specifically a heatmap. There are numerous packages which exist to generate heatmaps. However, the majority of these are limited due to the requirements for proprietary hardware, associated premium costs, or lack of support for immersive virtual reality environments (limited to 2D multimedia experiences).

To circumvent these challenges, a novel custom Unity3D component was developed. Written in C#, the custom component took all eye tracking data (head movement and eye data) an input and interaction metrics for an individual participant. Each of the respective files were time-aligned within Unity3D utilizing the common timestamp. Once this occurred, playback of the user's data was possible. Positional, rotational, and distance metrics for head and eye movements are processed and interpreted to replay the test data captured for each individual participant.

This process is visualised in Figure 6.4, which displays a debug view in Unity3D. The image presents three individual lines moving from left (head and eye position) to right (immersive content). These lines are generated using a common approach in game development known as ray casting. Ray casting allows developers to draw paths in 3D space towards a specific direction, interactions which occur in the drawn line are reported and can be used to cause action within the experience [164]. In this application, ray casting was utilized to understand when users visual gaze intersected the immersive virtual content. The yellow and green lines display the origin position and direction of each respective eye. The red line represents the mean position and direction of each eye in an effort to estimate and visualise the central point of focus. As this specific path intersects the immersive content it begins to trigger the heatmap drawing process.

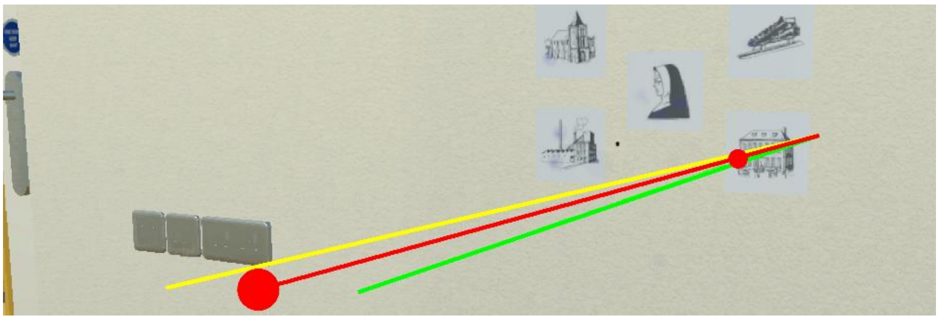


Figure 6.4: Eye Tracking – Oculomotor Convergence

The eye position and gaze direction is visualised in Figure 50 by the yellow (left eye) and green (right eye) lines, the point of convergence is presented by the red line.

Typically, heatmaps are generated using a hidden image layer which captures the gaze intersection with regions of interest. To replicate this, a hidden image plane was placed directly in front of the stimulus in 3D space. Unity3D does not provide a native mechanism which allows the drawing of heatmap data. Hence, another custom algorithm was developed to alter the image over time. Drawing any form of heatmap requires two fundamental components each of which rely on a time parameter.

First, gaze position is factored identifying the region of interest, as time progresses the size of the region increases. Second, and in parallel to this process, as the region of interest increases in size, a “heat” visualisation is generated via the use of a colour gradient. Thus, generating a graphical representation of a visual gaze pattern.

This process was coded into the heatmap generation algorithm (inspired by [114]). As per step 1, the position of a pixel, relevant to gaze direction is reported to the system. As time progresses, the script begins to note the initial pixel position and those around it in an attempt to highlight a region of interest over time. Each of these respective values are stored within an multidimensional “pixel sample array”. This process is visualised in Figure 6.5. Initially, due to the lack of detail the per-pixel sampling will produce a diamond like structure. This is evident in the first four images of Figure 6.5. To circumvent this, the algorithm was designed with numerous thresholds to produce a more circular image (last image in Figure 6.5).

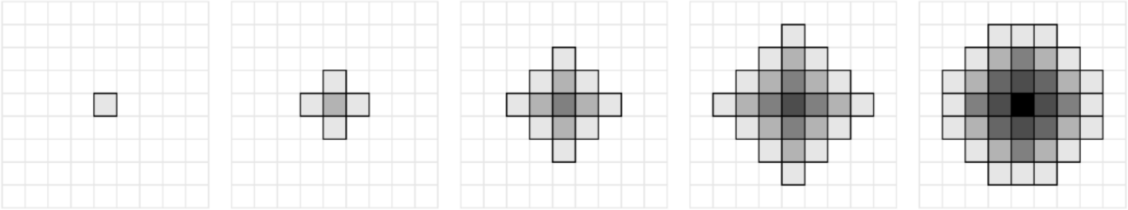


Figure 6.5: Heatmap Generation - Per-Pixel Sampling

As each pixel is sampled and placed into the “pixel sample array”, the second step, the colour data (inclusive of alpha) is noted and stored in a separate multidimensional “colour array”. As time progresses, each of these values are adjusted via a gradient fade as per Figure 6.6. Content which has not been viewed is represented by transparent areas of the image, whilst on the other end of the scale content which appears in red, reports the highest level of visual fixation.

Figure 6.6: Heatmap Gradient Fade

Combining each of these steps, a heatmap image can be generated for each of the respective slides associated with the immersive multimedia SLT assessment (An example can be seen in Figure 6.7). This oculomotor data capture and processing system was designed and developed in such a way that heatmap images can be generated on a per participant basis. A summative overview of user oculomotor response to each slide is presented in Appendix 12 and key findings are discussed later in this Chapter (6.3.4.2).

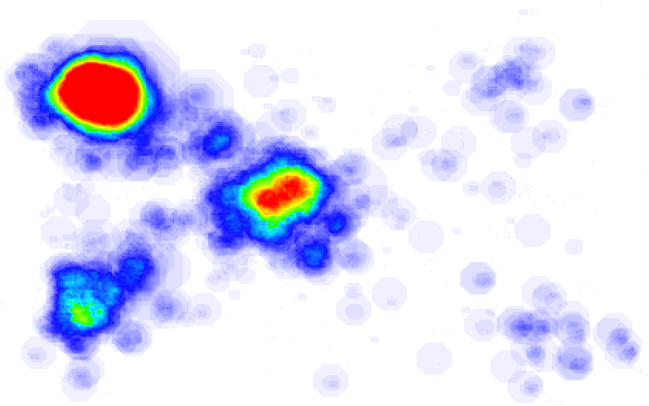


Figure 6.7: Generated Heatmap – Example

2) Oculomotor Reporting

Traditionally, visual tracking experiments capture and report a series of co-ordinates associated with eye position. This data can be difficult to understand and interpret. As a result, string-based AOI reporting [115] is often used as a methodology to summarise and report descriptive oculomotor movements within traditional 2D multimedia experiences (section 3.2.2.3 3) d)). Despite numerous benefits, to the best of the authors knowledge, string-based AOI approaches have not been applied within QoE evaluations of immersive VR experiences.

Inspired by already existing string-based reporting mechanics [115], a Unity3D framework was developed to complete the following tasks:

- Deliver a more descriptive overview of where a user is looking.
- Automatically classify fixation and saccadic events.
- Reduce the amount of data, thus providing a human readable ledger of visual activity.

Although the framework was developed to operate while participants were experiencing the system, or post-experience through the play back of oculomotor data as per section 6.2.2.3 (Part 1), here a description is provided of the post-experience use case. There are two pre-requisites required to run the component. First a ray cast must be supplied. In this PhD work, the summative ray cast associated with the left and right eye position as discussed in 6.2.2.3 (Part 1) is used. Following this, all content which is to be reported should have an appropriate descriptive name, and game-object collision box (also known as a collider [165]). The primary objective of a collider is to approximate the shape of a game-object and track virtual collisions within a physics engine. In the case of this experiment, individual collision boxes were applied to all virtual assessment content, and furniture within the immediate visual range.

Once these parameters have been configured the framework is operational. During playback, the system mimics the ray cast projection, and as this “collides” with each of the pre-defined game objects, a report is generated. This process is best described as “Object based reporting” as it identifies and reports the game object as opposed to the virtual world co-ordinates (Figure 6.8). Thus, altering the report from raw numerical data (as per Figure 6.8 (left)), to semantic descriptions as noted in Figure 6.8 (right).

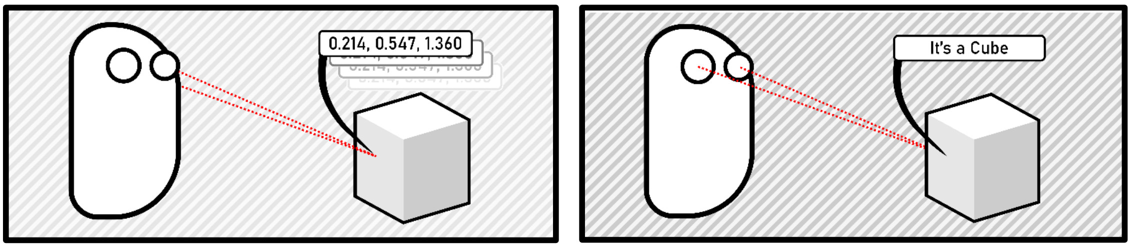


Figure 6.8: Object Based AOI Reporting

(Left) System reports co-ordinates of a visual collision; (Right) System reports a descriptive name of the 3D object.

Events are not generated instantly, instead they are conditional. A timer is instantiated when a participant begins the initial oculomotor fixation with a game object and ends when a ray cast exits the game object (a time difference is recorded). Following this, the time duration is categorized into a fixation or saccadic event. The definition of what quantifies a the length visual fixation differs across multiple works, it can be influenced by visual stimulus and age [109] [85] [166]. In this PhD work, a fixation event was defined as an event over 200ms, whilst saccadic events were classified by events under 200 ms [109]. Once this has been classified a report is generated and written to the file. This method generates a minimal timeline of events over the duration of the assessment, thus providing opportunity to evaluate visual attention via descriptive means.

Although there are many advantages to utilizing the component during an immersive experience, the true potential is unlocked via the post-experience playback. New, more refined visual objects within the virtual environments can be re-defined post-evaluation and reported with minimal effort. Such approaches deliver additional value to large scale immersive experience data sets of the future. As such allowing them the be further evaluated at a precise level, thus allowing for additional contributions within the immersive multimedia research community.

6.3 Results

The following section presents the results of experiment 2. Following a similar approach how the results were presented in experiment 1, explicit, implicit, and objective measures are reported. In addition, in experiment 2, the utility of eye tracking in an effort to further understand physiological response and gain insight into visual attention of the immersive VR SLT assessment is explored. Although all results are presented, an emphasis is placed on understanding oculomotor events with a view to comprehending the identified trends associated with physiological states of arousal.

6.3.1 *Explicit Measures*

Table 6.4 presents the results of the post-experience questionnaire discussed previously in section 5.2.3.1 and available in Appendix 2. The MOS values are presented both as one overall group and separated by gender to highlight some interesting findings. The results of a within-sample T-Test at a 95% confidence interval is also included in Table 6.4. This comparative analysis yielded one statistically significant difference between the genders for Question 14. Complementing this analysis, a collective overview of the QoE model category ratings is presented in Figure 6.9.

Table 6.4: Experiment 2 – Reported MOS – Summative & Gender Split Results

		Group		**Male**		**Female**			
		MOS	**SD**	**MOS**	**SD**	**MOS**	**SD**	**F**	**Sig**
Discomfort	**Q1**	3.821	1.467	3.929	1.385	3.714	1.590	0.877	0.707
	Q4	2.107	1.031	1.857	0.864	2.357	1.151	0.000	1.000
	Q8	2.429	1.103	2.500	1.092	2.357	1.151	0.014	0.848
	Q9	1.357	0.678	1.429	0.852	1.286	0.469	1.298	0.205
Immersion	**Q2**	4.786	0.418	4.786	0.426	4.786	0.426	0.037	0.564
	Q12	4.607	0.567	4.643	0.497	4.571	0.646	0.022	0.826
	Q14	1.786	0.833	1.786	0.699	1.786	0.975	40.847	0.003
Interaction	**Q3**	3.893	0.956	3.929	0.997	3.857	0.949	0.066	0.739
	Q7	2.643	1.496	3.429	1.651	1.857	0.770	1.611	0.587
	Q10	4.607	0.685	4.643	0.842	4.571	0.514	0.190	0.789
Enjoyment	**Q5**	4.500	0.638	4.571	0.646	4.429	0.646	7.374	0.103
	Q6	4.536	0.838	4.571	0.852	4.500	0.855	1.066	0.746
	Q11	3.750	1.041	3.429	1.222	4.071	0.730	2.774	0.651
	Q13	2.821	1.219	2.929	0.997	2.714	1.437	1.620	1.000

The collective MOS ratings presented in Figure 6.9 from a positive rating of 5 to a negative rating of 0. Negatively phrased questions (Questions 4, 8, 9, 13, and 14) have had their MOS values flipped to normalise the scale. A full breakdown of the original data used can be found in Appendix 11. Both gender group participants reported medium levels of comfort for wearing the VR HMD. This can be attributed to the low rating associated with question 8 which asked users if they felt restrictive in their movements while using the system. Similar to the findings reported in experiment 1, influential factors here are associated with the tethered nature of the HTC Vive HMD.

High levels of immersive were reported by both gender groups with an average MOS of 4.321. Due to the enclosed nature of the VR experience, users often reported "forgetting" they were in the real-world post-experience.

In terms of interaction, there is a core difference in the reported levels of enjoyment associated with interaction between genders. Male participants report a MOS of 4.00 with a SD of 0.610. While female participants report a MOS of 3.429 (SD of 1.407). This rather large difference can be attributed to question 7 which explored the learning curve associated with the system. The leap motion sensor used within this experiment to capture hand movement may have been an influence here. This form of technology relies on the interpretation of hand position and gestures via camera-based streams. As such, subject matters which are larger in size may have produced a more accurate system related response. Although only slightly, male and female hands differ in size [167], this may have been enough to produce a less accurate tracking experience and thus impact the perception of interaction . Lastly, Figure 6.9 reports a similar level of enjoyment experienced by both male and female groups with an average MOS rating of 4.009.

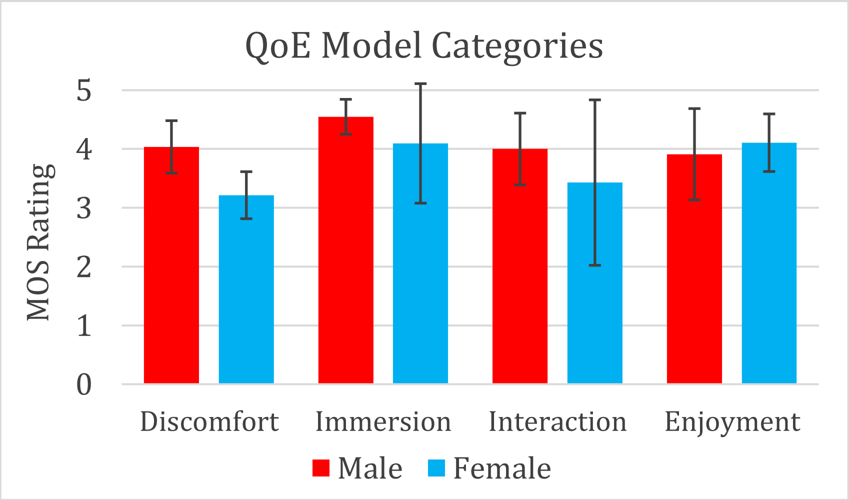


Figure 6.9: Experiment 2 – QoE Model

With the exception of questions 4, 7, and 11 (highlighted in Figure 6.10), there are only minor differences in the overall MOS ratings between the genders. Question 4 reveals a marginal increase in the level of annoyance associated with the HMD for female (MOS of 2.357) in contrast to male participants (MOS of 1.857). Similarly, question 11 asked about the perception of reality associated with the virtual environment. Female participants reported an average MOS of 4.071 (SD of 0.73), whilst male participants reported a MOS of 3.429 (SD of 1.222). This result is interesting as it reveals that female participants may have feel more virtually present within the immersive virtual world, a high level of agreement is reflected in the relatively low SD.

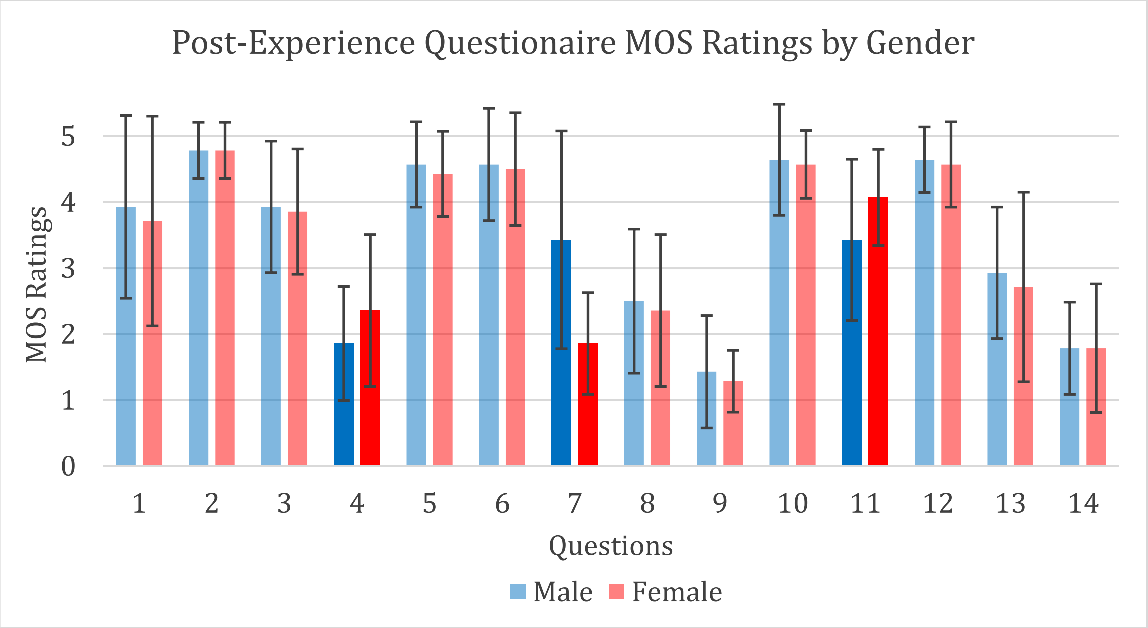


Figure 6.10: Experiment 2 – Post-Experience Questionnaire (Grouped by Gender)

Lastly, reporting the largest level of disagreement between genders. Question 7 asked participants to rate the learning curve associated with the system. Male participants a reported the highest MOS value (3.429), therefore agreeing that the learning curve was neither easy, nor difficult to adapt. There was however a large SD (1.651) associated with this result, thus demonstrating a large degree of disagreement among male participants. Opposing this, female participants reported a what is perceived to be a more challenging experience associated with the learning curve of the system with a MOS rating of 1.857 (SD: 0.855). Overall, the smaller SD is suggestive of a high level of agreement within the female sample size.

6.3.2 *Objective Measures*

6.3.2.1 Interaction

Objective measures of interaction error are summarised in Table 6.5. To begin, focus is placed on understanding assessment errors. Considering the results presented in experiment 1 (Table 5.7), the HTC Vive group in experiment 2 had the highest number of assessment errors of all of the devices evaluated. However, as the sample sizes differ slightly, similarities are revealed through looking at the average number of assessment errors. In experiment 2, participants experienced 1.143 errors on average. This measurement is almost identical to that of the Oculus DK2 group in experiment 1 reported in Table 5.7 which experienced an average of 1.15 errors (23 assessment errors ÷ 20 participants).

Consistent with the previous findings for all the other systems evaluated, slide 9 reports the highest level of assessment error, with an average of 0.286 (SD 0.600). A similarity can be seen with original VR group (Oculus DK2) in terms of the average number of assessment errors for slide 9 (0.35 for the Oculus DK2). Of intertest, Table 6.5 also reveals a large degree of error associated with slide 8. This finding was not reported in the previous assessments. These errors are evenly distributed across 4 individual participants (3 female and 1 male). The practice slide (P) and slide 3 each had a count of 4 errors. Each of these respective slides solicited a large degree of error in the Oculus DK2 group (Table 5.7).

It was hypothesised that the increase in device specifications (i.e. screen resolution and display technology) would alleviate some of the ad-hoc reported challenges associated with the perception of detailed stimuli. Considering these results, this was not the case. Despite this, it would appear that the perception of the virtual cursor may have hindered interaction performance. This is reflected in the total number of interaction errors reported in Table 6.5.

The HTC Vive group report a total of 5 interaction errors encountered throughout the virtual assessment. This is a significant reduction in contrast to that of the Oculus DK2 group who reported a total of 14 interaction errors. This demonstrates the advantage of using higher screen resolutions. The increase in visual clarity, and thus reduced impact from the screen door effect [156] [157] more than likely delivered a higher level of awareness for the virtual cursor. Thus, facilitating in a reduced level of interaction error which falls more in line with the AR group results presented in Experiment 1.

Table 6.5: Experiment 2 – Interaction Error Statistics

Slide	**Assessment Error**			**Interaction Error**		
	Count	***Average***	***SD***	***Count***	***Average***	***SD***
Slide P	4	0.143	0.356	1	0.036	0.189
Slide 1	2	0.071	0.262	1	0.036	0.189
Slide 2	1	0.036	0.189	0	0	0
Slide 3	4	0.143	0.448	0	0	0
Slide 4	1	0.036	0.189	0	0	0
Slide 5	3	0.107	0.416	1	0.036	0.189
Slide 6	0	0	0	0	0	0
Slide 7	2	0.071	0.262	1	0.036	0.189
Slide 8	6	0.214	0.418	0	0	0
Slide 9	8	0.286	0.600	0	0	0
Slide 10	1	0.036	0.189	1	0.036	0.189
Total	32	1.143	0.303	5	0.179	0.086

6.3.2.2 Progression Time

The average progression time for both the HTC Vive group and the Oculus DK2 group on a per slide basis is reported in Figure 6.11. Participants using the HTC Vive made a successful selection every 5.497 seconds on average (SD 3.694 seconds). This result delivers an increased performance in contrast to that of the Oculus DK2 group who progressed every 7.893 seconds on average (SD of 5.557 seconds). There are similar trends for both groups with a slight hesitation associated with the start of the assessment (Slide P). This is followed by a higher level of confidence which reduces the average progression time. Slide 3 reports an increased progression time for both groups, this is also reflected in the previously reported interaction metrics discussed in part 1 of this section (6.3.2.1).

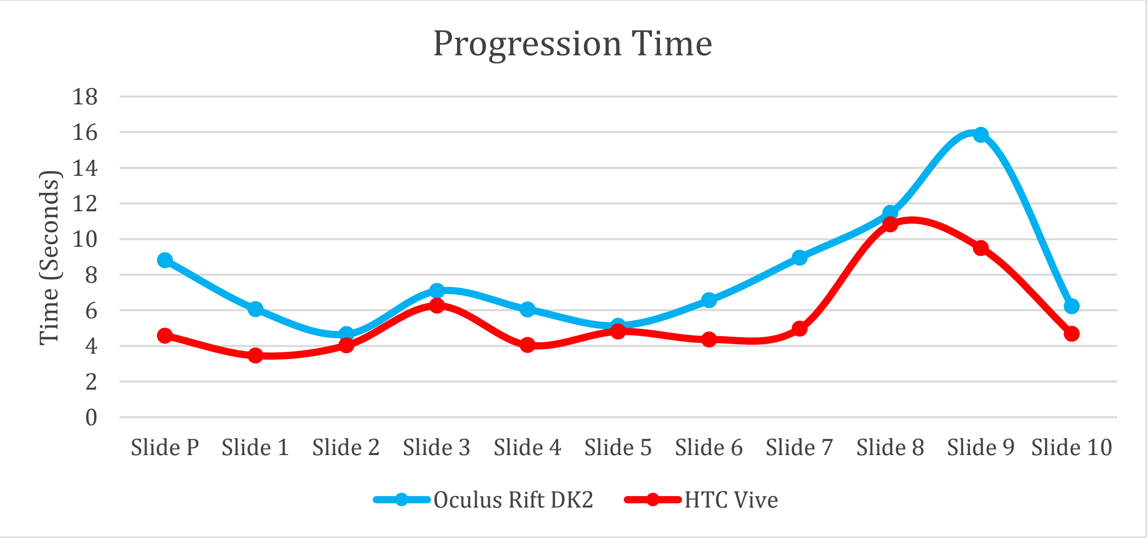


Figure 6.11: HTC Vive – Progression Time

Later in the assessment, the HTC Vive group report a large increase in progression time for slide 8. This is again consistent although not as exaggerated as that of the DK2 VR group. Reflecting on the complete data set, 10 participants experience a delay of 10 seconds or more. As this is over a third of participants, it’s believed that this increase is not by chance. Similarly, slide 9 also reports a large increase in the average progression time. Once again, this increase cannot be attributed to an individual as nine participants reported the increased progression delay.

In an attempt to further understand these interaction errors and delays in progression time in a deeper manner, section 6.3.3 and section 6.3.4 explores the associated physiological and oculomotor responses respectively. More specifically, a comparative analysis is carried out on task-evoked physiological measures. The oculomotor data is evaluated in section 6.3.4 to gain insight into the cognitive selection processes and associated levels of indecisiveness which occurred during the assessment. Thus, gaining previously unattainable insight into the human, system, and context factors which may have altered the user perceived QoE within this immersive virtual SLT assessment.

6.3.3 *Physiological Response*

6.3.3.1 Empatica E4 Data – A Statistical Evaluation

Table 6.6 presents a statistical analysis of the data captured from the Empatica E4. Data associated with the training phase has been excluded from this analysis and the within analysis compares the waiting (a baseline reading) and test recordings. More specifically, the results focus on the mean score of EDA, HR, HRV, and skin temperature for both the waiting and test phases. A paired samples t-test was carried out at a 95% confidence interval utilizing the complete sample size of 28. The paired results evaluated physiological changes as participants progressed from the waiting (baseline) and test phases of the experiment. There are two statistically significant changes for EDA and HR between the baseline and test phases.

Table 6.6: Physiological Data – Paired Samples T-Test (95% C.I)

	Waiting		**Test**			
	Mean	***SD***	***Mean***	***SD***	***T***	***Sig***
EDA (µS)	0.898	1.505	1.641	2.091	-3.248	0.003
HR (BPM)	78.017	10.715	81.814	12.546	-3.151	0.004
HRV	-0.038	0.131	0.002	0.168	-0.901	0.375
Temperature (°C)	31.160	1.388	31.384	1.446	-1.048	0.304

First, an increase of 0.607 μS is reported in EDA over the baseline (0.898 μS). The change results in a statistical significance of p=0.003 as users progressed through the assessment. Furthermore, an increase in SD of 0.586 μS is noted in the test phase, this result is indicative of a large degree of physiological response occurring due conditions occurring within the immersive VR SLT assessment.

The statistical analysis of HR reveals an increase of 3.797 BPM during the test when compared to the baseline. Like EDA, an increase can be seen in the average SD (1.831 BPM) as users progressed through the phases. These conditions resulted in a statistical significance of p=0.004. Although this result is of interest, it’s important to also reflect on the overall user experience within the immersive VR experience. Participants remained still during when baseline measures were captured. In contrast to this, the gesture-based interaction system may have resulted in an increase in blood flow contributing somewhat to the reported result. Due to the nature of interactive and immersive multimedia experiences, this can be difficult to control.

In parallel, these external factors associated with movement may explain why there is a high level of significance associated with HR and not with HRV, despite both measuring the heart in an effort to capture emotional states. Lastly, Table 6.6 reports the average change in temperature as users progressed throughout the virtual assessment. There were no significant changes.

The data presented in Table 6.6 delivers a summative overview of physiological response throughout the experiment. Overall, the results prove promising with early indicators of EDA and to an extent, HR displaying physiological responses which may be associated with changes in arousal. To further validate and understand these findings, the results can be complemented by measures of interaction to deliver a more precise overview of physiological response. Although such approaches have proved to be of significance in previous experiments (section 5.3.2), they can be further supported through the consideration of not only by understanding progression, but through the immersive task context.

6.3.3.2 Task-Evoked Physiological Response

EDA and pupillary response are time series measurements by nature. There are numerous methodologies which aim to explore, interpret, and evaluate physiological data, some of which transform physiological responses on the time domain to the frequency domain [168] [169]. Immersive multimedia experiences are by nature interactive. The capture of task driven data within complex virtual worlds delivers an abundance of information relating to user QoE. Thus, the synchronization of application data and physiological response can enable the extraction of key insights into task-evoked measures of arousal. Such an approach delivers an awareness of user interaction within the environment to automate the reporting of events, thus providing a supporting objective measure of user QoE.

Due to the sandboxed nature of the virtual assessment application, there are three primary events which have been explored to be a reliable trigger for emotional arousal. They are: (a) user progression (via correct response); and two measures of error: in the form of (b) assessment error; and (c) application error (user interacts with incorrect content). It is hypothesized that as users encounter error within the immersive multimedia assessment, a physiological response which indicates a stress related trigger can be captured within the VR environment. Complementing this, a correct response and positive progression within the immersive virtual assessment should produce an increase in phasic change and thus, physiological response as per real world conditions.

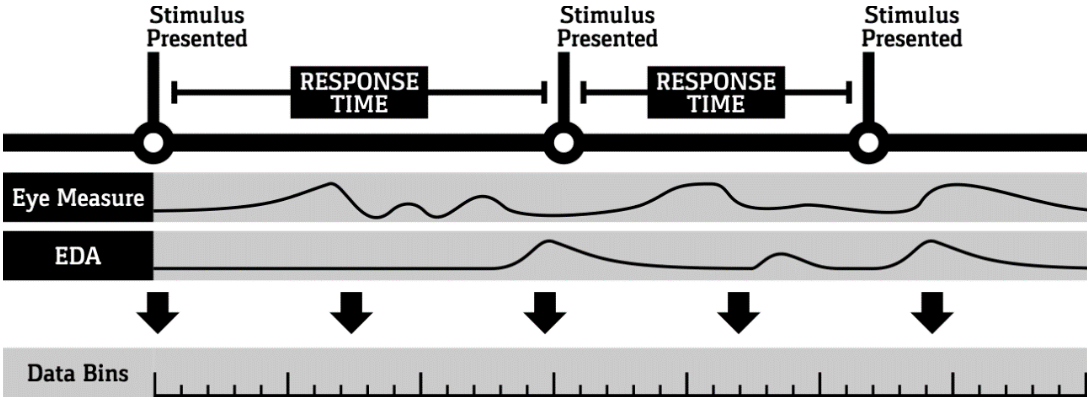


Figure 6.12: Task-Evoked Physiological Data Processing

1) Task-Evoked Electrodermal Activity

Guided by the methodological approaches described in section 6.3.3, Figure 6.13 presents physiological responses in EDA based on the various “triggers” introduced in section 5.2.3.2 (Part 2). Recent works have highlighted that phasic changes in EDA are indicative of emotional arousal occur between 1 and 3 seconds after an event [142]. Considering this, a sample size of 5 seconds was chosen. This allows for the maximum delay in physiological response and an additional 2 second time window to capture the change. Thus, Figure 6.13 reports a total of 20 samples of EDA over a duration of 5 seconds (capture rate of 4Hz) in an effort to capture signal trend.

The data extraction process resulted in a sample size of 373 measures of physiological response as outlined in Table 6.7. As the sample population consisted of the typically developing population, there was a reduced level of error. Users who experience an assessment error or interaction error received an audible prompt to try again. Due to the level of similarity between these two scenarios, it was assumed that each of these independent interactions invoke a similar level of physiological response associated with the stress of making a mistake and thus not progressing. Therefore, both sample bins are also presented as a single measure known as “all errors”. This approach resulted in a more reliable measure of stress related physiological response within the immersive experience.

Table 6.7: Task-Evoked Physiological Response – Sample Overview

	Correct	**Assessment Error**	**Interaction Error**	**All Errors**
Number of Samples	336	32	5	37

Due to the nature of the data, the discussion here places an emphasis not on the signal changes over time, but on the trends associated with physiological response. This approach factors for variability associated with each participant independent delay and onset of emotional arousal (if any). Thus, it facilitates a fair interpretation of the sample.

Focusing on data associated with user errors (assessment error and interaction error), Figure 6.13 reveals a large degree of variation in both signals post interaction (0 seconds) exists. Although each of these independent measures alter within a unique manner, there is consistency associated with the signal trend post-interaction. Both errors and interaction errors produce a positive trend as they traverse the 5 second time window. This can be seen in the trend line associated with the summative measure of all errors (red). Typically, these positive trends can be associated with user stress [97] [170] [171] as they experience an increase in phasic response over time.

Figure 6.13 also reports the signal trend of participants when the experienced a correct answer and progressed through the assessment. Unlike the reporting of errors, the correct responses signal trend does not produce a clear positive or negative trend. Instead a more horizontal response is noted over the duration of the 4 second sample. This response is to be expected as physiological responses associated with skin conductance are typically correlated to cognitive and emotional states of arousal (i.e. the fight-or-flight response) as opposed to relaxed states (i.e. the rest-and-digest response) [102]. Arguably these core differences in trend are only minor. However, the signal trend associated with the correct responses also reveals a much lower SD of 0.042 in contrast to the much larger signal variation correlated with interaction errors (0.153). This large degree of signal variation associated with user error is indicative of user specific independent variations associated with phasic changes occurring at varied stages across the 5 second sample [97] [129] [170] [171]. Whilst the stability associated with the correct answers delivers insight into a more unified phasic response in EDA, which may be indicative of the rest-and-digest responses of the skin [171].

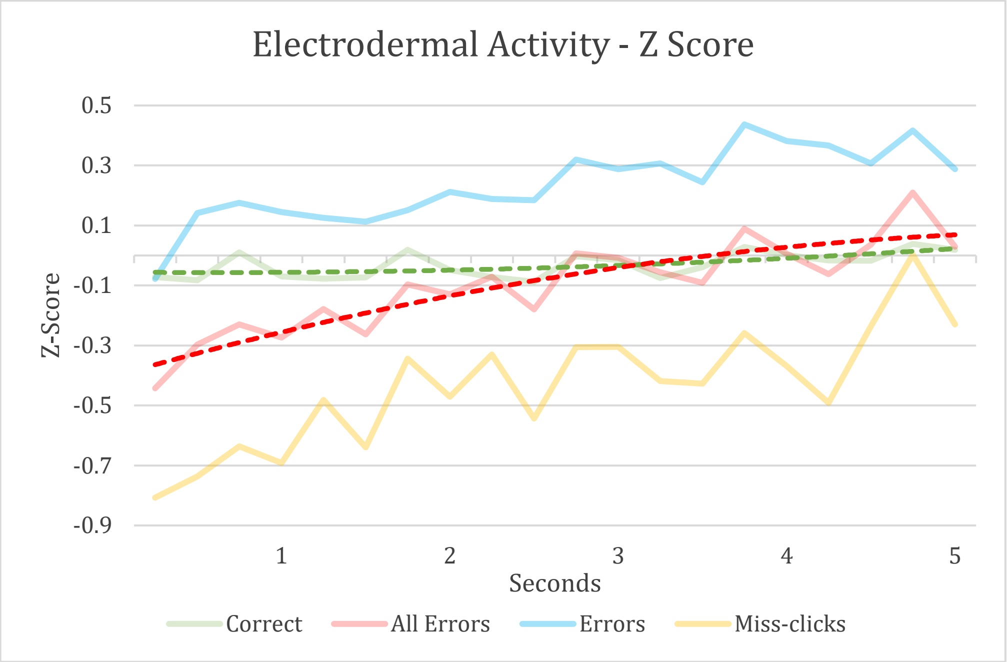


Figure 6.13: Electrodermal Activity – Z-Scores (Post-Interaction)

2) Task-Evoked Pupillary Response

The onset of a pupillary response typically occurs 500 ms after a task-evoked event [172]. Similar to EDA, there are varied degrees of delay associated with each individual participant. Guided by the recommendations outlined in [172], measures of pupillary response were extracted 500 ms prior and 1500 ms after a task related interaction occurred within the immersive virtual environment. The Tobii Pro HMD captures eye measurements at a rate of 120 Hz, however pupillary response was sub sampled at an interval of 100 ms (10 Hz). This sample rate was deemed sufficient as the focus on the results presented here is on the identification and interpretation of trends in pupillary response as opposed to a deep understanding of minute oculomotor movements. Figure 6.14 presents an average pupillary response for both left and right eyes. Although subtle variations exist between each eye, pupil diameter is highly correlated [173].

Similar to the results presented in Figure 6.13, pupillary response was extracted under three conditions (correct response, user error, and interaction error). As part of signal processing, the raw data was converted to Z-score in an effort to normalise the data [172]. Through this process, the captured data set produced two unique findings in VR which can be linked to changes in states of arousal.

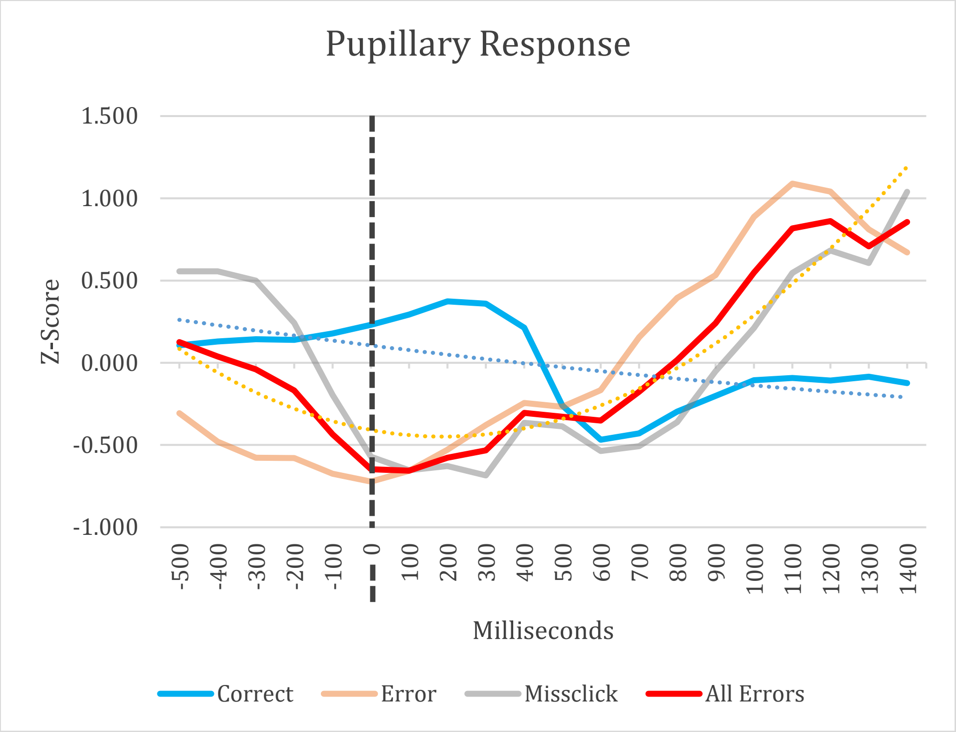


Figure 6.14: Pupillary Response – Z-Score (Pre and Post-Interaction)

First, focusing on the average pupillary response associated with correct answers, we can see a significant change. 500 ms after interaction occurs, there is a rapid decrease in pupil size. This is indicative of the pupil contracting overtime, therefore revealing a state of relaxation occurring post-interaction. From a multimedia perspective, the visual change in stimulus and virtual SLP providing an audio experience are more than likely influential factors. Both deliver an indication of immediate success, therefore inciting the physiological response.

Of interest, assessment errors and interaction errors follow an opposing trend. Both of these measures have been aggregated to an “all errors” metric as they produce a similar response from the immersive virtual assessment. Unlike the physiological change associated with correct answer, errors cause the pupil to contract in size. Thus, visualising the activation of the fight-or-flight response associated with experiencing an error [174].

6.3.4 *Oculomotor Visualizations*

6.3.4.1 Visual Attention

Visual attention is a fundamental component in the understanding of cognitive processing and user perception of multimedia content. The virtual STL assessment is by nature limited by the number of visuospatial elements within the immersive multimedia environment. As a result, the presentation of scan path places an emphasis on representing visual attention via summative graphical representation.

6.3.4.2 Fixation & Saccades

Table 6.8 delivers a summative overview of visual fixations throughout the immersive multimedia experience. There are two primary columns, first the targets column reports visual stimuli (key and target pairs), these have been separated as user progression relied on the interpretation of each of these respective components. Second to this, the distractors column reports the close, distant, and unrelated semantic distractors as per the CAT guidelines.

Table 6.8: Visual Fixation – Time Related Statistics

Slide	**Targets**		**Distractors**			**Other**	**Average**
	Key	***Target***	***Close***	***Distant***	***Unrelated***		
Launch Screen	5.204					0.264	5.204
Practice Slide	1.198	**1.760**	0.171	0.514	0.083	0.258	0.745
Slide 1	0.614	**1.659**	0.248	0.103	0.167	0.271	0.558
Slide 2	0.829	**1.905**	0.205	0.209	0.202	0.207	0.670
Slide 3	**2.221**	1.912	0.437	0.356	0.468	0.204	1.079
Slide 4	1.061	**1.774**	0.118	0.139	0.295	0.218	0.677
Slide 5	1.268	**1.562**	0.572	0.396	0.436	0.141	0.847
Slide 6	1.002	**1.398**	0.503	0.326	0.392	0.197	0.724
Slide 7	0.754	**1.851**	0.508	0.640	0.392	0.254	0.829
Slide 8	**2.596**	2.007	1.364	2.215	1.549	0.176	1.946
Slide 9	**2.905**	2.567	0.982	1.224	0.995	0.142	1.735
Slide 10	1.211	**1.317**	0.493	0.529	0.436	0.177	0.797
Average	1.424	1.792	0.509	0.605	0.492	0.204	0.964
SD	0.781	0.336	0.373	0.616	0.424	0.044	0.506

A descriptive overview of the stimulus, and object classification (i.e. key, target, or distractor) can be found earlier in this thesis in Table 5.1. The “other” column reports the average time users were fixated on environmental content (walls, doors, furniture etc.). These forms of visual fixation only accounted for an average of 0.204 seconds across the virtual assessment, as a result they are not reported in a high level of detail. Lastly, stimuli which reports the largest average period of fixation is highlighted in bold on a per slide basis.

The results reveal that users spent on average 0.964 seconds (SD: 0.506 seconds) focusing on each of the respective stimuli (all 5 images). Out of the five possible options, participants focused their attention on the target stimuli (answer) in 8 of the 11 slides. This finding is interesting as it reveals users spent less time consuming the visual representation of concepts displayed on the key objects. In addition, the increase in fixation associated with the target stimulus highlights a delay associated with cognitive processing as users no-doubt reflect on the key object prior to selecting the target object.

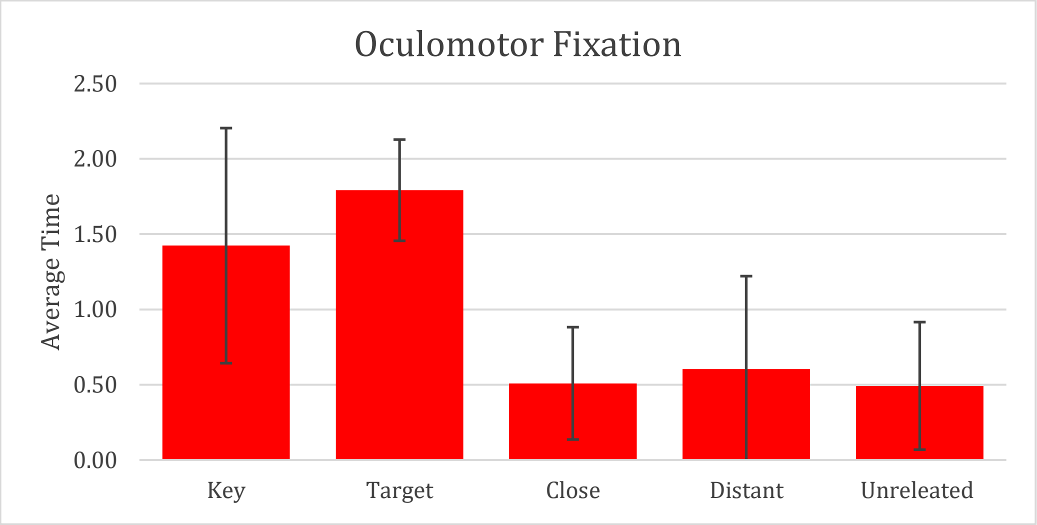


Figure 6.15: Oculomotor Fixation

The above figure presents an overview of the average fixation time for each of the respective stimulus types.

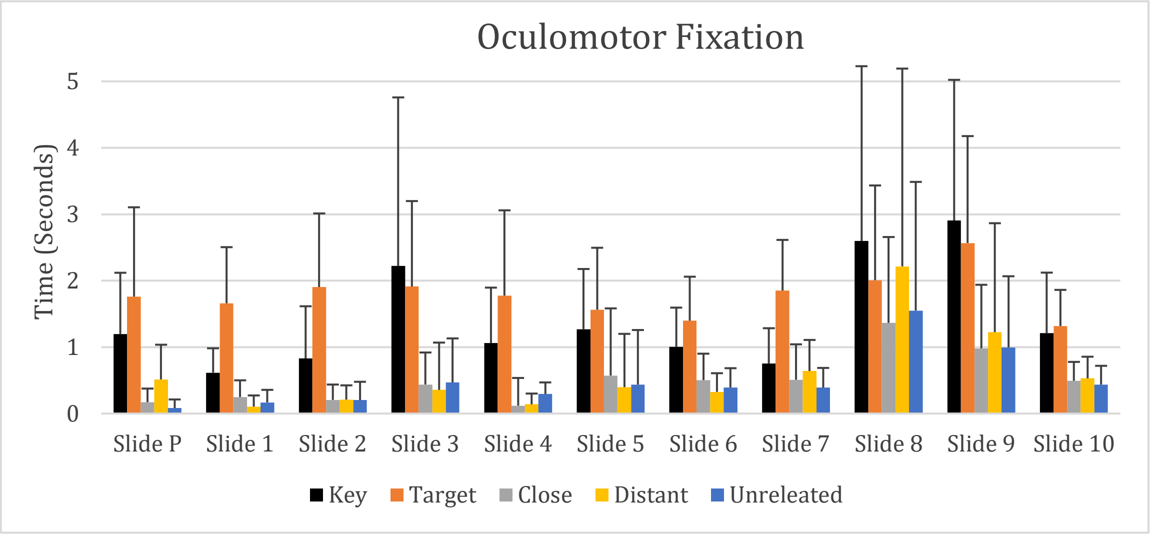


Figure 6.16: Overview of Fixation Data per Slide

For the three remaining slides (3, 8, and 9), a longer period of visual fixation occurred on the key object (Figure 6.16). This finding delivers fundamental insight into Figure 6.11 which also reports that slides 3, 8, and 9 have the highest progression times from an assessment perspective. An increase in visual stimulus may reflect numerous events occurring as a result of an increase in oculomotor and cognitive processing.

It is well accepted that gaze density maps of eye data can deliver fundamental insight into oculomotor attention within multimedia experiences [114]. However, due to the repeated stimulus and focused interactive approaches within the context of this research, it has not been used as an analytical approach. Instead, heatmaps have been generated as a supportive visual component, a full overview of all slide content inclusive of the overlaid heatmap can be seen in Appendix 12.

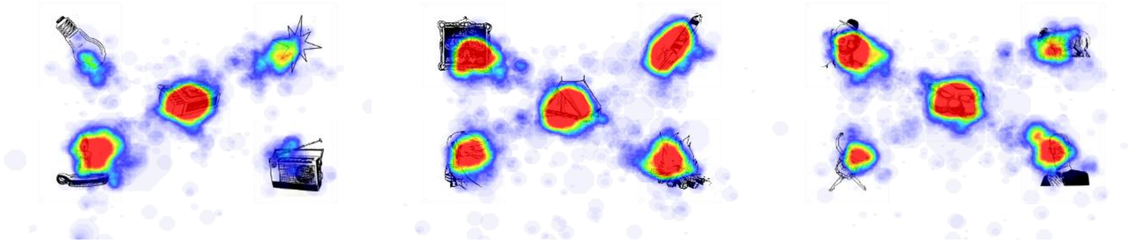


Figure 6.17: Generated Heatmaps

Slide 3 (Left), Slide 8 (Middle), and Slide 9 (Right)

Figure 6.17 presents an overview of the heatmaps generated for slides 3, 8, and 9. Immediately, the summative information delivers core insight into the visual content which captured the participants attention. This information is however limited buy the ambiguity associated with the time series nature of a gaze density map. Visual features which gained attention are clear and highlighted by colours ranging from clear to blue, yellow, and then to red. Nevertheless, there is a crucial loss of data as heatmaps occlude the cognitive decision-making process which is required to progress throughout the immersive virtual assessment.

To further understand this phenomenon, Table 6.9 and Figure 6.18 present supporting information in the form of the average number of fixation events which occurred on a per stimulus basis. Fixations are counted when a participant’s eye remained fixated on a stimulus within each of the respective slides for a duration of 200 ms or more [109]. In Table 6.9, summary statistics are presented in the form of an average, and total number of fixations. This provides core insight into significant oculomotor movement within the immersive multimedia experience. Values which are highlighted in green represent data which is one SD beyond the mean, whilst those highlighted in amber represent values two SD beyond the mean for each of the respective stimulus categories. From Table 23, it is clear that both slide 8 and 9 are significant within this context. This follows a similar trend to previously discussed data.

Table 6.9: Visual Attention – Fixation Count

Slide	**Targets**		**Distractor**			**Summary**	
	Key	**Target**	**Close**	**Distant**	**Unrelated**	**Average**	**Total**
Launch Screen	2.179					2.179	2.179
Practice Slide	2.500	2.107	0.536	1.214	0.321	1.336	6.679
Slide 1	1.607	2.286	0.821	0.321	0.607	1.129	5.643
Slide 2	1.679	2.429	0.643	0.786	0.643	1.236	6.179
Slide 3	3.071	2.500	1.357	0.821	1.071	1.764	8.821
Slide 4	1.929	2.321	0.321	0.536	0.964	1.214	6.071
Slide 5	2.214	1.821	1.179	1.107	0.929	1.450	7.250
Slide 6	1.821	1.929	1.250	0.821	1.071	1.379	6.893
Slide 7	1.571	2.643	1.250	1.536	1.000	1.600	8.000
Slide 8	4.643	3.321	2.643	2.929	2.571	3.221	16.107
Slide 9	4.500	3.500	1.929	2.000	1.857	2.757	13.786
Slide 10	2.536	2.357	1.250	1.393	1.393	1.786	8.929
Average	2.552	2.474	1.198	1.224	1.130	1.716	8.045
SD	1.100	0.523	0.657	0.738	0.627	0.729	3.702

With respect to slide 8 (Figure 6.19), an average of 16.107 fixation events occurred as users consumed the multimedia. This is significant as it delivers further insight into a large degree of indecision in terms of choosing the correct content. On average, participants glanced at each of the respective objects 3.221 times. Both the key (tent) and target (fire) are reported as the objects with the largest number of fixations. These results are to be expected as assessment progression relies on visually consuming each of these. Of more interest is the large quantity of views that each of the respective semantic distractors obtained. Users glanced at the distant (rocket) distractor on average 2.929 times, the unrelated distractor (picture) 2.571 times, and lastly the close semantic distractor (torch) an average of 2.643 times. Such increases are significant as each of these values are two SD within the norm.

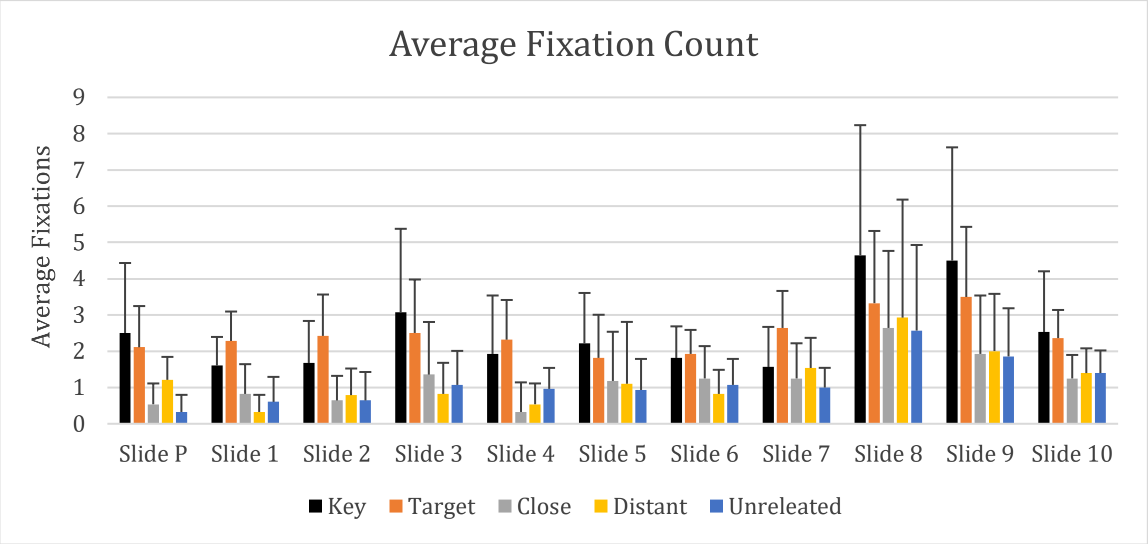


Figure 6.18: Average Fixation Count

There are numerous factors which could have influenced such a change. For example, the clip art images associated with of each of the semantic distractors, or key differences in visual perception through the HMD in contrast to the real-world. In addition, there is an inherent similarity between the target (fire) and close semantic distractor (torch) as each of these respective images contained a pictographic representation of a flame. Through ad-hoc discussion post-experience, participants did note that there was a minor challenge on the slide 8. Of interest, most mentioned the interpretation of the picture (Figure 6.19 (top left)).

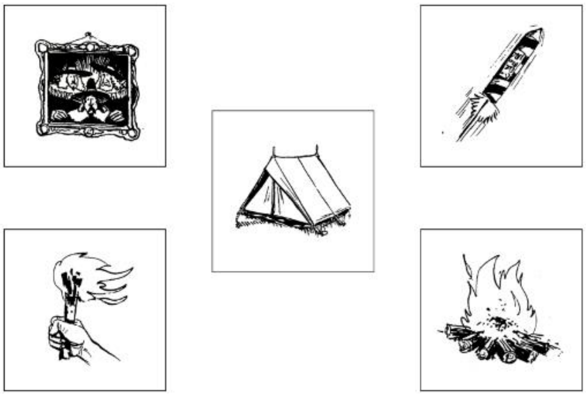


Figure 6.19: Slide 8 – Stimulus

There are, however, other factors which must be considered. Although the content sourced from the CAT has been clinically validated, the assessment was published in 2004 [69]. Societal changes have the potential to impact the perception of associative knowledge. A well-known example of this is to present a participant from generation Z, or generation Alpha an audio cassette and a pencil challenging them to find the association.

As each of these respective generations grew up within the digital age, the shift from analogue to digital has the potential to interrupt the recollection of such knowledge.

Therefore, the influence of human factors from a generational perspective cannot be ruled out within the context of this experiment. For example, the association within slide 8 (fire and tent) may have been not been an obvious choice for some participants. Reflecting on the content from a pathological perspective, the question is posed as to how often image based content associated with visual assessment batteries should be updated. Should generational content be considered moving forward or can adaptive systems which already exist today in other fields be used prior in an effort to generate user specific assessments. Thus, delivering an overall higher QoE to the end users and therefore higher assessment standard tailored to the individual.

6.4 Discussion & Conclusion

Real world application of immersive multimedia allows users full interactive control over virtual worlds. As a result, it is becoming more challenging to evaluate and understand the perception of visuospatial components. To circumvent these challenges, this experiment presents an overview of an extension to the methodology presented in Chapter 5. The extension focused on the capture and interpretation of oculomotor movement. Two custom Unity3D components were created to: (1) derive semantic reports of visual attention and generate heat maps, and (2) to automate the classification of oculomotor fixation and saccadic events as users experienced the immersive virtual world. This novel approach alleviates the requirement for post-processing, and to a large extent reduces the effort required to interpret complex data sets. As such, simplifying the reporting of visuospatial attention and oculomotor measures into precise data sets for QoE evaluation. This approach, in parallel to the objective measure of user interaction, derives key insight into user attention, and thus context. Therefore, allowing for an in-depth understanding of the user perception of virtual stimulus within the immersive environment.

Prior to the experiment, it was hypothesized that the user's visual perception reports a similar level of media fixation prior to making a selection within the assessment. Instead, in some cases it would appear that the semantic recollection of conceptual links takes priority over visual-spatial consumption of the multimedia content. Thus, not all of the stimuli are considered prior to making an interaction within the immersive multimedia assessment.

Implicit measures of EDA reported in Chapter 5, highlight an increase in physiological arousal associated with an increase in cognitive load, and thus stress is captured for the VR group. In this experiment (Chapter 6), supporting measures of pupillary response were captured to further understand and validate the initial findings. The results presented here factor objective measures of user interaction, in an effort to derive more precise measures of task-evoked physiological arousal. These precise measures demonstrate an activation of the fight-or-flight response in measures of both EDA, and pupillary response when users encounter an error (Assessment, or Interaction error). This implicit response is believed to be an indicator of stress. Similarly, when users progress through the virtual assessment, a physiological response associated with the rest-and-digest state (relaxation) is noted in each of the implicit measures. Although the application within the context explored here may be limited, the fundamental components of interaction (both correct and incorrect actions), and potential to capture precise measures of physiological response exist in the majority immersive multimedia applications. Thus, it is possible that such a novel task-evoked approaches could be applied in a general capacity within other QoE evaluations in an effort to derive a deeper understanding of emotional arousal.

Overall, the combination of chosen metrics delivers a holistic overview of user perception of multimedia, informative overview of physiological response, and therefore fundamental insight into overall user QoE. The approaches utilized within this experiment could theoretically be used to derive context from data captured live in future work. Such approaches would require manual programming, however recent advantages in ML have demonstrated the benefit of black box solutions utilizing rigorously tested mathematical models. Thus, the future of QoE evaluations may rely on such novel data processing approaches in an effort to automatically derive context from extensive data sets in the future.

Chapter 7: CONCLUSION

In this Chapter, the conclusions of this PhD research are presented. Each of the respective research sub-questions are addressed in section 0. Following this, some limitations of the work are highlighted and discussed. Finally, three proposals are presented as avenues for future work.

7.1 Conclusion

Contributions of this work have focused on the QoE evaluation of immersive multimedia devices (AR, VR and Tablet) as a potential delivery method of speech and language assessment. An experimental methodology was developed in which user perceived QoE was evaluated. Traditional explicit and novel implicit metrics were captured as an insightful measure of QoE. Objective measures of interaction were also explored for the purpose of evaluating the utility of such a measure within the context of QoE evaluations. In addition, the capture of these novel measures aimed to explore opportunities for the delivery of more insightful measures of speech and language assessment performance within the clinic.

Analysis of results suggests that the tablet technology supports a higher overall QoE when compared to immersive multimedia HMDs. The results, however, reveal only a slight advantage when compared with the AR solution. The tablet and immersive AR technologies deliver very similar and, in some cases, identical results. What is interesting here is that for both the tablet and AR systems, the user is operating in the real world and interacting with digital content. Thus, within the context of an immersive multimedia application for SLT assessment, it’s hypothesized that remaining within a real world surrounding (as opposed to virtual) may play a fundamental role within the perception of quality in such a novel experience.

Initial results for the VR group (Oculus DK2) as presented in Chapter 5, reveal a lower level of perceived QoE. Objective measures of user performance are not on par with that of the AR and tablet groups. In addition, physiological measures within the VR group reveal a notable change in EDA which is believed to coincide with an increase in cognitive activity. These findings, however, may have been attributed to limitations associated with system factors. For example, the lack of native hand-gesture support (i.e. reliance on a 3rd party controller), and lower screen resolution may have been a key influence.

To further evaluate and understand the objective and implicit measures for the VR DK2 group, a second experiment was carried out. In Chapter 6, the results of another VR experiment (HTC Vive group) is presented in which users experienced a device more representative of the current state-of-the-art. In addition to the clinical-grade capture of heart rate and EDA, physiological measures of oculomotor response were recorded in an effort to provide a supportive measure of the physiological responses as noted in the initial results (Chapter 5). Data capture and analysis explores more precise measures of task-evoked implicit responses in an effort to gain a deeper understanding of physiological states of arousal, and thus QoE. The multimodal approach reports similarities in physiological measures of both EDA, and pupillary response, both of which can be utilized to gain insight into user emotional states.

Considering the objective data captured by the novel QoE methodology, user emotional states were classified into two independent categories. Informed by the literature, physiological response associated with what is perceived to be stress (fight-or-flight) was extracted when users experienced an assessment error or interaction error. Similarly, physiological responses associated with states of relaxation (rest-and-digest) were identified when users got a correct answer and progressed through the assessment.

Considering the above, each of the three multimedia platforms could be employed in future SLT evaluations, although minor shortfalls exist with the current generation of VR HMD, this will no doubt resolve overtime as the technology matures. The novel QoE methodology explored the capture of explicit, implicit, and objective measures to derive user perception of quality. The holistic approach to data capture, and processing presents a methodology which allows for the evaluation of user-centric active multimedia experiences in the future. In short, the novelty of this work can be summarised as follows:

- A first of its kind QoE evaluation of three multimedia platforms (AR, tablet, and VR) as potential candidates for the delivery of an immersive virtual SLT assessment.
- A novel QoE methodology which considers explicit, implicit, and objective measures of user interaction.
- The novel use of precise, task-evoked physiological measures of emotional arousal within an active user-centric immersive multimedia experience.

7.2 Research Sub-Questions

In this section, each of the respective research sub-questions (as outlined in section 1.2.1) have been answered.

- **SubRQ1:** Can immersive multimedia systems invoke a similar level of performance when compared to conventional paper-based semantic memory assessments?

Considering the results of experiment 1 (section 5.3.3), Table 5.8 reveals that two immersive multimedia technologies invoke a similar level of assessment performance when compared to traditional paper based assessments. Both AR and tablet report an average score which is one standard deviation within the mean of the normative results. VR did not perform as well (two SD from the mean). It is believed the next generation of VR HMDs which support higher screen resolution and native hand gesture recognition will be able to resolve this.

- **SubRQ2:** Do immersive multimedia experiences result in a higher degree of QoE when compared to that of a traditional 2D multimedia experience?

Comparing explicit, implicit, and objective measures of user performance presented in Chapter 5 the results reveal that AR can invoke a similar level of QoE when compared to traditional 2D experiences (tablet). Each of these respective groups experience a high level of a similarity associated with physiological measures of EDA, HR, and objective measures of user performance (progression time and assessment performance). User self-reporting MOS ratings demonstrate a statistically significant difference associated with discomfort and immersion. In terms of discomfort the results were in favour of the tablet group, as they report the lowest level of system annoyance. Alternative to this the AR group reported a higher MOS rating, therefore revealing that they would like to experience the environment again. Reflecting on these explicit measures, they are to be expected as the there is a core difference in device form factor (HMD vs no HMD).

- **SubRQ3:** Can objective measures of user performance support a deeper understanding of user and physiological response in the context of a QoE evaluation of an interactive and immersive multimedia speech and language therapy assessment application?

In Chapters 5 and 6, increases associated with user performance (progression time) are paired with application interaction logs in an effort to derive context. The results reveal what is perceived to be an increase in cognitive load. Utilizing the same objective measures of user performance, accurate measures of physiological response can be extracted. As such supporting the initial findings associated with cognitive load as reflected through physiological response in EDA. Considering these findings, objective measures of user performance can support a deeper understanding of the user, and physiological response within a QoE evaluation of an active immersive multimedia experience.

- **SubRQ4:** Can precise measures of physiological response be used as a reliable measure of emotion and as such provide a better understanding of user response within the context of a QoE evaluation of an immersive virtual reality experience?

In experiment 1 (Chapter 5 and 6), implicit measures of EDA and HR are aggregated with objective measures of user interaction (progression time). This approach derives user-centric context, not only from the application perspective, but in terms of the understanding of precise windows of physiological response. As a result, deriving a holistic overview of emotional states within active user-centric immersive multimedia experiences.

7.3 Limitations

There are numerous limitations which must be acknowledged throughout this work. Some of which revolved around hardware and the influence of gender bias. The initial experiment (as detailed in Chapter 5) compares three multimedia platforms as potential candidates for an immersive SLT application. Although both AR and tablet groups experience state-of-the-art systems which support native hand gesture recognition, the VR group did not.

In Chapter 6, a more recent HMD is evaluated (HTC Vive), as such allowing for a comparison between system related factors. Some minor increases in user perceived QoE is noted across explicit, implicit, and objective metrics. It is believed that more recent HMDs which have been brought to the market in the last year may be more suited to such an evaluation. Examples of such include the oculus quest [160] which currently supports native hand gestures and presents a completely wireless experience.

In addition to this, experiment 1 (Chapter 5) does not present an even distribution of age and gender for AR, tablet, or VR groups (Table 5.3). As such, the influence of gender (and age) bias is also to be considered as a limitation. To an extent, this was later addressed in experiment 2 (Chapter 6), which reports an even gender balance of 50% male and 50% female participants. Due to the similarity associated with objective measures of user performance between the VR groups in Chapters 5 and 6, it is believed this will have minimum impact on the results. However, future work aims to explore the utility of new and emerging technologies within the context of immersive SLT evaluations. Therefore, providing opportunity to address any concerns in full.

7.4 Future Work

The work presented within this thesis has demonstrated the utility of a novel methodology for the explicit, implicit, and objective capture of user perceived QoE. Utilizing such an approach, task-evoked measures of physiological response report accurate measures of arousal, and application context. Such novel approaches derive a fundamental understand of QoE at a deeper level, thus allowing developers to make more informed decisions. In parallel, the automated user performance reporting methodologies used in this work may be of significance within clinical evaluations in the future. As such, the primary focus of future work focuses on the following proposals.

7.4.1 *Continuous Feedback and Automated Reporting of QoE*

To a large degree, the methodology explored within this work is semi-automated apart from the post-processing requirements of implicit measures. In future work, the existing experiment could be extended to utilize ML in an effort to derive instant classification of emotional states as a user progresses through the application. This no doubt alleviates the efforts required with respect to post-processing of physiological signals from a QoE perspective. In addition to this, the advanced approach could lead to the development of adaptive experiences, which can change based on user requirements within the specified domain. By programmatically altering the presentation of content to derive more confidence from the user or present more challenging content based on current emotional state. Such an approach, however, may rely on the further capture of stress related physiological measures in an effort to avoid challenges associated with over or under fitting a ML classification model.

7.4.2 *Body Movement: A Supportive Measure of Physiological Response*

Immersive multimedia applications are driven by user-specific engagement which results in unique experiences. In this work, a novel methodology is proposed to assist in the QoE evaluation of these active, user centric experiences. Although physiological measures of arousal can be captured utilizing discreet sensor technologies, there are other measures which can be utilized to derive supporting measures of emotion. Examples of such include utilizing head movement [11], hand, or arm movement [175] as indicators of emotion within immersive VR experiences. Considering these novel measures, a multimodal feedback system which can derive more reliable measures of physiological response could be possible. As such allowing for a more accurate measure of user experience within QoE evaluations. In parallel, supporting a potential increase in performance for predictive algorithms in the future.

7.4.3 *A QoE Focused Clinical Evaluation and Validation*

The immersive multimedia technologies presented in this work demonstrate a great degree of suitability for the evaluation of receptive language disorders. Results for both AR and tablet groups (as described in Chapter 5) perform within one SD of the mean performance score of the paper-based assessment. This however was evaluated with participants who did not have a speech or language impairment. Future work may extend the evaluation to consider the clinical population in an effort to derive user perceived QoE from such a novel system from both the clinical population and clinician’s perspective. Therefore, delivering a fundamental overview of the quality perception process within an immersive health context. To the best of the authors knowledge, there has been no QoE evaluation which presents a holistic overview of a proposed system. It is proposed that the collective novel data set be presented to the community to further understand user perception.

APPENDICES

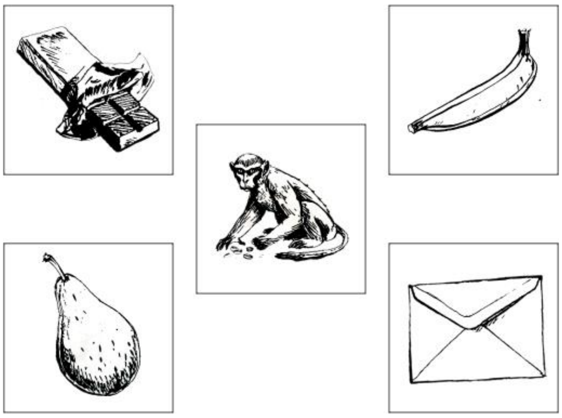
In the following pages, the appendices of this work are presented. Table A.1 provides a quick overview of each of the respective Appendix. Appendix 1 – 4, present the content, and participant documentation for the experimental evaluation carried in Chapters 5 and 6. Following this Appendix 4 – 8 contain a full overview of the statistical analysis for data captured in Chapter 5. These are presented as a comparative analysis between: (a) all three groups, (b) AR vs Tablet, (c) Tablet vs VR, and (d) AR vs VR. Appendix 9 and 10 present an overview of data captured throughout the training exercise in Chapter 5. These represent introductory data analysis and are provided for the sake of completeness.

Appendix 11 presents the MOS scores from the HTC Vive group (Chapter 6) in a corrected format (negative questions are inverted) to present a similar scale. The gaze density maps (heat maps) generated for each of the respective test stimulus is presented in appendix 12.

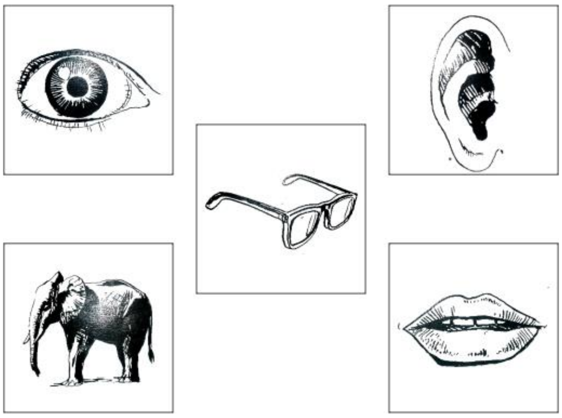
Table A.1: Appendices Overview

	Description
Appendix 1	Experiment 1 – Semantic Memory Assessment Content
Appendix 2	Experiment 1 – Post-Experience Questionnaire
Appendix 3	Experiment 1 – Information Sheet
Appendix 4	Experiment 1 – Participant Consent Form
Appendix 5	Experiment 1 – Post-Experience Questionnaire Analysis
Appendix 6	Experiment 1 – Heart Rate Analysis
Appendix 7	Experiment 1 – EDA Analysis
Appendix 8	Experiment 1 – Progression Time
Appendix 9	Experiment 1 – Participant EDA throughout the training phase
Appendix 10	Experiment 1 – Participant HR throughout the training phase
Appendix 11	HTC Vive – MOS Scores (Inverted)
Appendix 12	Gaze Density Maps (Heatmaps)

APPENDIX 1 – SEMANTIC MEMORY ASSESSMENT CONTENT

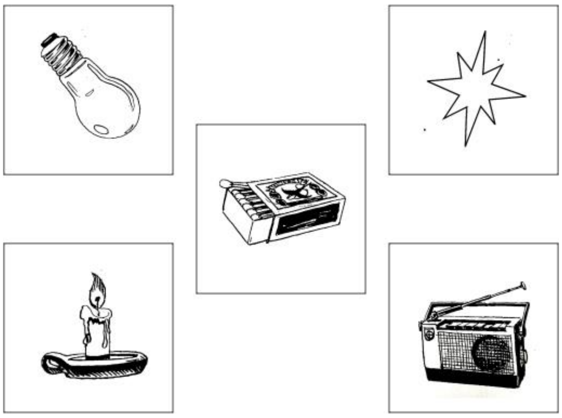


Slide P (Practice)

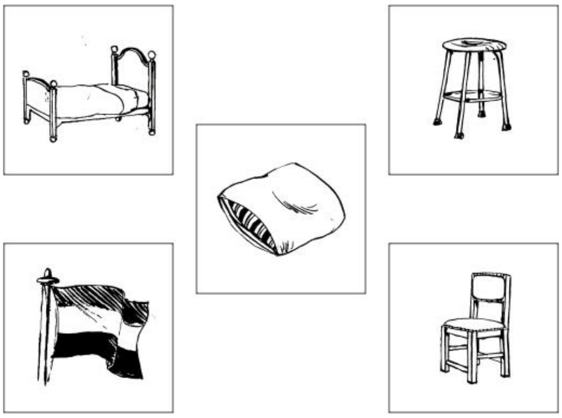


Slide 1

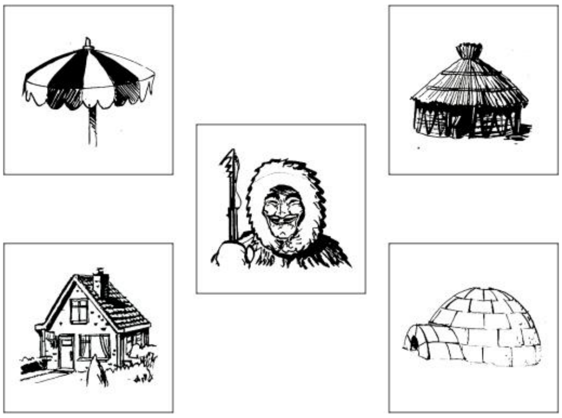
Slide 2



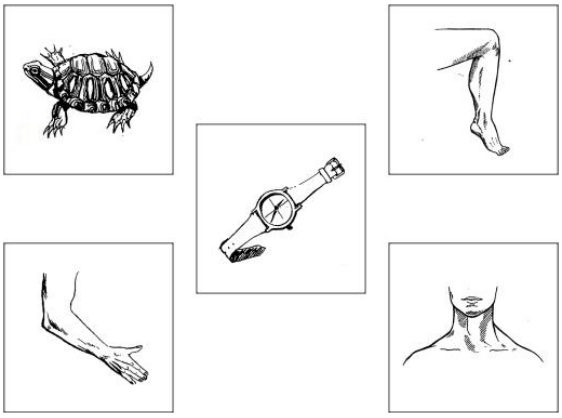
Slide 3



Slide 4

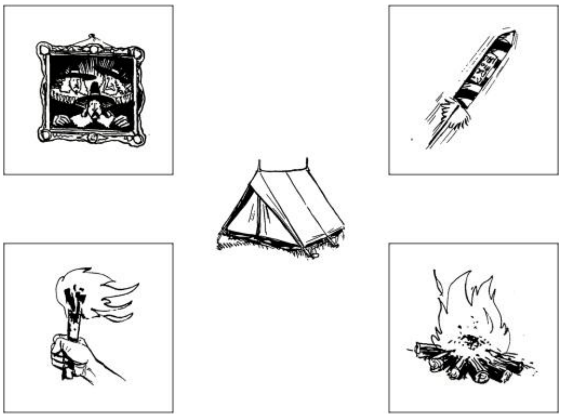


Slide 5

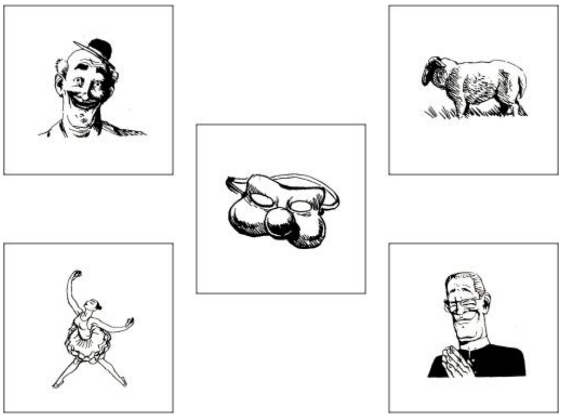


Slide 6

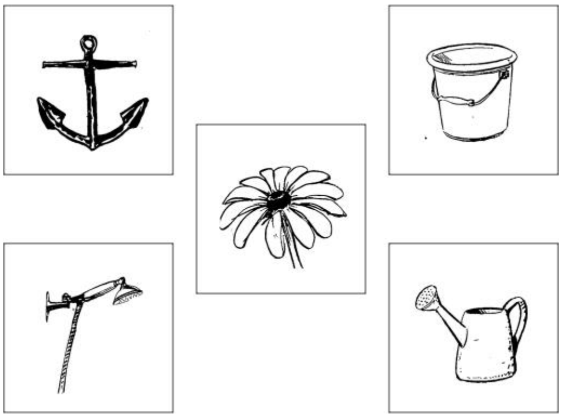
Slide 7



Slide 8



Slide 9



Slide 10

Figure A.1: Semantic Memory Assessment Content

APPENDIX 2 – POST-EXPERIENCE QUESTIONNAIRE

Q1: I did not feel any discomfort while using the system

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q2: I enjoyed the experience

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q3: My interaction with the assessment was natural

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q4: The device was annoying

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q5: I was immersed in the activity

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q6: I was engaged with the system while using it

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q7: The learning curve was not too great

Strongly Disagree	**Disagree**	**Neither Disagree or Agree**	**Agree**	**Strongly Agree**

Q8: I was restricted in my movements using the device

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q9: The system made me feel nauseous

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q10: The system was easy to use

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q11: The environment I was interacting with was real

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q12: I would like to experience this environment again

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q13: I did not feel a strong sense of presence whilst experiencing the system

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

Q14: My experience did not meet my expectations

Strongly Disagree	Disagree	Neither Disagree or Agree	Agree	Strongly Agree

APPENDIX 3 – INFORMATION SHEET

Information Sheet

Principle Investigator: Conor Keighrey

"Quality of Experience Evaluation of an Interactive Digital Speech & Language Assessment"

Brief explanation of title

In this experiment, we aim to evaluate user quality of experience whilst using multimedia devices to interact with a virtual representation of a speech and language assessment. The speech and language assessment in question focuses entirely on how participants form process language in terms of thoughts rather than expression through speech. Evaluation objectives include a comparative analysis of user performance when compared to that of a traditional paper-based form of this test. Furthermore, additionally metrics captured from non-invasive devices will provide more insight into the study fitting in the spectrum of Quality of experience assessment.

Introduction

I am inviting you to take part in a research experiment to be carried out in the Software Research Institute in Athlone Institute of Technology. The aim of this document is to explain why the research is being carried out and what it will involve.

If you are not clear on any points, please do not hesitate ask questions. Thank you for reading this.

What is the purpose of the project?

In this experiment, we aim to evaluate if a higher quality of experience is experienced using a different multimedia devices for an Interactive Speech and Language assessment. Results gathered from throughout this sub-study will be compared against previously captured data to determine if a multimedia approach to this type of assessment is more suitable.

Do I have to take part?

It is entirely up to you to decide whether you wish to take part in this experiment. Refusal to take part is entirely at your discretion. If you decide to take part, you can keep this information sheet and will be required to sign a consent form.

What does the experiment involve?

This experiment should last a minimum of 30 mins and no more than 40 minutes. Participants will be seated in a laboratory in the AIT Engineering Building. The lab will consist of a chair, table, multimedia device, PIP biosensor which the participant will hold in their hand, a wristband, and a PC to monitor participant progress. The test will involve participant interaction with a virtual representation of a Speech and Language assessment which are normally paper based.

The participant will be required to use hand gestures to interact with the virtual assessment while wearing non-intrusive devices to monitor heart rate and electrodermal activity. Lastly, participants will be asked to complete a questionnaire at the end of the experiment to give their thought on the quality of experience.

Appendices

What do I have to do?

Participants will undergo a visual screening to ensure they are eligible for the test. The visual screening process involves testing the participant's visual clarity using a Snellen chart, and testing a participant's colour perception using the Ishihara test.

If you are pregnant or suspect that you may be pregnant, please let the administrator of the test know.

What are the possible disadvantages and risks of taking part?

Due to the nature of this type of test there are no risks involved in taking part, participants will be requested to interact with a Tablet application which involves simple hand gestures.

Will my taking part in this project be kept confidential?

Any information collected during this test will be strictly confidential. All data will be stored in a secure manor and it will not be possible to recognise you from this experiment.

What will happen to the results of the research project?

The results of this experiment will be used to produce a paper for publication as part of my research.

Thanks!

Just like to say, thank you very much for your time and help with this experiment.

APPENDIX 4 – CONSENT FORM

Consent Form

Title of Project: *Quality of Experience Evaluation of an Interactive Digital Speech & Language Assessment.*

Name of Researcher: *Conor Keighrey*

Please Tick the Box

1. I confirm that I have read the information sheet dated __/__/20__ (Ver:) for the above study and have had the opportunity to ask questions. ☐
2. I am satisfied that I understand the information provided and have had enough time to consider the information. ☐
3. I do not suffer from photosensitive epilepsy or any other form of epilepsy. ☐
4. I'm not pregnant and/or I am not experiencing any symptoms of pregnancy. ☐
5. I understand that my participation is voluntary and that I am free to withdraw at any time, without giving any reason, without my legal rights being affected. ☐
6. I agree to take part in the above study. ☐
7. I understand that participant within this experiment will have no impact on grade should I be in attendance of a module carried out by the researcher. ☐
8. What age are you? (Optional) ____

________________________	________________	____________________
Name of Participant	Date	Signature
________________________	________________	____________________
Name of Person taking consent (if different from researcher)	Date	Signature
________________________	________________	____________________
Researcher	Date	Signature

APPENDIX 5 – POST-EXPERIENCE QUESTIONNAIRE ANALYSIS

Table A.2: Subjective Ratings – MANOVA (95% C.I.) – All Groups

	AR		**Tablet**		**VR**			
	Mean	SD	Mean	SD	Mean	SD	F	Sig
Question 1	3.900	1.021	4.450	1.099	3.550	1.468	1.475	0.238
Question 2	4.550	0.510	4.650	0.587	4.600	0.503	0.016	0.984
Question 3	3.650	0.988	4.450	0.686	4.050	0.945	1.624	0.207
Question 4	2.500	1.000	1.700	0.923	2.000	0.858	3.221	0.048
Question 5	4.350	0.587	4.150	0.813	4.350	0.489	1.223	0.302
Question 6	4.700	0.470	4.500	0.513	4.600	0.503	1.427	0.249
Question 7	3.350	1.663	3.050	1.761	3.350	1.599	0.074	0.929
Question 8	2.550	0.887	1.900	1.119	2.400	0.995	1.739	0.185
Question 9	1.150	0.366	1.250	0.550	1.300	0.470	0.103	0.902
Question 10	4.600	0.598	4.950	0.224	4.500	0.607	4.213	0.020
Question 11	3.450	1.146	3.900	1.119	3.900	0.641	0.067	0.935
Question 12	4.500	0.513	3.900	0.788	4.050	0.605	5.531	0.007
Question 13	2.650	0.875	3.050	1.317	3.300	1.174	2.256	0.115
Question 14	2.300	0.865	1.900	1.021	1.850	0.813	1.648	0.202

Table A.3: Subjective Ratings – MANOVA (95% C.I.) – AR vs Tablet Group

	AR		**Tablet**			
	Mean	SD	Mean	SD	F	Sig
Question 1	3.900	1.021	4.450	1.099	0.267	0.608
Question 2	4.550	0.510	4.650	0.587	0.012	0.914
Question 3	3.650	0.988	4.450	0.686	3.519	0.069
Question 4	2.500	1.000	1.700	0.923	5.770	0.022
Question 5	4.350	0.587	4.150	0.813	1.017	0.320
Question 6	4.700	0.470	4.500	0.513	2.532	0.120
Question 7	3.350	1.663	3.050	1.761	0.055	0.816
Question 8	2.550	0.887	1.900	1.119	2.741	0.106
Question 9	1.150	0.366	1.250	0.550	0.060	0.808
Question 10	4.600	0.598	4.950	0.224	3.386	0.074
Question 11	3.450	1.146	3.900	1.119	0.003	0.960
Question 12	4.500	0.513	3.900	0.788	10.380	0.003
Question 13	2.650	0.875	3.050	1.317	2.356	0.134
Question 14	2.300	0.865	1.900	1.021	2.131	0.153

Table A.4: Subjective Ratings – MANOVA (95% C.I.) – Tablet vs VR Group

	Tablet		**VR**			
	Mean	SD	Mean	SD	F	Sig
Question 1	4.450	1.099	3.550	1.468	2.345	0.134
Question 2	4.650	0.587	4.600	0.503	0.004	0.950
Question 3	4.450	0.686	4.050	0.945	1.414	0.242
Question 4	1.700	0.923	2.000	0.858	0.693	0.411
Question 5	4.150	0.813	4.350	0.489	2.094	0.157
Question 6	4.500	0.513	4.600	0.503	1.527	0.225
Question 7	3.050	1.761	3.350	1.599	0.141	0.710
Question 8	1.900	1.119	2.400	0.995	2.154	0.151
Question 9	1.250	0.550	1.300	0.470	0.039	0.844
Question 10	4.950	0.224	4.500	0.607	10.304	0.003
Question 11	3.900	1.119	3.900	0.641	0.091	0.765
Question 12	3.900	0.788	4.050	0.605	2.159	0.150
Question 13	3.050	1.317	3.300	1.174	0.275	0.603
Question 14	1.900	1.021	1.850	0.813	0.012	0.913

Table A.5: Subjective Ratings – MANOVA (95% C.I.) – AR vs VR Group

	AR		**VR**			
	Mean	SD	Mean	SD	F	Sig
Question 1	3.900	1.021	3.550	1.468	1.268	0.268
Question 2	4.550	0.510	4.600	0.503	0.036	0.851
Question 3	3.650	0.988	4.050	0.945	0.408	0.527
Question 4	2.500	1.000	2.000	0.858	2.945	0.095
Question 5	4.350	0.587	4.350	0.489	0.152	0.699
Question 6	4.700	0.470	4.600	0.503	0.150	0.701
Question 7	3.350	1.663	3.350	1.599	0.017	0.896
Question 8	2.550	0.887	2.400	0.995	0.024	0.878
Question 9	1.150	0.366	1.300	0.470	0.262	0.612
Question 10	4.600	0.598	4.500	0.607	0.985	0.328
Question 11	3.450	1.146	3.900	0.641	0.140	0.711
Question 12	4.500	0.513	4.050	0.605	4.126	0.050
Question 13	2.650	0.875	3.300	1.174	5.114	0.030
Question 14	2.300	0.865	1.850	0.813	3.102	0.087

APPENDIX 6 – EXPERIMENT 1 – HEART RATE ANALYSIS

Table A.6: Heart Rate – MANOVA – (95% C.I.) - All Groups

	AR		**Tablet**		**VR**			
	Mean	SD	Mean	SD	Mean	SD	F	Sig
Baseline								
Average	77.845	13.000	78.454	8.169	78.860	9.987	0.150	0.861
Test								
Slide P	75.526	13.775	79.663	10.506	78.225	9.866	1.193	0.311
Slide 1	75.764	13.600	80.621	10.978	78.380	9.859	1.353	0.267
Slide 2	76.195	13.414	80.770	10.455	78.960	9.843	1.105	0.338
Slide 3	76.880	13.397	80.439	10.493	78.945	9.627	0.627	0.538
Slide 4	77.641	13.465	80.325	10.637	79.446	9.624	0.359	0.700
Slide 5	77.714	13.471	80.201	10.663	79.348	9.378	0.359	0.700
Slide 6	78.177	13.714	79.942	10.654	79.387	9.323	0.301	0.741
Slide 7	78.779	14.212	79.902	10.633	79.454	9.050	0.234	0.792
Slide 8	78.899	14.210	80.116	10.485	80.033	9.360	0.305	0.738
Slide 9	79.900	13.768	81.152	13.200	80.615	9.726	0.228	0.797
Slide 10	79.899	13.854	80.512	11.970	80.694	9.840	0.191	0.827

Table A.7: Heart Rate – MANOVA – (95% C.I.) – AR vs Tablet Group

	AR		**Tablet**			
	Mean	SD	Mean	SD	F	Sig
Baseline						
Average	77.845	13.000	78.454	8.169	0.258	0.615
Test						
Slide P	75.526	13.775	79.663	10.506	2.122	0.154
Slide 1	75.764	13.600	80.621	10.978	2.393	0.131
Slide 2	76.195	13.414	80.770	10.455	1.973	0.169
Slide 3	76.880	13.397	80.439	10.493	1.104	0.300
Slide 4	77.641	13.465	80.325	10.637	0.626	0.434
Slide 5	77.714	13.471	80.201	10.663	0.621	0.436
Slide 6	78.177	13.714	79.942	10.654	0.517	0.477
Slide 7	78.779	14.212	79.902	10.633	0.383	0.540
Slide 8	78.899	14.210	80.116	10.485	0.515	0.478
Slide 9	79.900	13.768	81.152	13.200	0.329	0.570
Slide 10	79.899	13.854	80.512	11.970	0.305	0.584

Table A.8: Heart Rate – MANOVA – (95% C.I.) – Tablet vs VR Group

	Tablet		**VR**			
	Mean	SD	Mean	SD	F	Sig
			Baseline			
Average	78.454	8.169	78.860	9.987	0.209	0.650
			Test			
Slide P	79.663	10.506	78.225	9.866	0.624	0.435
Slide 1	80.621	10.978	78.380	9.859	0.886	0.353
Slide 2	80.770	10.455	78.960	9.843	0.622	0.435
Slide 3	80.439	10.493	78.945	9.627	0.335	0.567
Slide 4	80.325	10.637	79.446	9.624	0.162	0.689
Slide 5	80.201	10.663	79.348	9.378	0.209	0.650
Slide 6	79.942	10.654	79.387	9.323	0.213	0.647
Slide 7	79.902	10.633	79.454	9.050	0.247	0.622
Slide 8	80.116	10.485	80.033	9.360	0.260	0.613
Slide 9	81.152	13.200	80.615	9.726	0.304	0.585
Slide 10	80.512	11.970	80.694	9.840	0.211	0.649

Table A.9: Heart Rate – MANOVA – (95% C.I.) – AR vs VR Group

	AR		**VR**			
	Mean	SD	Mean	SD	F	Sig
			Baseline			
Average	77.845	13.000	78.860	9.987	0.018	0.894
			Test			
Slide P	75.526	13.775	78.225	9.866	0.683	0.414
Slide 1	75.764	13.600	78.380	9.859	0.637	0.430
Slide 2	76.195	13.414	78.960	9.843	0.595	0.445
Slide 3	76.880	13.397	78.945	9.627	0.358	0.553
Slide 4	77.641	13.465	79.446	9.624	0.231	0.633
Slide 5	77.714	13.471	79.348	9.378	0.195	0.662
Slide 6	78.177	13.714	79.387	9.323	0.137	0.714
Slide 7	78.779	14.212	79.454	9.050	0.064	0.802
Slide 8	78.899	14.210	80.033	9.360	0.115	0.737
Slide 9	79.900	13.768	80.615	9.726	0.019	0.891
Slide 10	79.899	13.854	80.694	9.840	0.039	0.845

APPENDIX 7 – EXPERIMENT 1 – EDA ANALYSIS

Table A.10: Electrodermal Activity – MANOVA (95% C.I.) – All Groups

	AR		**Tablet**		**VR**			
	Mean	SD	Mean	SD	Mean	SD	F	Sig
Baseline								
Average	3.142	1.294	2.997	1.128	2.896	1.205	0.617	0.543
Test								
Slide P	3.584	1.054	4.539	1.448	3.278	1.148	6.252	0.004
Slide 1	3.691	1.095	4.579	1.432	3.297	1.105	6.750	0.002
Slide 2	3.656	1.143	4.551	1.418	3.293	1.115	6.524	0.003
Slide 3	3.650	1.151	4.498	1.434	3.301	1.120	5.928	0.005
Slide 4	3.616	1.181	4.465	1.444	3.291	1.134	5.795	0.005
Slide 5	3.570	1.192	4.482	1.481	3.300	1.124	5.754	0.005
Slide 6	3.589	1.202	4.424	1.425	3.339	1.133	4.850	0.012
Slide 7	3.575	1.241	4.389	1.405	3.371	1.165	4.282	0.019
Slide 8	3.561	1.246	4.408	1.437	3.360	1.157	4.714	0.013
Slide 9	3.621	1.186	4.402	1.432	3.398	1.183	4.499	0.016
Slide 10	3.618	1.177	4.414	1.449	3.422	1.173	4.374	0.017

Table A.11: Electrodermal Activity – MANOVA (95% C.I.) – AR vs Tablet Group

	AR		**Tablet**			
	Mean	SD	Mean	SD	F	Sig
Baseline						
Average	3.142	1.294	2.997	1.128	0.000	0.991
Test						
Slide P	3.584	1.054	4.539	1.448	5.110	0.030
Slide 1	3.691	1.095	4.579	1.432	5.011	0.031
Slide 2	3.656	1.143	4.551	1.418	4.896	0.033
Slide 3	3.650	1.151	4.498	1.434	4.476	0.041
Slide 4	3.616	1.181	4.465	1.444	4.339	0.044
Slide 5	3.570	1.192	4.482	1.481	4.776	0.035
Slide 6	3.589	1.202	4.424	1.425	3.920	0.055
Slide 7	3.575	1.241	4.389	1.405	3.563	0.067
Slide 8	3.561	1.246	4.408	1.437	4.280	0.046
Slide 9	3.621	1.186	4.402	1.432	3.922	0.055
Slide 10	3.618	1.177	4.414	1.449	3.968	0.054

Table A.12: Electrodermal Activity – MANOVA (95% C.I.) – Tablet vs VR Group

	Tablet		**VR**			
	Mean	SD	Mean	SD	F	Sig
Baseline						
Average	2.997	1.128	2.896	1.205	1.079	0.306
Test						
Slide P	4.539	1.448	3.278	1.148	10.554	0.003
Slide 1	4.579	1.432	3.297	1.105	11.883	0.001
Slide 2	4.551	1.418	3.293	1.115	11.816	0.001
Slide 3	4.498	1.434	3.301	1.120	10.686	0.002
Slide 4	4.465	1.444	3.291	1.134	10.621	0.002
Slide 5	4.482	1.481	3.300	1.124	10.238	0.003
Slide 6	4.424	1.425	3.339	1.133	8.888	0.005
Slide 7	4.389	1.405	3.371	1.165	7.989	0.008
Slide 8	4.408	1.437	3.360	1.157	8.451	0.006
Slide 9	4.402	1.432	3.398	1.183	7.923	0.008
Slide 10	4.414	1.449	3.422	1.173	7.545	0.009

Table A.13: Electrodermal Activity – MANOVA (95% C.I.) – AR vs VR Group

	AR		**VR**			
	Mean	SD	Mean	SD	F	Sig
Baseline						
Average	3.142	1.294	2.896	1.205	0.857	0.361
Test						
Slide P	3.584	1.054	3.278	1.148	1.176	0.285
Slide 1	3.691	1.095	3.297	1.105	1.574	0.218
Slide 2	3.656	1.143	3.293	1.115	1.405	0.244
Slide 3	3.650	1.151	3.301	1.120	1.253	0.270
Slide 4	3.616	1.181	3.291	1.134	1.223	0.276
Slide 5	3.570	1.192	3.300	1.124	0.856	0.361
Slide 6	3.589	1.202	3.339	1.133	0.792	0.379
Slide 7	3.575	1.241	3.371	1.165	0.626	0.434
Slide 8	3.561	1.246	3.360	1.157	0.468	0.498
Slide 9	3.621	1.186	3.398	1.183	0.592	0.447
Slide 10	3.618	1.177	3.422	1.173	0.484	0.491

APPENDIX 8 – EXPERIMENT 1 – PROGRESSION TIME

Table A.14: Progression Time – MANOVA (95% C.I.) – All Group

	AR		**Tablet**		**VR**			
	Mean	SD	Mean	SD	Mean	SD	F	Sig
Test								
Slide P	4.387	2.219	5.115	5.372	8.808	9.424	2.316	0.108
Slide 1	7.177	14.104	3.571	1.153	6.066	6.226	0.713	0.495
Slide 2	4.494	1.607	3.543	1.253	4.656	2.526	1.958	0.151
Slide 3	4.758	1.811	3.845	1.458	7.079	4.177	8.729	0.001
Slide 4	4.104	1.366	3.459	1.339	6.051	7.476	4.279	0.019
Slide 5	4.145	1.495	3.605	1.427	5.135	3.117	2.507	0.091
Slide 6	4.694	2.577	3.529	1.480	6.555	5.428	8.852	0.000
Slide 7	4.990	1.974	5.023	4.803	8.960	7.580	5.137	0.009
Slide 8	6.811	3.797	5.066	2.068	11.459	11.479	9.103	0.000
Slide 9	7.425	5.675	6.634	2.646	15.838	15.866	4.619	0.014
Slide 10	5.319	2.780	4.438	1.861	6.216	2.945	2.869	0.065

Table A.15: Progression Time – MANOVA (95% C.I.) – AR vs Tablet Group

	AR		**Tablet**			
	Mean	SD	Mean	SD	F	Sig
Test						
Slide P	4.387	2.219	5.115	5.372	0.112	0.739
Slide 1	7.177	14.104	3.571	1.153	0.505	0.482
Slide 2	4.494	1.607	3.543	1.253	2.085	0.157
Slide 3	4.758	1.811	3.845	1.458	3.928	0.055
Slide 4	4.104	1.366	3.459	1.339	1.763	0.193
Slide 5	4.145	1.495	3.605	1.427	1.762	0.193
Slide 6	4.694	2.577	3.529	1.480	1.717	0.198
Slide 7	4.990	1.974	5.023	4.803	0.240	0.627
Slide 8	6.811	3.797	5.066	2.068	6.677	0.014
Slide 9	7.425	5.675	6.634	2.646	0.006	0.936
Slide 10	5.319	2.780	4.438	1.861	2.093	0.157

Table A.16: Progression Time – MANOVA (95% C.I.) – Tablet vs VR Group

	Tablet		**VR**			
	Mean	SD	Mean	SD	F	Sig
Test						
Slide P	5.115	5.372	8.808	9.424	2.170	0.149
Slide 1	3.571	1.153	6.066	6.226	6.222	0.017
Slide 2	3.543	1.253	4.656	2.526	3.464	0.071
Slide 3	3.845	1.458	7.079	4.177	13.061	0.001
Slide 4	3.459	1.339	6.051	7.476	5.211	0.028
Slide 5	3.605	1.427	5.135	3.117	3.936	0.055
Slide 6	3.529	1.480	6.555	5.428	13.968	0.001
Slide 7	5.023	4.803	8.960	7.580	6.096	0.018
Slide 8	5.066	2.068	11.459	11.479	12.957	0.001
Slide 9	6.634	2.646	15.838	15.866	5.347	0.027
Slide 10	4.438	1.861	6.216	2.945	6.224	0.017

Table A.17: Progression Time – MANOVA (95% C.I.) – AR vs VR Group

	AR		**VR**			
	Mean	SD	Mean	SD	F	Sig
Test						
Slide P	4.387	2.219	8.808	9.424	3.250	0.080
Slide 1	7.177	14.104	6.066	6.226	0.069	0.795
Slide 2	4.494	1.607	4.656	2.526	0.488	0.489
Slide 3	4.758	1.811	7.079	4.177	5.626	0.023
Slide 4	4.104	1.366	6.051	7.476	3.415	0.073
Slide 5	4.145	1.495	5.135	3.117	1.255	0.270
Slide 6	4.694	2.577	6.555	5.428	6.609	0.014
Slide 7	4.990	1.974	8.960	7.580	5.975	0.020
Slide 8	6.811	3.797	11.459	11.479	5.576	0.024
Slide 9	7.425	5.675	15.838	15.866	4.481	0.041
Slide 10	5.319	2.780	6.216	2.945	0.776	0.384

APPENDIX 9 – EXPERIMENT 1 – EDA

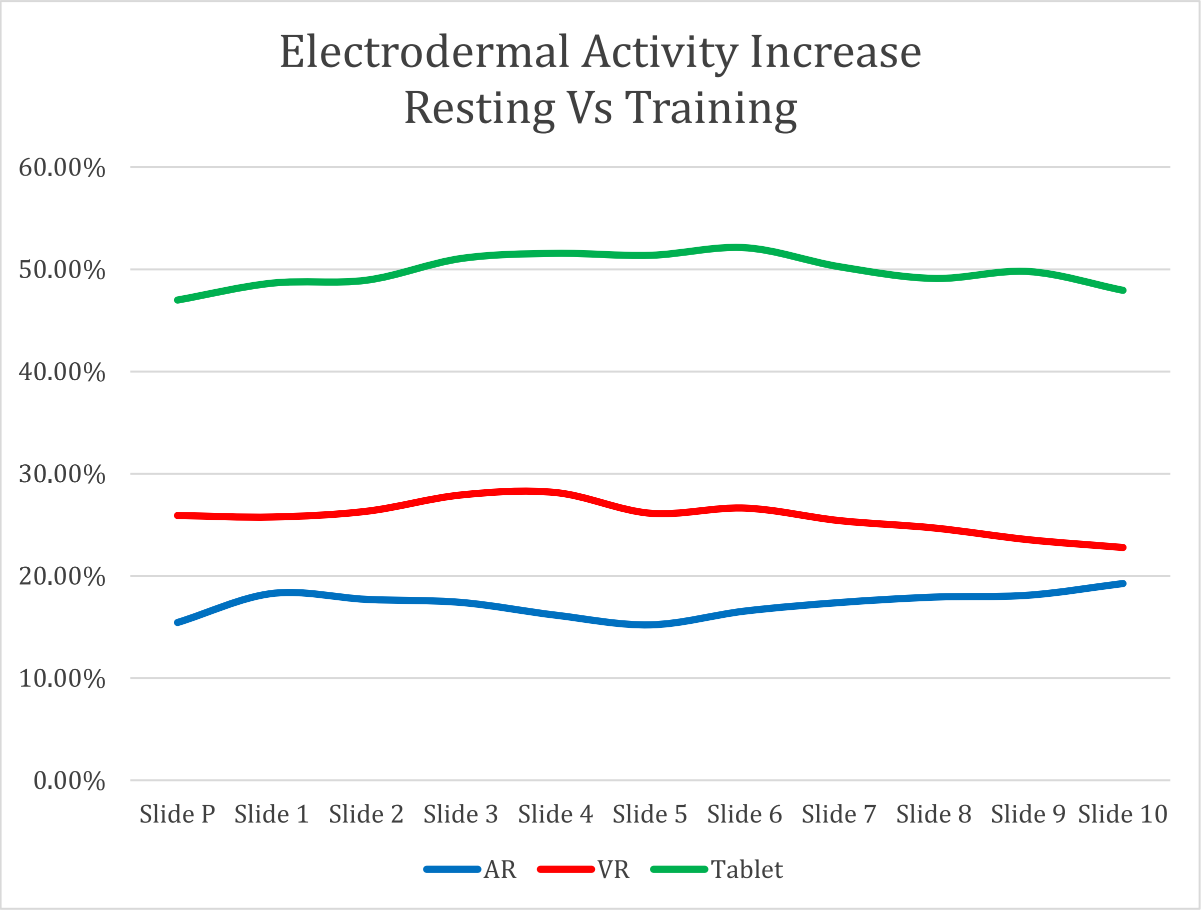


Figure A.2: Experiment 1 – Average EDA (Training Phase)

NB: These findings are presented for the sake of completeness. As such, the data within the table only serves as a preliminary overview to provide insight into signal trend. Future work will further validate and explore the statistical significative of the data.

APPENDIX 10 – EXPERIMENT 1 – HEART RATE

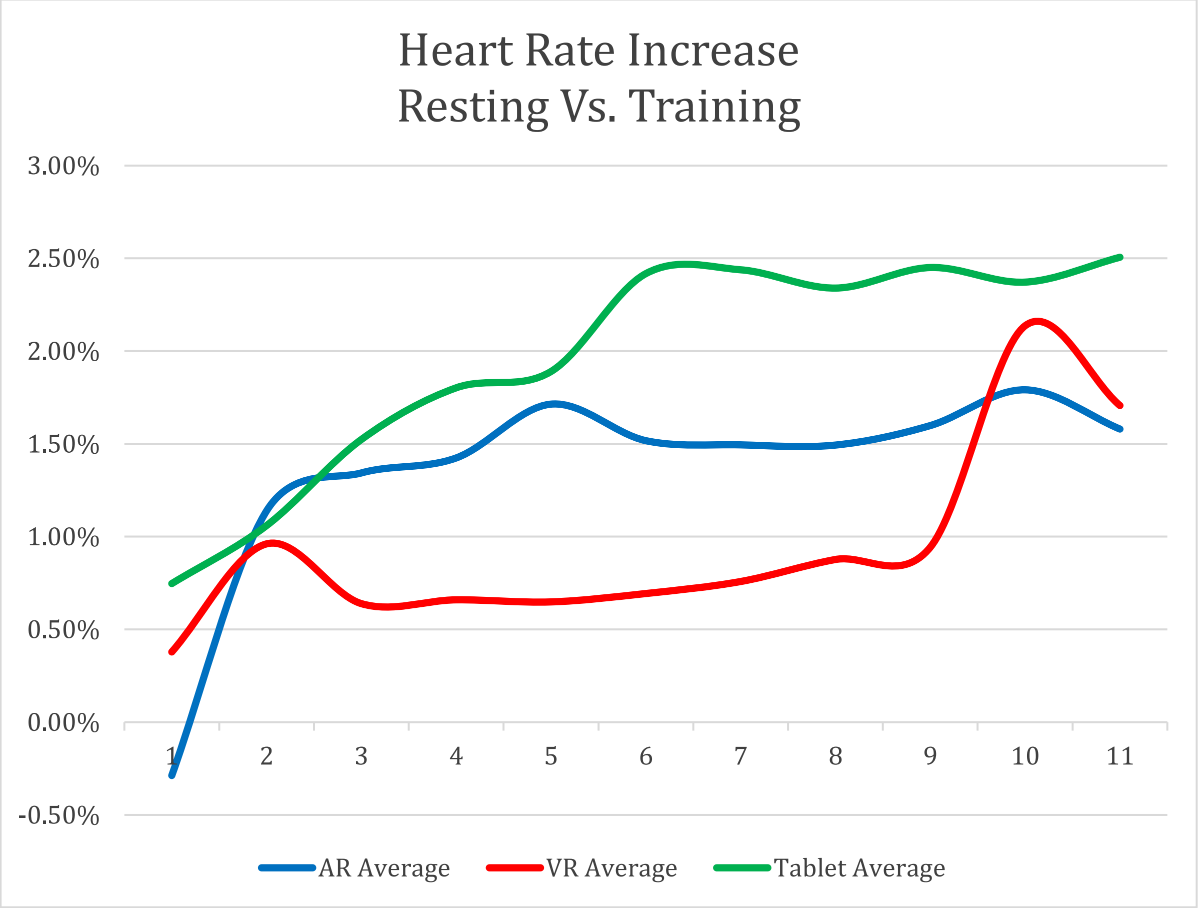


Figure A.3: Experiment 1 – Average Heart Rate (Training Phase)

NB: These findings are presented for the sake of completeness. As such, the data within the table only serves as a preliminary overview to provide insight into signal trend. Future work will further validate and explore the statistical significative of the data.

Appendix 11 – HTC Vive – MOS Scores (Inverted)

Table A.18: HTC Vive – Inverted MOS – Summative & Gender Split Results

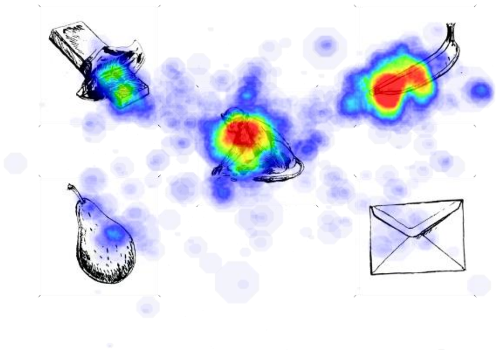
		Group		**Male**		**Female**			
		MOS	**SD**	**MOS**	**SD**	**MOS**	**SD**	**F**	**Sig**
Discomfort	**Q1**	3.821	1.467	3.929	1.385	3.714	1.590	0.877	0.707
	Q4	3.893	1.031	3.143	0.864	2.643	1.151	0.000	1.000
	Q8	3.571	1.103	2.500	1.092	2.643	1.151	0.014	0.848
	Q9	4.643	0.678	3.571	0.852	3.714	0.469	1.298	0.205
Immersion	**Q2**	4.786	0.418	4.786	0.426	4.786	0.426	0.037	0.564
	Q12	4.607	0.567	4.643	0.497	4.571	0.646	0.022	0.826
	Q14	4.214	0.833	3.214	0.699	3.214	0.975	40.847	0.003
Interaction	**Q3**	3.893	0.956	3.929	0.997	3.857	0.949	0.066	0.739
	Q7	2.643	1.496	3.429	1.651	1.857	0.770	1.611	0.587
	Q10	4.607	0.685	4.643	0.842	4.571	0.514	0.190	0.789
Enjoyment	**Q5**	4.500	0.638	4.571	0.646	4.429	0.646	7.374	0.103
	Q6	4.536	0.838	4.571	0.852	4.500	0.855	1.066	0.746
	Q11	3.750	1.041	3.429	1.222	4.071	0.730	2.774	0.651
	Q13	3.179	1.219	2.071	0.997	2.286	1.437	1.620	1.000

NB: Those highlighted in green report inverted values for negatively phrased questions, the content of this table has been utilized to create Figure 6.9.

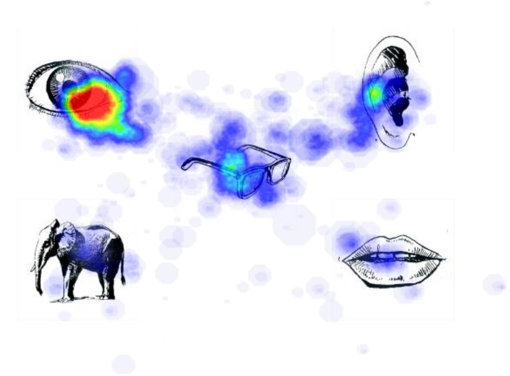
APPENDIX 12 – GAZE DENSITY MAPS (HEATMAPS)

TAP THE IMAGE WHEN READY TO BEGIN

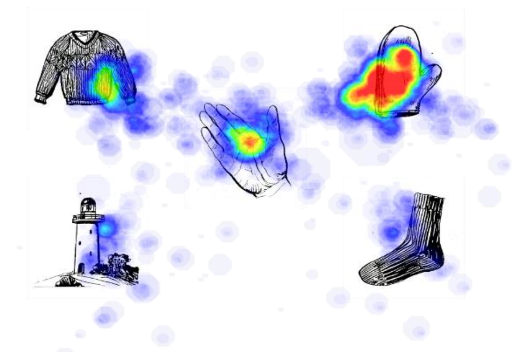
Launch Screen



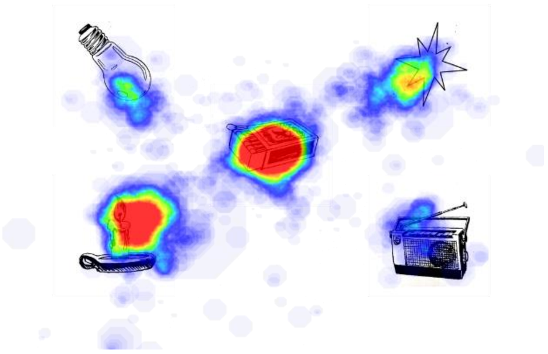
Practice Slide



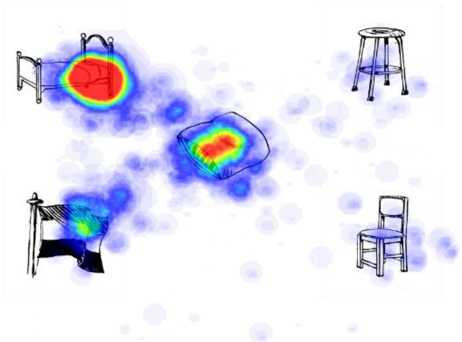
Slide 1



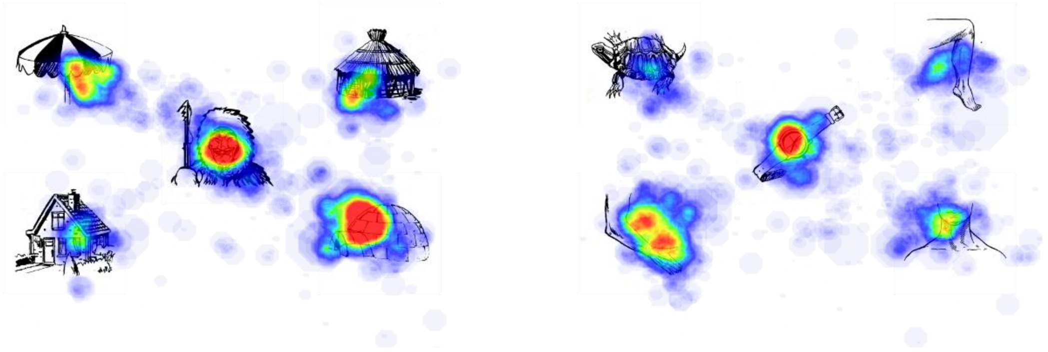
Slide 2



Slide 3

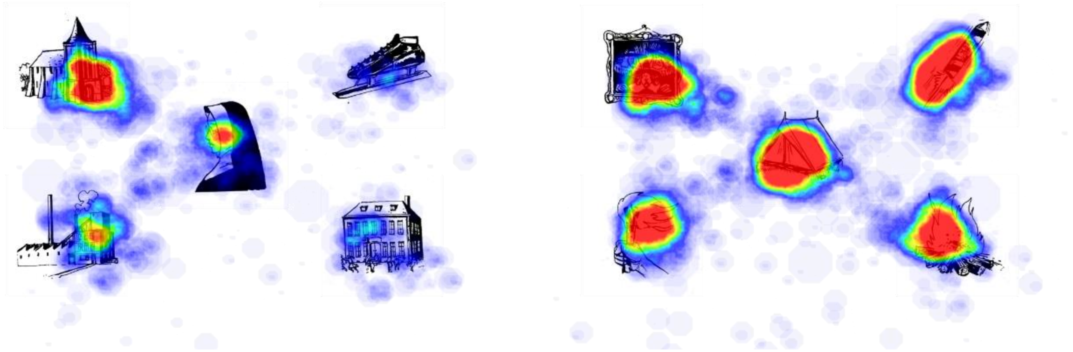


Slide 4



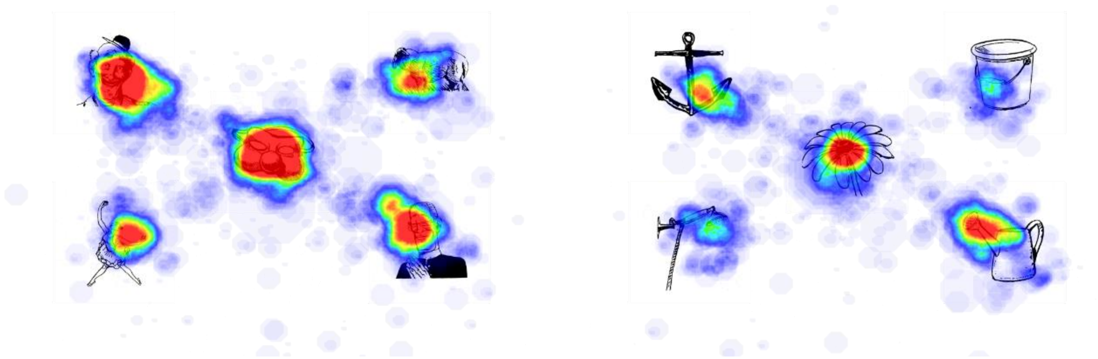
Slide 5

Slide 6



Slide 7

Slide 8



Slide 9

Slide 10

Figure A.4: Semantic Memory Assessment – Gaze Density Maps

REFERENCES

[1] “VIVE™ | Discover Virtual Reality Beyond Imagination,” HTC, [Online]. Available: https://www.vive.com/eu/. [Accessed 21 March 2017].

[2] “Microsoft HoloLens,” Microsoft, [Online]. Available: https://www.microsoft.com/microsoft-hololens/en-us. [Accessed 02 February 2017].

[3] HP, “HP Windows Mixed Reality Headset | HP® Official Site,” HP Development Company, [Online]. Available: https://www8.hp.com/us/en/campaigns/mixedrealityheadset/overview.html. [Accessed 12 September 2018].

[4] A. Perkis, C. Timmerer and et al., “QUALINET White Paper on Definitions of Immersive Media Experience (IMEx),” in *14th QUALINET meeting (online), May 25, 2020,* Athlone Ireland (Online), 2020.

[5] A. J. A. M. van Deursen, S. ben Allouch and L. P. Ruijter , “Tablet Use In Primary Education: Adoption Hurdles and Attitude Determinants,” *Education and Information Technologies,* vol. 21, pp. 971 - 991, 2016.

[6] J. Incao, “How VR is Transforming the Way We Train Associates,” Walmart, 20 9 2018. [Online]. Available: https://blog.walmart.com/innovation/20180920/how-vr-is-transforming-the-way-we-train-associates. [Accessed 25 10 2018].

[7] Audi AG, “Audi uses virtual reality to train Logistics employees | Audi MediaCenter,” Audi AG, [Online]. Available: https://www.audi-mediacenter.com/en/press-releases/audi-uses-virtual-reality-to-train-logistics-employees-9259. [Accessed 15 September 2018].

[8] O. Soto-Martin, A. Fuentes-Porto and J. Martin-Gutierrez, “A Digital Reconstruction of a Historical Building and Virtual Reintegration of Mural Paintings to Create an Interactive and Immersive Experience in Virtual Reality,” *Applied Sciences,* vol. 10, no. 2, pp. 1-22, 2019.

[9] D. Egan, C. Keighrey, J. Barret, Y. Qiao, S. Brennan, C. Timmerer and N. Murray, “Subjective Evaluation of an Olfaction Enhanced Immersive Virtual Reality Environment,” in *AltMM '17 Proceedings of the 2nd International Workshop on Multimedia Alternate Realities,* Mountain View, California, USA, 2017.

[10] S. Arndt, M. Ervik and A. Perkis, “Factors of Immersion in Interactive Digital Storytelling,” in *International Conference on Interactive Digital Storytelling,* Funchal, Madeira, Portugal, 2017.

[11] E. Hynes, R. Flynn, B. Lee and N. Murray, “A QoE Evaluation of an Augmented Reality Procedure Assistance Application,” in *Twelfth International Conference on Quality of Multimedia Experience (QoMEX),* Athlone, Ireland, 2020.

[12] American Speech-Language-Hearing Association, “Treatment Efficacy Summary,” ASHA, 2008.

[13] J. Law, A. van der Gaag, B. Hardcastle, J. Beck, A. MacGregor and C. Plunkett, “Communication Support Needs: a Review of the Literature,” Scottish Executive Social Research, Edinburgh, 2007.

[14] “Scope of Practice in Speech-Language Pathology,” [Online]. Available: http://www.asha.org/uploadedFiles/SP2016-00343.pdf. [Accessed 20 January 2017].

[15] K. G. Shipley and J. G. McAfee, Assessment in Speech-Language Pathology: A Resource Manual, Cengage Learning, 2016.

[16] N. Hinds and C. M. Wiles, “Assessment of Swallowing and Referral to Speech and Language Therapists in Acute Stroke,” *QJM - Monthly Journal of the Association of Physicians,* vol. 91, no. 12, pp. 829 - 835, 1998.

[17] L. M. M. S. A. Qassem, H. A. Hawai, S. A. Shehhi, J. Zemerly and J. W. Ng, “AIR-EDUTECH: Augmented Immersive Reality (AIR) Technology for High School Chemistry Education,” in *7th IEEE Global Engineering Education Conference,* Abu Dhabi, UAE, 2016.

[18] O. G. D., B. Martin-Gorriz, I. B. I., M. M. A., A. S. G. and B. Miguel, “Development and Assessment of a Tractor Driving Simulator with Immersive Virtual Reality for Training to Avoid Occupational Hazards,” *Computers and Electronics in Agriculture* , vol. 143, p. 111–118, 2017.

[19] H. Z. Jahromi, B. Rooney, D. T. Delaney and A. Hines, “Establishing Waiting Time Thresholds in Interactive Web Mapping Applications for Network QoE Management,” in *Irish Signals and Systems Conference (ISSC),* Maynooth, Ireland, Ireland , 2019.

[20] N. Côté and J. Berger, “Speech Communication,” in *Quality of Experience: Advanced Concepts, Applications and Methods,* Switzerland, Springer International Publishing, 2014, pp. 165 - 178.

[21] A. Drachen, L. E. Nacke, G. Yannakakis and A. L. Pedersen, “Correlation between Heart Rate, Electrodermal Activity and Player Experience in First-Person Shooter Games,” in *SIGGRAPH Symposium on Video Games,* Los Angeles, California, 2010.

[22] J. Puig, A. Perkis, F. Lindseth and T. Ebrahimi, “Towards an Efficient Methodology for Evaluation of Quality of Experience in Augmented Reality,” in *Quality of Multimedia Experience (QoMEX),* Yarra Valley, VIC, Australia, 2012.

[23] M. Varela, L. Skorin-Kapov, K. De Moor and P. Reichl, “QoE−Defining a User-Centric Concept for Service Quality,” in *Multimedia Quality of Experience (QoE): Current Status And Future Requirments,* Wiley, 2016, pp. 5-28.

[24] T. B. Rodrigues, C. Ó Catháin, N. E. O’Connor and N. Murray, “A Quality of Experience Assessment of Haptic and Augmented Reality Feedback Modalities in a Gait Analysis System,” *PLOS ONE,* vol. 15, no. 3, pp. 1 - 19, 2020.

[25] S. Arndt, A. Perkis and J.-N. Voigt-Antons, “Using Virtual Reality and Head-Mounted Displays to Increase Performance in Rowing Workouts,” in *1st*

International Workshop on Multimedia Content Analysis in Sports (MMSports), Seoul, Republic of Korea, 2018.

[26] T. Ebrahimi, “Quality of Multimedia Experience: Past, Present and Future,” in *Proceedings of the 17th ACM international conference on Multimedia*, New York, United States, 2009.

[27] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chat, N. Ramzan and K. Brunnström, “Psychophysiology-Based QoE Assessment: A Survey,” *IEEE Journal of Selected Topics in Signal Processing,* vol. 11, no. 1, pp. 6 - 21, 2017.

[28] G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall and C.-W. Wallesch, Linguistic Disorders and Pathologies: An International Handbook, Berlin: Mouton De Gruyter, 1993.

[29] G. Conti-Ramsden, K. Durkin, U. Toseeb, N. Botting and A. Pickles, “Education and Employment Outcomes of Young Adults With a History of Developmental Language Disorder,” *International Journal of Language & Communication Disorders* , vol. 53, no. 2, pp. 237-255, 2018.

[30] B. E. Murdoch, Acquired Speech and Language Disorders: A neuroanatomical and functional neurological approach, Bristol: Chapman and Hall, 1990.

[31] K. G. Shipley and J. G. McAfee, “Assessment of Neurocognitive Disorders,” in *Assessment in Speech-Language Pathology: A Resource Manual,* 2016, Cengage Learning, 2016, pp. 457 - 506.

[32] World Health Organization, “ICD-10 Version:2010,” [Online]. Available: http://apps.who.int/classifications/icd10/browse/2010/en#/F80. [Accessed 08 November 2017].

[33] American Speech Language Hearing Association, “Preschool Language Disorders,” American Speech Language Hearing Association, [Online]. Available: https://www.asha.org/public/speech/disorders/Preschool-Language-Disorders/. [Accessed 13 November 2017].

[34] J. Boyle, E. McCartney, J. Forbes and A. O'Hare, “A Randomised Controlled Trial And Economic Evaluation Of Direct Versus Indirect And Individual Versus Group Modes Of Speech And Language Therapy For Children With Primary Language Impairment.,” *Health Technology Assessment,* vol. 11, no. 25, 2007.

[35] N. Botting and G. Conti-Ramsden, “Social and behavioural difficulties in children with language impairment,” *Child Language Teaching and Therapy,* vol. 16, no. 2, pp. 105-120, 2017.

[36] A. Hines and J. D. Kelleher, “A Framework for Post-Stroke Quality of Life Prediction using Structured Prediction,” in *International Conference on Quality of Multimedia Experience*, Erfurt, Germany, 2017.

[37] B. E. Murdoch, “Bostonian and Lurian aphasia syndromes,” in *Acquired Speech and Language Disorders: A neuroanatomical and functional neurological approach,* Springer, 1990, pp. 60-96.

[38] A. Cotoi, K. Salter, A. Jinah, J. Brar, S. Macaluso, J. Orange and R. Teasell, “Aphasia and Apraxia,” in *Evidence-Based Review of Stroke Rehabilitation*, Evidence-Based Review of Stroke Rehabilitation, 2016.

[39] R. Lesser, “Aphasia Therapy,” in *Linguistic Disorders and Pathologies: An International Handbook*, Berlin, Walter de Gruyter & Co, 1993, pp. 352-372.

[40] The National Aphasia Association, “Aphasia Definitions - National Aphasia Association,” The National Aphasia Association, [Online]. Available: https://www.aphasia.org/aphasia-definitions/. [Accessed 10 August 2018].

[41] “Samsung Gear VR,” Samsung, [Online]. Available: http://www.samsung.com/global/galaxy/gear-vr/. [Accessed 26 March 2017].

[42] “Oculus Rift Development Kit 2,” Oculus, [Online]. Available: https://www.oculus.com/en-us/dk2/. [Accessed 02 February 2016].

[43] “PlayStation VR | PlayStation.com,” PlayStation, [Online]. Available: https://www.playstation.com/en-ie/explore/playstation-vr/. [Accessed 20 March 2017].

[44] “Unity - Game Engine,” Unity3D, [Online]. Available: https://unity3d.com/. [Accessed 07 April 2017].

[45] “Leap Motion,” Leap Motion, [Online]. Available: https://www.leapmotion.com/. [Accessed 07 April 2017].

[46] Ultraleap, “Tracking | Leap Motion Controller | Ultraleap,” Ultraleap, [Online]. Available: https://www.ultraleap.com/product/leap-motion-controller/. [Accessed 20 January 2020].

[47] Acer, “Windows Mixed Reality Headset | Virtual & Mixed Reality | Acer United States,” Acer, [Online]. Available: https://www.acer.com/ac/en/US/content/series/wmr. [Accessed 18 September 2018].

[48] “Epson Moverio Augmented Reality Smart Glasses,” Epson, [Online]. Available: https://epson.com/moverio-augmented-reality. [Accessed 2017 March 26].

[49] Vuzix, “Vuzix Blade® | Augment Reality (AR) Smart Glasses for the Consumer,” Vuzix, [Online]. Available: https://www.vuzix.com/products/blade-smart-glasses. [Accessed 20 October 2018].

[50] Magic Leap, “Welcome | Magic Leap,” Magic Leap, [Online]. Available: https://www.magicleap.com/. [Accessed 12 October 2018].

[51] P. T. Kovács, N. Murray, G. Rozinaj , Y. Sulema and R. Rybárová, “Application of Immersive Technologies for Education: State Of The Art,” in *International Conference on Interactive Mobile Communication Technolo,* Thessaloniki, 2015.

[52] A. Covaci, G. Ghinea, C.-H. Lin, S.-H. Huang and J.-L. Shih, “Multisensory Games-based Learning - Lessons Learnt From Olfactory Enhancement Of A Digital Board Game,” *Multimedia Tools and Applications,* vol. 77, no. 16, p. 21245–21263, 2018.

[53] M. Obrist, C. Velasco, C. Vi, N. Ranasinghe, A. Israr, A. Cheok, C. Spence and P. Gopalakrishnakone, “Sensing the future of HCI: touch, taste, and smell user interfaces,” Sussex Research Online, 2016.

[54] C. Timmerer, M. Waltl, R. Benjamin and N. Murray, “Sensory Experience: Quality of Experience beyond Audio-Visual,” in *Quality of Experience: Advanced Concepts, Applications and Methods,* Springer, 2014, pp. 351-365.

[55] G. Ghinea, C. Timmerer, W. Lin and S. R. Gulliver , “Mulsemedia: State-of-the-Art, Perspectives and Challenges,” *ACM Transactions on Multimedia Computing, Communications and Applications,* vol. 11, no. 1S, 2014.

[56] A. Covaci, E. B. Saleme, G. A. Mesfin, N. Hussain, E. Kani-Zabihi and G. Ghinea, “How Do We Experience Crossmodal Correspondent Mulsemedia Content?,” *IEEE Transactions on Multimedia,* vol. 22, no. 5, pp. 1249-1258, 2019 (Early Access Article).

[57] Z. Yuan, T. Bi, G.-M. Muntean and G. Ghinea, “Perceived Synchronization of Mulsemedia Services,” *IEEE Transactions on Multimedia,* vol. 17, no. 7, pp. 957 - 966, 2015.

[58] N. Ranasinghe, T. N. Tram Nguyen, Y. Liangkun, L.-Y. Lin, D. Tolley and E. Yi-Luen Do, “Vocktail: A Virtual Cocktail for Pairing Digital Taste, Smell, and Color Sensations,” in *ACM Multimedia,* Mountain View, California, 2017.

[59] L. Shams and A. R. Seitz, “Benefits of Multisensory Learning,” *Trends in Cognitive Sciences,* vol. 12, no. 11, pp. 411 - 417, 2008.

[60] H. Gardner, Frames of Mind: The Theory of Multiple Intelligences, Fontana Press, 1993.

[61] U. Jekosch, Voice and Speech Quality Perception: Assessment and Evaluation (Signals and Communication Technology), Berlin, Heidelberg: Springer, 2005.

[62] “COST | European Network on Quality of Experience in Multimedia Systems and Services (QUALINET),” COST - European Cooperation in Science and Technology, [Online]. Available: http://www.cost.eu/COST_Actions/ict/IC1003?. [Accessed 15 August 2017].

[63] “QUALINET - COST IC 1003,” Qualinet - European Network on Quality of Experience in Multimedia Systems and Services, [Online]. Available: http://www.qualinet.eu/. [Accessed 15 August 2018].

[64] P. Le Callet, S. Möller and A. Perkis, “Qualinet White Paper on Definitions of Quality of Experience,” European Network on Quality of Experience in Multimedia Systems and Services (COST Action IC 1003), Lausanne, Switzerland, 2012.

[65] A. Raake and S. Möller, “Quality and Quality of Experience,” in *Quality of Experience: Advanced Concepts, Applications and Methods,* Switzerland, Springer International Publishing, 2014, pp. 11 - 33.

[66] U. Reiter, K. Brunnström, K. De Moor, M.-C. Larabi, M. Pereira, A. Pinheiro, J. You and A. Zgank, “Factors Influencing Quality of Experience,” in *Quality of*

Experience: Advanced Concepts, Applications, and Methods, Springer, 2014, pp. 55-72.

[67] K. Lukas and P. Le Callet, “Emerging science of QoEin multimedia applications: Concepts, experimentalguidelines, and validationof models,” *Academic Press Library in Signal Processing,* vol. 6, pp. 163-209, 2018.

[68] A. Sackl and R. Schatz, “Evaluating the Influence of Expectations, Price and Content Selection on Video Quality Perception,” in *Sixth International Workshop on Quality of Multimedia Experience (QoMEX)*, Singapore, 2014.

[69] K. Swinburn, G. Porter and D. Howard, Comprehensive Aphasia Test, Psychology Press, 2004.

[70] L. M. Dunn and D. M. Dunn, Peabody Picture Vocabulary Test, Pearson Education, 2007.

[71] D. Howard and K. Patterson, The Pyramids and Palm Trees Test, Bury St. Edmunds, UK: Thames Valley Test Company, 1992.

[72] J. Willis, “Teaching the Brain to Read,” in *Strategies for Improving Fluency, Vocabulary, and Comprehension,* Alexandria, Virginia USA, Association for Supervision and Curriculum Development, 2008, pp. 103-105.

[73] Pearson Education, “Q-interactive Overview | Pearson Assessment,” Pearson Education, [Online]. Available: https://www.pearsonclinical.co.uk/q-interactive/q-interactive.aspx. [Accessed 2018 September 21].

[74] Pearson Education, “Pearson Clinical | Home,” Pearson, [Online]. Available: https://www.pearsonclinical.com/. [Accessed 21 September 2018].

[75] D. Wahlstrom, “Want to Ditch Paper and Pencil Tests? | Psych Central Professional,” Psych Central, 06 August 2016. [Online]. Available: https://pro.psychcentral.com/want-to-ditch-paper-and-pencil-tests/. [Accessed 21 August 2018].

[76] Therapy Figures, “Scoring Software for the Comprehensive Aphasia Test (CAT),” Therapy Figures, [Online]. Available: http://www.therapyfigures.com/. [Accessed 02 May 2018].

[77] Microsoft, “Microsoft Excel Software,” Microsoft, [Online]. Available: https://products.office.com/en-ie/excel. [Accessed 15 September 2018].

[78] PLORAS, “eCAT User Manual,” [Online]. Available: https://www.ucl.ac.uk/drupal/site_ploras/sites/ploras/files/eCAT_User_Manual-2.1_eCAT-1.7_Nov2017.pdf. [Accessed 25 September 2018].

[79] PLORAS, “Project Information | Predicting Language Outcome and Recovery After Stroke (PLORAS) - UCL - London's Global University,” PLORAS, [Online]. Available: https://www.ucl.ac.uk/ploras/project-information-versions/project-info-standard. [Accessed 25 October 2018].

[80] D. Egan, S. Brennan, J. Barrett, Y. Qiao, C. Timmerer and N. Murray, “An Evaluation of Heart Rate and Electrodermal Activity as an Objective QoE Evaluation Method for Immersive Virtual Reality Environments,” in *8th*

International Conference on Quality of Multimedia Experience, Lisbon, Portugal, 2016.

[81] "ITU-T P.910 : Subjective video quality assessment methods for multimedia applications," [Online]. Available: https://www.itu.int/rec/T-REC-P.910-199909-S/en. [Accessed 01 February 2017].

[82] "ITU-T P.913 : Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment," ITU-T, [Online]. Available: https://www.itu.int/rec/T-REC-P.913/en. [Accessed 01 February 2017].

[83] W. M. Vagias, "Likert-type Scale Response Anchors," Clemson International Institute for Tourism & Research Development, Department of Parks, Recreation andTourism Management., Clemson University, 2006.

[84] T. Hoßfeld, R. Schatz and S. Egger, "SOS: The Mos Is Not Enough!," in *Quality of Multimedia Experience (QoMEX),* Mechelen, Belgium, 2011.

[85] T. Vigier, M. Perreira da Silva and P. Le Callet, "Impact of Viewing Immersion on Visual Behavior in Videos," in *Sino-French Workshop on information and communication technology,* Qingdao, China., 2017.

[86] A.-N. Moldovan, I. Ghergulescu and S. Weibelzahl, "User-centered EEG-based multimedia quality assessment," in *Broadband Multimedia Systems and Broadcasting (BMSB),* London, UK, 2013.

[87] W. Janig, "Autonomic Nervous System," in *Human Physiology,* Berlin, Springer, 1989, pp. 333-369.

[88] R. F. Schmidt and G. Thews, Human Physiology, Berlin: Springer-Verlag, 1989.

[89] A.-M. Brouwer, M. Neerincx, V. L. Kallen and M. t. Brinke, "EEG alpha asymmetry, heart rate variability and cortisol in response to Virtual Reality induced stress," *Journal of CyberTherapy & Rehabilitation,* vol. 4, no. 1, pp. 27-40, 2011.

[90] EMOTIV, "EMOTIV | Brain Data Measuring Hardware and Software Solutions," Emotiv, [Online]. Available: https://www.emotiv.com/. [Accessed 20 November 2017].

[91] D. L. Schacter, "EEG Theta Waves and Psychological Phenomena: a Review and Analysis," *Biological Psychology,* vol. 5, no. 1, pp. 47 - 82, 1977.

[92] D. Pereira Salgado, F. Roque Martins, T. Braga Rodrigues, C. Keighrey, R. Flynn, E. L. Martins Naves and N. Murray, "A Qoe Assessment Method Based on Eda, Heart Rate and Eeg of a Virtual Reality Assistive Technology System," in *9th ACM Multimedia Systems Conference*, Amsterdam, 2018.

[93] S. Davis, K. Nesbitt and E. Nalivaiko, "Comparing the onset of cybersickness using the Oculus Rift and two virtual roller coasters," in *Proceedings of the 11th Australasian Conference on Interactive Entertainment* , Sydney, 2015.

[94] C. Keighrey, R. Flynn, S. Brennan, S. Murray and N. Murray, "Comparing User QoE via Physiological and Interaction Measurements of Immersive AR and VR

Speech and Language Therapy Applications," in *Thematic Workshops '17 Proceedings of the on Thematic Workshops of ACM Multimedia 2017* , Mountain View, California, USA, 2017.

[95] "Fitbit Charge HR," Fitbit, [Online]. Available: https://www.fitbit.com/chargehr. [Accessed 10 February 2017].

[96] Fitbit, Inc, "Technology," Fitbit, Inc, 2020. [Online]. Available: https://www.fitbit.com/ie/technology#purepulse. [Accessed 10 January 2020].

[97] D. G. Kilpatrick, "Differential responsiveness of two electrodermal indices to psychological stress and performance of a complex cognitive task," *Psychophysiology,* vol. 9, no. 2, pp. 218-226, 1972.

[98] W. Boucsein, D. C. Fowles, S. Grimnes, G. Ben-Shakhar, W. T. Roth, M. E. Dawson and D. L. Filion, "Publication Recommendations for Electrodermal Measurements," *Psychophysiology,* vol. 49, no. 8, pp. 1017 -1034, 2012.

[99] iMotions, "Galvanic Skin Response: The Complete Pocket Guide," iMotions, 2017.

[100] F. Paas, J. E. Tuovinen, H. Tabbers and P. W. M. Van Gerven , "Cognitive Load Measurement as a Means to Advance Cognitive Load Theory," *Educational Psychologist,* vol. 38, no. 1, pp. 63-71, 2003.

[101] Y. Shi, N. Ruiz, R. Taib, E. Choi and E. Choi, "Galvanic Skin Response (GSR) as an Index of Cognitive Load," in *ACM Conference on Human Factors in Computing Systems,* San Jose, California, 2007.

[102] H. Critchley and Y. Nagai, "Electrodermal Activity (EDA)," in *Encyclopedia of Behavioral Medicine*, New York, NY, Springer, 2013, pp. 666 - 669.

[103] R. Buettner, "Cognitive Workload of Humans Using Artificial Intelligence Systems: Towards Objective Measurement Applying Eye-Tracking Technology," in *KI 2013: Advances in Artificial Intelligence,* Springer, 2013, pp. 37-48.

[104] PyGaze, "PyGaze | Open source eye-tracking software and more," P.S. Dalmaijer, [Online]. Available: http://www.pygaze.org/. [Accessed 15 April 2018].

[105] A. Voßkühler, "OGAMA (OpenGazeAndMouseAnalyzer): An open source software designed to analyze eye and mouse movements in slideshow study designs | open gaze and mouse analyzer," OGAMA, [Online]. Available: http://www.ogama.net/. [Accessed 15 June 2018].

[106] Tobii Pro, "Eye tracking technology for research - Tobii Pro," Tobii Pro, [Online]. Available: https://www.tobiipro.com/. [Accessed 18 October 2018].

[107] iMotions, "Webcam-Based Eye Tracking vs. an Eye Tracker [Pros & Cons]," iMotions, [Online]. Available: https://imotions.com/blog/webcam-eye-tracking-vs-an-eye-tracker/. [Accessed 18 October 2018].

[108] J. A. Büttner-Ennever, "Anatomy of the Oculomotor System," *Neuro-Ophthalmology,* vol. 40, pp. 1 - 14, 2007.

[109] K. Holmqvist and R. Andersson, "Eye movements: biology, neurology, and psychology," in *Eye-tracking: A comprehensive guide to methods, paradigms and*

measures, Lund, Sweeden, Lund Eye-Tracking Research Institute, 2017, pp. 12-33.

[110] J. Beatty, “Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources,” *Psychological Bulletin,* vol. 91, no. 2, pp. 276 - 292, 1982.

[111] M. M. Bradley, L. Miccoli, M. A. Escrig and P. J. Lang, “The Pupil as a Measure of Emotional Arousal and Autonomic Activation,” *Psychophysiology,* vol. 45, pp. 602 - 607, 2008.

[112] P. Reeves, “Rate of Pupillary Dilation and Contraction,” *Psychological Review,* vol. 25, no. 4, pp. 330 - 340, 1918.

[113] F. A. Young and W. R. Biersdorf, “Pupillary Contraction and Dilation in Light and Darkness,” *Journal of Comparative and Physiological Psychology,* vol. 47, no. 3, pp. 264 - 268, 1954.

[114] K. Holmqvist and R. Andersson, “Gaze Density Maps - Scientific Tools or Fancy Visualizations?,” in *Eye Tracking: A Comprehensive Guide to Methods and Measures*, Lund, Sweden:, Lund Eye-Tracking Research Institute, 2017, pp. 302 - 326.

[115] K. Holmqvist and R. Andersson, “Areas of Interest,” in *Eye Tracking : A Comprehensive Guide to Methods and Measures,* Lund, Sweden, Lund Eye-Tracking Research Institute, 2017, pp. 254 - 298.

[116] K. Holmqvist and R. Andersson, “Scanpaths – Theoretical Principals and Practical Application,” in *Eye-tracking: A comprehensive guide to methods, paradigms and measures,* Lund, Sweeden, Lund Eye-Tracking Research Institute, 2017, pp. 327-359.

[117] R. S. Hessels, C. Kemner, C. van den Boomen and I. T. C. Hooge , “The area-of-interest problem in eyetracking research: A noise-robust solution for face and sparse stimuli,” *Behavior Research Methods,* vol. 48, no. 4, pp. 1694-1712, 2015.

[118] Tobii AB, “Tobii Pro Glasses 2 Wearable Eye Tracking,” Tobii AB, 20 March 2020. [Online]. Available: https://www.tobiipro.com/product-listing/tobii-pro-glasses-2/. [Accessed 20 March 20].

[119] A. Adams and A. L. Cox, “Questionnaires, In-depth Interviews and Focus Groups,” in *Research Methods for Human Computer Interaction,* Cambridge, UK, Cambridge University Press, 2008, pp. 17-34.

[120] R. Schatz, S. Egger-Lampl and K. Masuch, “The Impact of Test Duration on User Fatigue and Reliability of Subjective Quality Ratings,” *Journal of the Audio Engineering Society,* vol. 60, no. 1, pp. 63 - 73, 2012.

[121] M. M. Bradley and P. J. Lang, “Measuring Emotion: The Self-assessment Manikin and the Semantic Differential,” *Journal of Behavior Therapy and Experimental Psychiatry,* vol. 25, no. 1, pp. 49-59, 1994.

[122] A. Mehrabian, Basic Dimensions for a General Psychological Theory Implications for Personality, Social, Environmental, and Developmental Studies, Oelgeschlager,Gunn & Hain Inc, 1980.

[123] S. Rigby and R. Ryan, “The Player Experience of Need Satisfaction (PENS),” Immersyve Inc., 2007.

[124] A. Denisova, A. I. Nordin and P. Cairns, “The Convergence of Player Experience Questionnaires,” in *Proceedings of the 2016 Annual Symposium on Computer-Human Interaction in Play,* Texas, 2016.

[125] C. Jennett, A. L. Cox, P. Cairns, S. Dhoparee, A. Epps, T. Tijs and A. Walton, “Measuring and Defining the Experience of Immersion in Games,” *International Journal of Human-Computer Studies,* vol. 66, no. 9, pp. 641-661, 2008.

[126] J. H. Brockmyer, C. M. Fox, K. A. Curtiss, E. McBroom, K. M. Burkhart and J. N. Pidruzny, “The Development of the Game Engagement Questionnaire: a Measure of Engagement in Video Game-playing,” *Journal of Experimental Social Psychology,* vol. 45, no. 4, pp. 624-634, 2009.

[127] C. Zhang, A. Perkis and . S. Arndt , “Spatial Immersion Versus Emotional Immersion, Which Is More Immersive?,” in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX),* Erfurt, Germany, 2017.

[128] M. Waltl, C. Timmerer and H. Hellwagner, “Improving the Quality of Multimedia Experience Through Sensory Effects,” in *Quality of Multimedia Experience (QoMEX),* Trondheim, Norway, 2010.

[129] F. Chen, J. Zhou, Y. Wang, K. Yu, S. Z. Arshad, A. Khawaji and C. Dan, “Stress and Cognitive Load,” in *Robust Multimodal Cognitive Load Measurement,* Springer International Publishing, 2016, pp. 185-194.

[130] G. F. Wilson, “An Analysis of Mental Workload in Pilots During Flight Using Multiple Psychophysiological Measures,” *The International Journal of Aviation Psychology,* vol. 12, no. 1, pp. 3-18, 2002.

[131] F. Danieau, J. Fleureau, P. Guillotel, N. Mollet, M. Christie and A. Lecuyer, “Toward Haptic Cinematography: Enhancing Movie Experiences with Camera-Based Haptic Effects,” *IEEE Multimedia,* vol. 21, no. 2, pp. 11-21, 2014.

[132] O. Baus and S. Bouchard, “Exposure to an Unpleasant Odour Increases the Sense of Presence in Virtual Reality,” *Virtual Reality,* vol. 21, no. 2, pp. 59-74, 2017.

[133] K. Williams, K. Moffatt, D. McCall and L. Findlater, “Designing Conversation Cues on a Head-Mounted Display to Support Persons with Aphasia,” in *ACM Conference on Human Factors in Computing Systems,* Seoul, South Korea, 2015.

[134] C. Stapleton, J. Whiteside, J. Davies, D. Mott and J. Vick, “Transforming Lives Through Story Immersion: Innovation of Aphasia Rehabilitation Therapy through Storytelling Learning Landscapes,” in *Proceedings of the 2nd ACM International Workshop on Immersive Media Experiences* , Orlando, Florida, 2014.

[135] K. Williams, K. Moffatt, J. Hong, Y. Faroqi-Shah and L. Findlater, “The Cost of Turning Heads: A Comparison of a Head-Worn Display to a Smartphone for Supporting Persons with Aphasia in Conversation,” in *Proceedings of the 18th International ACM SIGACCESS Conference on Computers and Accessibility* , Reno, Nevada, USA, 2016.

[136] S.-E. Moon and L. Jong-Seok, “Perceptual Experience Analysis for Tone-Mapped HDR Videos Based on EEG and Peripheral Physiological Signals,” *IEEE Transactions on Autonomous Mental Development* , vol. 7, no. 3, pp. 236 - 247, 2015.

[137] T. Holmdahl, “BUILD 2015: A closer look at the Microsoft HoloLens hardware - Microsoft Devices Blog,” Microsoft, 30 April 2015. [Online]. Available: https://blogs.windows.com/devices/2015/04/30/build-2015-a-closer-look-at-the-microsoft-hololens-hardware/#rzHeAAFYz5itGCHr.97. [Accessed 05 September 2018].

[138] O. Kreylos, “Doc-Ok.org | A developer's perspective on immersive 3D computer graphics,” Doc-Ok, 7 May 2018. [Online]. Available: http://doc-ok.org/?p=1274. [Accessed 10 October 2018].

[139] Microsoft, “Personal Digital Assistant - Cortana Home Assistant - Microsoft,” Microsoft, [Online]. Available: https://www.microsoft.com/en-us/cortana. [Accessed 05 September 2018].

[140] Lenovo, “Lenovo TAB 2 A10-70 | 10" Gaming Entertainment Tablet | Lenovo Ireland,” Lenovo, [Online]. Available: https://www.lenovo.com/ie/en/tablets/android-tablets/tablets-a-series/Lenovo-TAB-2-A10-70F/p/ZZITZTAA01F. [Accessed 05 September 2018].

[141] S. Möller, K.-P. Engelbrecht, C. Kühnel, I. Wechsung and B. Weiss, “A Taxonomy Of Quality Of Service And Quality Of Experience Of Multimodal Human-machine Interaction,” in *First IEEE International Workshop on Quality of Multimedia Experience*, San Diego, 2009.

[142] M. E. Dawson, A. M. Schell and D. L. Filion, “The Electrodermal System,” in *Handbook of Psychophysiology*, Cambridge, England, Cambridge University Press, 2007, pp. 159-181.

[143] Fitbit, “Fitbit Charge HR™ Wireless Heart Rate + Activity Wristband,” Fitbit, Inc., [Online]. Available: https://www.fitbit.com/ie/chargehr. [Accessed 17 July 2018].

[144] “PIP Biosensor,” PIP, [Online]. Available: https://thepip.com/en-eu/. [Accessed 01 February 2017].

[145] PIP Data Recorder User Guide, Galvanic Ltd., 2015.

[146] M. Handler, R. Nelson, D. Krapohl and C. R. Honts, “EDA Primer for Polygraph Examiners,” *Polygraph,* vol. 39, no. 2, pp. 68-108, 2010.

[147] National Eye Institute, “Facts About Color Blindness | National Eye Institute,” National Eye Institute, [Online]. Available: https://nei.nih.gov/health/color_blindness/facts_about. [Accessed 15 October 2018].

[148] Committee on Vision, Assembly of Behavioral and Social Sciences, National Research Council, “Procedures for Testing Color Vision: Report of Working Group 41,” National Academy Press, 1981.

[149] J. Birch, “Efficiency of the ishihara test for identifying red-green coiour deficiency,” *Ophthalmic & Physiological Optics,* vol. 17, no. 5, pp. 403-408, 1997.

[150] H. Snellen, Probebuchstaben zur Bestimmung der Sehschärfe, H. Peters, 187..

[151] EyeBulletin, “Eye Chart: Download Free Snellen Chart For Eye Test - Eye Bulletin,” EyeBulletin, 10 08 2017. [Online]. Available: https://eyebulletin.com/eye-chart/. [Accessed 10 August 2017].

[152] Fitbit, “Fitbit Charge HR Product Manual 1.2,” Fitbit, 14 04 2016. [Online]. Available: https://staticcs.fitbit.com/content/assets/help/manuals/manual_charge_hr_en_US.pdf. [Accessed 15 August 2018].

[153] “ISO 8589:2007 Sensory analysis — General guidance for the design of test rooms,” International Standards Organisation, [Online]. Available: https://www.iso.org/obp/ui/#iso:std:iso:8589:ed-2:v1:en. [Accessed 24 April 2017].

[154] “Pokémon Go,” Niantic, [Online]. Available: http://www.pokemongo.com/. [Accessed 02 February 2017].

[155] “IBM SPSS - IBM Analytics,” IBM, [Online]. Available: https://www.ibm.com/analytics/us/en/technology/spss/. [Accessed 07 April 2017].

[156] T. Zhan, K. Yin, J. Xiong, Z. He and S.-T. Wu, “Augmented Reality and Virtual RealityDisplays: Perspectives and Challenges,” *iScience,* vol. 23, no. 8, pp. 1 - 13, 2020.

[157] V. Angelov, E. Petkov, G. Shipkovenski and T. Kalushkov , “Modern Virtual Reality Headsets,” in *International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA),* Ankara, Turkey, Turkey, 2020.

[158] Tobii Pro AB, “Calibration - Tobii Pro SDK documentation,” Tobii Pro AB, [Online]. Available: http://developer.tobiipro.com/commonconcepts/calibration.html. [Accessed 10 July 2020].

[159] Tobii Pro AB, “Coordinate systems - Tobii Pro SDK documentation,” Tobii Pro AB, 2020. [Online]. Available: http://developer.tobiipro.com/commonconcepts/coordinatesystems.html. [Accessed 10 July 2020].

[160] Facebook Technologies, LLC., “Oculus Quest: All-in-One VR Headset | Oculus,” Oculus, [Online]. Available: https://www.oculus.com/quest/. [Accessed 10 July 2020].

[161] Tobii 2020, “Tobii Pro Lab - Software for eye tracking and biometrics,” [Online]. Available: https://www.tobiipro.com/product-listing/tobii-pro-lab/. [Accessed 15 May 2020].

[162] Tobii AB, “Tobii Pro SDK documentation,” Tobii, [Online]. Available: http://developer.tobiipro.com/. [Accessed 15 May 2020].

[163] Unity Technologies, “Unity - Manual: Order of Execution for Event Functions,” Unity Technologies, 13 07 2020. [Online]. Available: https://docs.unity3d.com/Manual/ExecutionOrder.html. [Accessed 18 June 2020].

[164] Unity Technologies, “Unity - Scripting API: Physics.Raycast,” Unity Technologies, [Online]. Available: https://docs.unity3d.com/ScriptReference/Physics.Raycast.html. [Accessed 15 October 2019].

[165] Unity Technologies, “Unity - Manual: Colliders,” Unity Technologies, [Online]. Available: https://docs.unity3d.com/Manual/CollidersOverview.html. [Accessed 10 June 2020].

[166] U. Engelke, H. Liu, J. Wang, P. Le Callet, I. Heynderickx, H.-J. Zepernick and A. Maeder, “A Comparative Study of Fixation Density Maps,” *IEEE Transactions on Image Processing,* vol. 22, no. 3, pp. 1121 - 1133, 2013.

[167] G. Amayeh, G. Bebis and M. Nicolescu, “Gender Classification from Hand Shape,” in *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops,* Anchorage, AK, USA, 2008.

[168] M. Nakayama and Y. Shimizu, “Frequency Analysis of Task Evoked Pupillary Response and Eye-movement,” in *Proceedings of the 2004 Symposium on Eye Tracking Research & Applications,* San Antonio, Texas, 2004.

[169] L. Boon-Leng, L. Dae-Seok and L. Boon-Giin , “Mobile-based Wearable-Type of Driver Fatigue Detection by GSR and EMG,” in *IEEE Region 10 Conference (TENCON),* Macao, China, 2015.

[170] W. Boucsein, Electrodermal Activity, Springer, 2012.

[171] L. Shu, J. Xie, M. Yang, Z. Li, Z. Li, D. Liao, X. Xu and X. Yang, “A Review of Emotion Recognition Using Physiological Signals,” *Sensors,* vol. 18, no. 7, pp. 1 - 41, 2018.

[172] M. B. Winn, D. Wendt, T. Koelewijn and S. E. Kuchinsky, “Best Practices and Advice for Using Pupillometry to Measure Listening Effort: An Introduction for Those Who Want to Get Started,” *Trends in Hearing,* vol. 22, pp. 1 - 32, 2018.

[173] M. E. Kret and E. E. Sjak-Shie, “Preprocessing pupil size data: Guidelines and code,” *Behavior Research Methods,* vol. 51, p. 1336–1342, 2019.

[174] S. Mathôt, “Pupillometry: Psychology, Physiology, and Function,” *Journal of Cognition,* vol. 1, no. 16, 2018.

[175] I.-O. Stathopoulou and G. A. Tsihrintzis, “Emotion Recognition from Body Movementsand Gestures,” in *Proceedings of the 4th International Conference on Intelligent Interactive Multimedia Systems and Services (IIMSS 2011),* Piraeus-Athens, Greece, 2011.