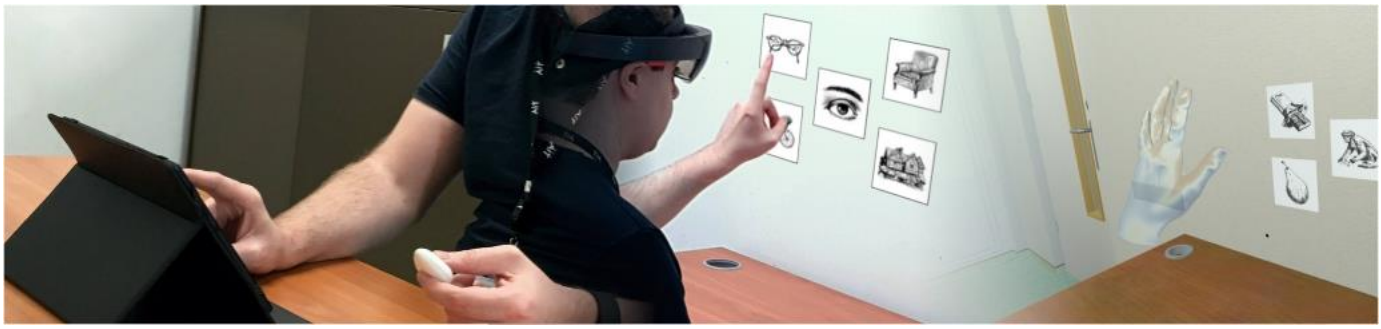
# A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications

Conor Keighrey, *Member, IEEE*, Ronan Flynn *Member, IEEE*, Siobhan Murray and Niall Murray, *Member, IEEE*



***Abstract*—The availability of affordable head-mounted display technology has facilitated new, potentially more immersive, interactive multimedia experiences. These technologies were traditionally focused on entertainment; however, academia and industry are now exploring applications in other domains such as health, learning and training. Key to the success of these new multimedia experiences is the understanding of a user's perceived quality of experience (QoE). Subjective user ratings have been the primary mechanism to capture insights into a user's experience. Such ratings have generally been captured post experience and reflected using a mean opinion score (MOS). However, user perception is multifactorial and subjective ratings alone do not express the true measure of an experience. As a result, recent efforts to capture QoE have included exploring the use of implicit metrics (e.g. physiological measures).**

**This article presents the results of an experimental QoE evaluation and comparison of immersive applications delivered across three multimedia platforms. The platforms compared were augmented reality, tablet and virtual reality. The QoE methodology employed considered explicit (post-test questionnaire) and implicit (heart rate and electrodermal activity) assessment methods. The results indicate comparatively higher levels of QoE for users of the augmented reality and tablet platforms.**

***Index Terms*—Quality of experience, augmented reality, virtual reality, physiological, speech language pathology, aphasia.**

This work was supported by the Irish Research Council - Government of Ireland Postgraduate Scholarship (grant number: GOIPG/2018/1314)

C. Keighrey is with the Athlone Institute of Technology, Athlone, Ireland. (email: c.keighrey@research.ait.ie)

R. Flynn is with the Athlone Institute of Technology, Athlone, Ireland. (email: rflynn@ait.ie)

S. Murray is with the Health Service Executive, Primary Care Centre, Longford, Ireland. (email: siobhan.murray1@hse.ie)

N. Murray is with the Athlone Institute of Technology, Athlone, Ireland. (email: nmurray@research.ait.ie)

## I. Introduction

Advances in technology have resulted in multimedia experiences that are increasingly more interactive and immersive. Although traditionally targeted towards the entertainment industry, recent works highlight the utility of head mounted display (HMD) technologies in clinical [1], training [2] and educational [3] environments. Applications within these domains attempt to mimic or enhance traditional learning outcomes by replicating training methodologies on a virtual platform. The move to a virtual platform presents numerous challenges, which include the limitations associated with current generation HMDs and the traversal of the digital/virtual divide in an ecologically valid manner.

Key to the success of multimedia applications, systems or services is the identification and understanding of the human, system and context factors [4] that influence the perceived quality of experience (QoE) [5]. QoE has emerged as an important research field, however, it is a complex paradigm to measure and quantify. It has been applied across a wide array of multimedia domains, examples of which include multimedia delivery [6] [7], panoramic multimedia systems [8] [9] and multi-sensory media experiences [10] [11]. Traditionally, post-experience surveys or questionnaires provided insight into a multimedia experience through the aggregation of mean opinion scores (MOS) [12]. However, these approaches have been shown to be limited [13] [14]. Recent research has attempted to capture and understand QoE through the monitoring of physiological and psychophysiological responses [15] [16]. These approaches aim to create a synergy between explicit and implicit measures, which produces a qualitative result supporting the experimental capture of quantitative data.

The work presented in this paper reports the results of a novel QoE comparison for three immersive technologies, namely, augmented reality (AR), tablet, and virtual reality (VR). A multimedia speech and language assessment application, which evaluates a user's ability to understand sematic links, was developed for each of the platforms. The comparison between AR, tablet, and VR is based on explicit

metrics (post-experience questionnaire) and implicit metrics (physiological metrics of heart rate and electrodermal activity (EDA)). The results indicate a higher QoE for the AR and tablet groups, based on the comparative implicit and explicit analysis.

## II. Related Work

### A. Evaluating Quality of Experience

Broadly speaking, there have been two fundamental approaches to capturing and understanding of user QoE of multimedia. These have involved the use of explicit and implicit measurements. Explicit measurements, through questionnaires or surveys, are presented and analyzed post-experience. However, there are many issues with these explicit measurements. These include the fact that they are based on post-experience only reporting; there is a loss of detail obtained by using the MOS [13] [17]; there is the question of how data captured using such methodologies can be analyzed [18].

Alternative approaches exist that aim to alleviate some of these concerns. In [19], using a multimodal feedback method, users were able to continuously reflect on their perception of a 3D video experience. At key moments during the experience user perception was captured using a tablet interface as opposed to reflecting on the overall (post) experience by completing a questionnaire. The use of implicit metrics (e.g. physiological) has in recent times generated significant interest as a mechanism to provide insight into a user's QoE [15] [20]. The human body consists of a complex network of systems that communicate using electrical signals. Research is trying to identify and formulate links between the human experience and these signals [21] [22] [16]. The autonomic nervous system (ANS) is widely accepted as the core mechanism that controls the regulatory functionality within the body. The ANS is responsible for the fight-or-flight response, which is often experienced due to a change in psychological or physiological states. Research has indicated that the monitoring of these signals provides insight into user emotion, thus providing the opportunity to develop a better understanding of a user's perceived QoE [15].

It is important to recognise that these studies do not aim to completely replace the subjective capture of QoE. Instead they propose to move towards a concurrent deployment of each capture method within QoE assessments.

In [23], the authors surveyed QoE methodologies for AR visualisation. They proposed a methodology to evaluate the QoE of AR in neuro-navigation applications. A pilot study explored the capture of objective metrics such as task completion, interaction accuracy and error rates. In terms of user-perceived QoE, human and system factors, such as the interaction metrics, were reported as key to the success of an immersive multimedia experience.

### B. Immersive Multimedia

The evaluation of user QoE when comparing immersive VR and non-VR environments was described in [21]. 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, using a post-test questionnaire, and implicit measures, in the form of heart rate and EDA. The results of the study reveal a slight elevation in heart rate in participants partaking in the immersive VR experience. Research into the use of immersive multimedia technologies as a learning mechanism within education [3] and industry [2] has received significant attention in recent times. Developed to enrich the learning of chemistry within a classroom setting, [3] presents the findings of a field study that 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.

Since the vehicle for the comparison of the three immersive platforms in this work is a speech and language assessment, it is of interest to consider existing works related to this domain. The application of HMDs as an assistive technology for those who suffer from aphasia was explored in [1]. The authors carried out two studies using Google Glass. The first study, an interview-based study with 8 participants who had aphasia, investigated the benefits and challenges associated with the use of such a technology. 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. A follow on study developed and evaluated a prototype application, which required participants to complete specific conversational tasks. The results indicated that participants were able to maintain focus on the conversation without the need for external tools. However, in some scenarios, multitasking between system interaction and conversation proved challenging.

The novelty of the work presented in this paper is in the exploration and comparison, based on implicit and explicit metrics, of user QoE of three immersive multimedia delivery platforms.

## III. Experimental Design

In this section, an overview of the immersive multimedia technologies and virtual speech and language assessment is presented. The methodologies used to capture explicit and implicit measures of QoE are described. Lastly, a detailed overview of the QoE assessment protocol is presented, highlighting the sample population and experimental approach.

### A. Immersive Multimedia Systems

Both AR and VR were selected as each of these technologies represent the current state-of-the-art in terms of immersive and interactive multimedia experiences. In addition, a tablet experience was evaluated as the hand-held device format has become a widely adopted platform for multimedia consumption.

#### 1) Augmented Reality

The AR experience was presented to participants using the Microsoft HoloLens. The wireless self-contained unit provides

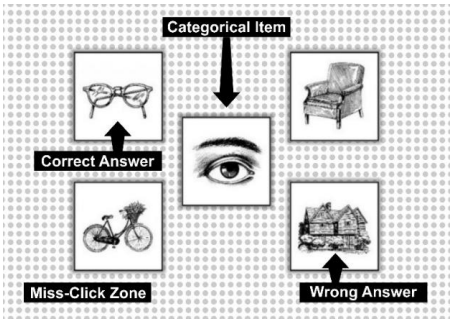



*Figure 1: Virtual semantic memory assessment*

users with an interactive and immersive holographic experience. It allows users to explore and interact with virtual content, whilst still being aware of their real-world surroundings. The device supports content interaction through a combination of head movement and hand gestures. Speakers, located above each ear, provide the wearer with a spatial audio experience.

*2) Virtual Reality*

The VR group experienced the virtual speech language therapy assessment using an Oculus Rift Development Kit 2 (DK2). A PC containing an Intel i7 6600K, 16GB of ram, and an NVIDIA 1080 GTX graphics card rendered the virtual content. The configuration of components ensured a stable frame rate of 60 frames per second. A limitation of current generation VR devices is a lack of natural human-computer interaction mechanisms. While controller-based systems do exist, there is no HMD on the market that has native support for gesture-based interaction. To facilitate gesture-based interaction, a Leap Motion controller was included in the system design. The VR environment was programmed to recognize hand gestures in a similar manner to the AR experience and thus facilitated a like-for-like experience.

*3) Tablet*

A Lenovo Tab 2 A10-70 was used to present the content for the tablet group. The device consisted of a full high-definition (HD) display with a resolution of 1200 x 1920 pixels. The device utilizes a capacitive touch screen to capture user input as they interact with the interface. Dual speakers located at the rear of the system provide audio feedback throughout the assessment. Tablet PCs are traditionally hand-held devices. In this experiment a smart cover was fitted to the device, allowing the device to stand freely on a table.

### *B. Virtual Speech and Language Assessment*

The virtual speech and language assessment employed in this work is based on the semantic memory assessment contained within the comprehensive aphasia test (CAT) [24]. The CAT is a battery of tests used to diagnose aphasia. The semantic memory assessment aims to identify the level of ease with which associative knowledge is retrieved. Poor performance on this type of test is indicative of an impairment to semantic knowledge about objects. The virtual speech language therapy assessment was developed in the Unity3D game engine.

*Table 1: Post-test questionnaire content*

| | | Questions |
|---|---|---|
| Discomfort | **Q1** | I did not feel any discomfort while using the system |
| | **Q4** | The device was annoying |
| | **Q8** | I was restricted in my movements using the system |
| | **Q9** | The system made me feel nauseous |
| Enjoyment | **Q2** | I enjoyed the experience |
| | **Q12** | I would like to experience this environment again |
| | **Q14** | My experience did not meet my expectations |
| Interaction | **Q3** | My interaction with the environment was natural |
| | **Q7** | The learning curve was not too great |
| | **Q10** | The system was easy to use |
| Immersion | **Q5** | I was immersed in the activity |
| | **Q6** | I was engaged with the system while using it |
| | **Q11** | The environment I was interacting with was real |
| | **Q13** | I did not feel a strong sense of presence whilst experiencing the system |

*Table 2: 5-point likert scale*

| 1 | 2 | 3 | 4 | 5 |
|---|---|---|---|---|
| STRONGLY DISAGREE | DISAGREE | NEITHER | AGREE | STRONGLY AGREE |

Assessment stimuli, as illustrated in Figure 1, consisted of one centralized image that was surrounded by four outer images. The four outer images contained a semantic target, an unrelated semantic distracter, a distant semantic distracter and a close semantic distracter. The objective of the assessment was to formulate a link between the central image and the semantic target. The test was made up of ten evaluation slides, along with one practice slide used at the beginning (eleven in total). In order for users to select their answer, a non-verbal response was used by participants (a hand gesture). To provide feedback to users during the assessment, audio was provided to indicate a correct or incorrect selection (akin to real life settings). An example of the type of audio response given is "That's correct, it's the glasses, because glasses help us to see" (as in the case of Figure 1) [24].

### *C. Quality of Experience Measures*

*1) Explicit Measures*

Inspired by the QoE model categories outlined in [5], a post-test questionnaire containing fourteen questions captured user response to aspects such as enjoyment, immersion, interaction and discomfort. The questionnaire [25] content is presented in Table 1. User ratings were captured using the absolute category rating (ACR) system as outlined in the ITU-T P.910 guidelines [26]. Questions were answered using a 5-point Likert scale that ranged from strongly disagree to strongly agree (Table 2).

Device discomfort was evaluated through questions 1, 4, 8 and 9. Questions 2, 12 and 14 gauged levels of satisfaction, expectations and interest in re-experiencing a system. As a collective, these three questions aimed to provide insight into enjoyment. User interaction was evaluated in questions 3, 7 and 10. User immersion was captured by questions 5, 6, 11 and 13.


*Figure 2: Experiment quality assessment protocol*

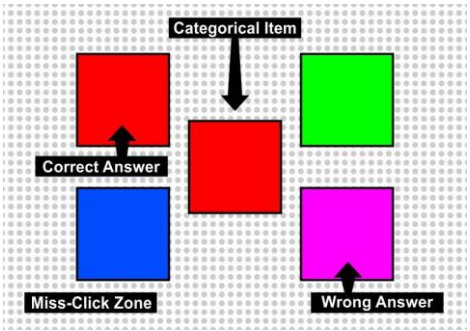



*Figure 3: Example of virtual training exercise*

*2) Implicit Measures*

In this research, heart rate (HR) and EDA were monitored continuously during phases 2, 3 and 4 of the QoE assessment protocol (see Section III.D). HR was monitored using a Fitbit Charge HR device, on which an onboard sensor measures blood volume pulse. Internal Fitbit algorithms interpret this data and present a calculated measure of beats per minute (BPM) at a rate of 1Hz.

Both physiological and psychological changes, in terms of emotional arousal, can be observed through the monitoring of the electrical characteristics of the skin. The standardized terminology for such responses is EDA [27] [28]. The Pip biosensor [29] was used to monitor changes in EDA. This is a non-invasive device that is held between the thumb and index finger. It samples EDA at a rate of 8Hz. EDA signals can be classified into two distinct categories, namely, tonic change and phasic change. Slow or steady changes in skin conductance are categorized as tonic changes. This type of variation is known as a skin conductivity level (SCL). Existing studies indicate that these slow changes are indicative of an increase in cognitive activity [30] [31]. Phasic events can be identified by a short peak in the signal, which is accompanied by varied rates of decline. These events are often referred to as the skin conductivity response (SCR). Changes in environmental stimuli, such as sound, smell or sight, often induce an SCR response. As a result, studies that consider EDA must gather additional metrics to facilitate an informed analysis of the physiological measure.

*D. QoE Assessment Protocol*

The QoE assessment protocol consisted of four phases as shown in Figure 2. These phases were designed to facilitate the capture of baseline metrics, address influences associated with the novelty of technology through training and perform the actual QoE evaluation. On average, the QoE assessment protocol took between 30-35 minutes to complete.

*1) Information and Screening Phase*

The initial phase of the experiment was divided into two key stages, information and screening. First, participants were greeted and brought to the test room. The test environment was inspired by the guidelines outlined in ISO 8589:2007 [32], only the participant and principal investigator were present during the experiment.

During the information stage participants were given an information document, with an opportunity to ask further questions. Following this, participants were required to provide written consent.

Since these immersive experiences are highly visual, as part of the screening process participants were assessed for any visual defects that could impact the results. An Ishihara test [33] was used to identify deficiencies in color perception, in which the scoring mechanism allows for a maximum of four errors. Visual acuity was evaluated using a Snellen test [34]. A score of 20/20 was required to pass the test. Five participants were deemed ineligible due to their scores during the screening phase.

*2) Resting Phase*

The anticipated use of a novel technology and experimental settings can often induce an unnatural degree of excitement. To offset any potential impact on physiological signal measurements, participants were asked to complete a resting phase. Each person has an independent self-regulating HR and EDA. To provide a fair comparison between subjects, a baseline measure of HR and EDA was captured during the resting phase. These measures were recorded over a 5-minute period. Throughout this process, participants were seated, asked to relax and to keep movements to a minimum.

*3) Training Phase*

To introduce the technology and provide an understanding of the experimental process, a series of training videos was created. Key elements such as assessment progression and system interaction were highlighted. To ensure that an understanding of the interaction mechanisms was acquired, a training phase was also developed. In this phase the AR or VR headset was fitted to the participant. For the tablet group, the device was placed in a pre-set position in front of each user.

The training phase then replicated the immersive experiences in the actual test in terms of user interaction and assessment progression. The training application required participants to formulate a link between objects. A series of colored blocks was used (Figure 3), rather than images as in the real test. A total of 11 matches was required to pass the training task. As part of the protocol, an opportunity was provided to repeat the training application if there was more than 6 errors. However, this error level was not obtained by any participant. To normalize the testing protocol, the participants' HR and EDA were captured during this phase.

*4) Testing Phase*

The fourth and final phase of the experiment consisted of two key steps. First, participants were required to interact with the immersive speech and language therapy assessment. A total of eleven slides (adhering to the CAT specification) were

*Table 3: Participant demographic*

| | | Age | | Gender (%) | |
|---|---|---|---|---|---|
| | Total | Avg. (μ) | SD (σ) | 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% |

evaluated (one practice, and ten test). HR and EDA were recorded throughout the assessment. Upon completion, the investigator removed the multimedia device (HMD/tablet), Fitbit and Pip Biosensor from the user. Participants were then asked to complete a post-test questionnaire [25], as described in Section III.C.

### *E. Participants*

A total of 67 participants was recruited using convenience sampling. Of these, 7 participants were omitted due to a combination of screening and technical errors noted during testing. Therefore, the findings on a total of 60 subjects are presented. To avoid the influence of pre-exposure to assessment stimuli, a between subject design was selected for this experiment. As such the participants were split evenly between the three test groups (AR, VR, and tablet). Information on participant age and gender can be found in Table 3.

None of the AR group had experienced an AR HMD before. However, it was noted that seven had experienced mobile AR in the form of applications such as Pokémon Go. Six of the VR group had previous experience with VR HMD systems but, again, these were on mobile platforms (i.e. non-interactive immersive experiences). Lastly, all participants within the tablet group had a high level of familiarization with the functionality of touch-screen devices.

## IV. Results and Discussion

In this section, we present the explicit and implicit data captured throughout the experiments. Statistical analysis was performed using IBM SPSS [35]. A multivariate analysis of variance (MANOVA) was performed with a 95% confidence level.

### *A. Explicit Data*

The MOS results of the post-test questionnaire are presented in Table 4. Measuring levels of discomfort, the MOS ratings of questions 1, 4 and 8 are favorable towards the tablet group. Tablet users reported that they experienced a greater level of comfort, found the device least annoying and were less restricted in their movements. This was somewhat expected as tablet users did not wear a display device on their head.

Levels of immersion were evaluated through questions 5, 6, 11 and 13. The results reveal that a marginally higher level of immersion was experienced within the VR and AR groups. This finding is expected as the presentation of virtual content directly within a user's field of view, as opposed to on a screen, is a fundamental component of immersing a user within a multimedia experience.

The results for enjoyment, which was evaluated through questions 2, 12 and 14, were mixed. 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 reported 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 speech and language therapy application was enjoyable for all groups.

Questions 3, 7 and 10 provided an overview of user interaction. The results reveal that two out of the three questions were favorable towards the tablet group. All participants had a familiarity with touchscreen tablet technology, because of this the reported MOS is somewhat expected. However, of most interest, are the similarities within the ratings. This highlights the adaptability of new and immersive experiences.

In the following sections we discuss and highlight the results of the MANOVA (95% confidence interval) for the self-reported measures (see Table 5), comparing each respective group in the form of AR vs tablet, AR vs VR, and

*Table 4: MOS ratings captured via post-test questionnaire*

| | | AR | | Tablet | | VR | |
|---|---|---|---|---|---|---|---|
| | | MOS | SD (σ) | MOS | SD (σ) | MOS | SD (σ) |
| **Discomfort** | **Q1** | 3.900 | 1.021 | 4.450 | 1.099 | 3.550 | 1.468 |
| | **Q4** | 2.500 | 1.000 | 1.700 | 0.923 | 2.000 | 0.858 |
| | **Q8** | 2.550 | 0.887 | 1.900 | 1.119 | 2.400 | 0.995 |
| | **Q9** | 1.150 | 0.366 | 1.250 | 0.550 | 1.300 | 0.470 |
| **Enjoyment** | **Q2** | 4.550 | 0.510 | 4.650 | 0.587 | 4.600 | 0.503 |
| | **Q12** | 4.500 | 0.513 | 3.900 | 0.788 | 4.050 | 0.605 |
| | **Q14** | 2.300 | 0.865 | 1.900 | 1.021 | 1.850 | 0.813 |
| **Interaction** | **Q3** | 3.650 | 0.988 | 4.450 | 0.686 | 4.050 | 0.945 |
| | **Q7** | 3.350 | 1.663 | 3.050 | 1.761 | 3.350 | 1.599 |
| | **Q10** | 4.600 | 0.598 | 4.950 | 0.224 | 4.500 | 0.607 |
| **Immersion** | **Q5** | 4.350 | 0.587 | 4.150 | 0.813 | 4.350 | 0.489 |
| | **Q6** | 4.700 | 0.470 | 4.500 | 0.513 | 4.600 | 0.503 |
| | **Q11** | 3.450 | 1.146 | 3.900 | 1.119 | 3.900 | 0.641 |
| | **Q13** | 2.650 | 0.875 | 3.050 | 1.317 | 3.300 | 1.174 |

*Table 5: MANOVA of post-test questionnaire at 95% confidence level*

| | AR vs Tablet | | Tablet vs VR | | VR vs AR | |
|---|---|---|---|---|---|---|
| | F | Sig | F | Sig | F | Sig |
| **Q1** | 0.267 | 0.608 | 2.345 | 0.134 | 1.268 | 0.268 |
| **Q2** | 0.012 | 0.914 | 0.004 | 0.95 | 0.036 | 0.851 |
| **Q3** | 3.519 | 0.069 | 1.414 | 0.242 | 0.408 | 0.527 |
| **Q4** | 5.77 | 0.022 | 0.693 | 0.411 | 2.945 | 0.095 |
| **Q5** | 1.017 | 0.32 | 2.094 | 0.157 | 0.152 | 0.699 |
| **Q6** | 2.532 | 0.12 | 1.527 | 0.225 | 0.15 | 0.701 |
| **Q7** | 0.055 | 0.816 | 0.141 | 0.71 | 0.017 | 0.896 |
| **Q8** | 2.741 | 0.106 | 2.154 | 0.151 | 0.024 | 0.878 |
| **Q9** | 0.06 | 0.808 | 0.039 | 0.844 | 0.262 | 0.612 |
| **Q10** | 3.386 | 0.074 | 10.304 | 0.003 | 0.985 | 0.328 |
| **Q11** | 0.003 | 0.96 | 0.091 | 0.765 | 0.14 | 0.711 |
| **Q12** | 10.38 | 0.003 | 2.159 | 0.15 | 4.126 | 0.05 |
| **Q13** | 0.118 | 0.733 | 2.697 | 0.109 | 5.114 | 0.03 |
| **Q14** | 2.131 | 0.153 | 0.012 | 0.913 | 3.102 | 0.087 |

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TMM.2020.2982046, IEEE Transactions on Multimedia



tablet vs VR. A Levene's test for equality of variances has been carried out on each of the respective results, all data reported meets the homogeneity requirements for a MANOVA.

*1) Augmented Reality vs Tablet*

An analysis of the MOS captured between the AR and tablet groups reveals two statistically significant results. 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 compared to the AR group who reported a MOS of 2.5. This result is statistically significant ($p$=0.022). This finding suggests that the users' preference was not to wear a HMD. 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 reported a higher MOS of 4.5 compared to the tablet group rating of 3.9. For the novel delivery of an AR experience, in terms of real-world projected content compared to a traditional screen, the users reported they would happy to experience it again.

*2) Tablet vs Virtual Reality*

The comparison of the tablet and VR groups reveals one statistically significant finding. In terms of interaction as a factor influencing user QoE, question 10, which was found to be statistically significant ($p$=0.003), asked users to rate how easy the system was to use. The tablet group provided a higher overall MOS of 4.95 compared to 4.5 for VR group.

*3) Virtual Reality vs Augmented Reality*

Comparison of the AR and VR group revealed two statistically significant findings. Question 12 asked users if they would like to experience the environment again. The statistical analysis revealed a borderline result of $p$=0.05. This is reflected in the MOS of 4.5 for the AR group compared to the MOS of 4.05 for the VR group. Question 13 aimed to evaluate the sense of presence felt by users whilst experiencing the system (the question is negatively phrased, as a result a lower MOS is better). Prior to the experiment it was hypothesized that the VR group would experience a greater sense of presence in the virtual environment. However, the results reveal that the AR group felt a greater presence within the virtual assessment with a MOS rating of 2.65 compared to 3.3 for the VR group. This difference is statistically significant ($p$=0.30). The standard deviation (σ) of 1.21 in the VR group reveals a level of disagreement between participants.

*B. Implicit Data*

*1) Electrodermal Activity*

Figure 4 presents the average electrodermal response as users interacted with the multimedia stimulus. To illustrate the trend between the groups, the values have been scaled using a min-max normalization approach.

In terms of EDA for both AR and tablet, there is a downward sloping trend as participants progressed throughout the assessment. This finding is interesting, as it suggests that their EDA is returning to the baseline levels captured in the resting phase. As such, there is a potential correlation associated with both of these groups experiencing the virtual assessment within the context of the real-world.

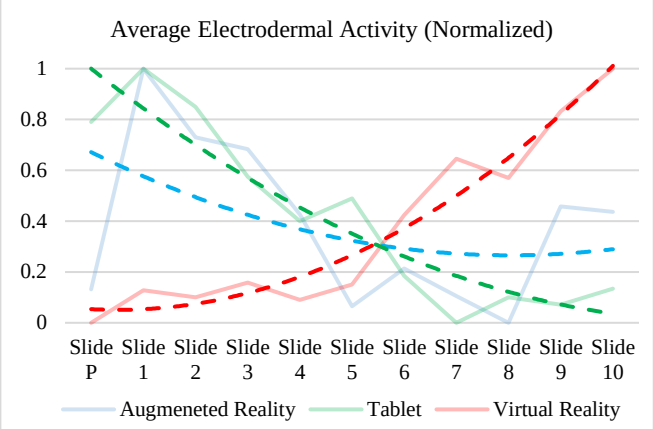



*Figure 4: Average increase in EDA when compared to the baselines.*

An opposite trend is observed in the VR group, in which participants report an average increase as they progress from slide P to slide 10. As reported in [16], this increase coincides with the number of errors (interaction and incorrect responses) experienced by the VR group. Previous research has indicated that a slow and steady increase in EDA (SCR) can be correlated with an increase in cognitive activity [30] [31] [36]. As outlined in Section III.B, semantic assessments aim to challenge the brain's 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 is believed that the change noted here highlights a difficulty in semantic categorization experienced by the VR group.

Table 6 presents an overview of the average measure of EDA at each slide throughout the experiment. The baseline value represents the average measure of EDA captured during the resting phase (Section III.D). Slides P through 10 provide insight into the average EDA as participants progressed through the assessment content. The test average (Test Avg.) in Table 6 represents the average EDA throughout the entire testing phase. Comparing the baseline value and the test average reveals that an increase was experienced across all groups.

Overall, the AR group experienced an increase of 0.474 (15%) microsiemens compared to the baseline. Similarly, the EDA from the VR group increased by 0.448 (15%) microsiemens. The similarities in these two groups is interesting, as both AR and VR participants experienced the delivery of multimedia content from a HMD. However, unlike the AR and VR groups, the tablet group reported a much larger increase of 1.459 (49%) microsiemens from the baseline. As a result of this, an investigation was carried out to explore the possibility of third-party influences. Factors considered were the influence of environmental factors, such as weather, and the impact of capacitive touch screen technology. Exploring each of these avenues validated the results i.e. the measure of EDA within the tablet group was not influenced by external factors and hence was comparable with the AR and VR groups.

*Table 6: MANOVA of electrodermal activity at a 95% confidence level*

| | Between Group | | | | | | | | Within Subject Between Group | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | AR | | Tablet | | VR | | | | AR | Tablet | VR | | |
| | Avg. (μ) | SD (σ) | Avg. (μ) | SD (σ) | Avg. (μ) | SD (σ) | F | Sig. | SD (σ) | SD (σ) | SD (σ) | F | Sig. |
| **Baseline** | 3.142 | 1.294 | 2.997 | 1.128 | 2.891 | 1.214 | 0.639 | 0.532 | 0.343 | 0.280 | 0.323 | 0.101 | 0.904 |
| **Slide P** | 3.578 | 1.055 | 4.539 | 1.448 | 3.278 | 1.148 | 6.258 | 0.004 | 0.089 | 0.119 | 0.067 | 0.627 | 0.538 |
| **Slide 1** | 3.691 | 1.095 | 4.579 | 1.432 | 3.297 | 1.105 | 6.750 | 0.002 | 0.064 | 0.076 | 0.045 | 1.587 | 0.214 |
| **Slide 2** | 3.656 | 1.143 | 4.551 | 1.418 | 3.293 | 1.115 | 6.522 | 0.003 | 0.066 | 0.084 | 0.042 | 2.621 | 0.082 |
| **Slide 3** | 3.650 | 1.152 | 4.498 | 1.434 | 3.301 | 1.120 | 5.929 | 0.005 | 0.084 | 0.054 | 0.050 | 0.726 | 0.488 |
| **Slide 4** | 3.616 | 1.181 | 4.465 | 1.444 | 3.291 | 1.134 | 5.796 | 0.005 | 0.067 | 0.057 | 0.041 | 0.424 | 0.657 |
| **Slide 5** | 3.570 | 1.192 | 4.482 | 1.481 | 3.300 | 1.124 | 5.754 | 0.005 | 0.066 | 0.077 | 0.049 | 0.454 | 0.637 |
| **Slide 6** | 3.589 | 1.202 | 4.424 | 1.425 | 3.339 | 1.133 | 4.850 | 0.012 | 0.050 | 0.089 | 0.064 | 1.225 | 0.302 |
| **Slide 7** | 3.575 | 1.241 | 4.389 | 1.405 | 3.371 | 1.165 | 4.283 | 0.019 | 0.049 | 0.075 | 0.055 | 1.139 | 0.328 |
| **Slide 8** | 3.561 | 1.246 | 4.408 | 1.437 | 3.360 | 1.156 | 4.713 | 0.013 | 0.066 | 0.084 | 0.055 | 0.731 | 0.486 |
| **Slide 9** | 3.621 | 1.186 | 4.402 | 1.432 | 3.398 | 1.183 | 4.499 | 0.016 | 0.064 | 0.086 | 0.078 | 0.667 | 0.517 |
| **Slide 10** | 3.618 | 1.177 | 4.414 | 1.449 | 3.422 | 1.173 | 4.374 | 0.017 | 0.057 | 0.081 | 0.053 | 1.490 | 0.235 |
| **Test Avg.** | 3.616 | 1.160 | 4.456 | 1.424 | 3.339 | 1.146 | 5.453 | 0.007 | 0.128 | 0.160 | 0.110 | 1.118 | 0.334 |

The results of a statistical analysis for EDA are also presented in Table 6. All of the results are statistically significant. 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 were significant. The key differences in display technology, in terms of the presentation on a HMD versus a table-top display, would have been a key factor in these high rates of significance.

Lastly, a comparison of EDA between the AR and VR groups reveals no statistically significant differences. This finding is interesting as it reveals that, in terms of emotional arousal, both the AR and VR groups had a similar experience throughout the evaluation. The wearing of a HMD is more than likely a key influential factor within these results.

To further validate these findings, Table 6 also presents a within-subject between-group analysis of EDA. The objective of this is to provide insight into the average standard deviation of each respective group's physiological response on a per subject basis. Presentation of these results further validate the previous findings by highlighting that there was no significant difference with respect to the standard deviation of EDA between the subjects. Therefore, supporting the validity of the assessment strategy.

*2) Heart Rate*

An analysis of heart rate (measured in BPM) captured during the testing phase revealed no statistically significant results. In Figure 5, a visual overview of this data is presented. Although diverse, each of the three groups follows a similar trend. Initial signals appear slightly offset from the average baseline, which is followed by an increase that is shared across groups. Despite this trend, the percentage change that occurs over the duration of the test is varied. A minor increase is noted in the tablet group as the mean HR increases by 1.8% over the test duration. The VR group experienced a slightly larger increase of 3.33%. Lastly, the AR group experienced the highest increase of 5.7%.

Interestingly, both the AR and VR group heart rates begin below their baseline thresholds. To a certain degree this was expected because the use of a new technology can often create high levels of anticipation. This increase has the potential to raise the user's heart rate above the normal. In addition, the impact of positive reinforcement must be considered. As participants progressed through the activity, a virtual speech and language therapist responded to each correct answer with a positive comment. These comments from the virtual therapist have the potential to invoke a level of enjoyment or satisfaction at performing well in the assessment. It is important to remember that because of the interactive element of the assessment, subtle movement has the potential to cause a deviation in heart rate as the blood circulates around the body. This could explain the lack of statistical significance between the groups, as all groups would have experienced similar interactive measures.

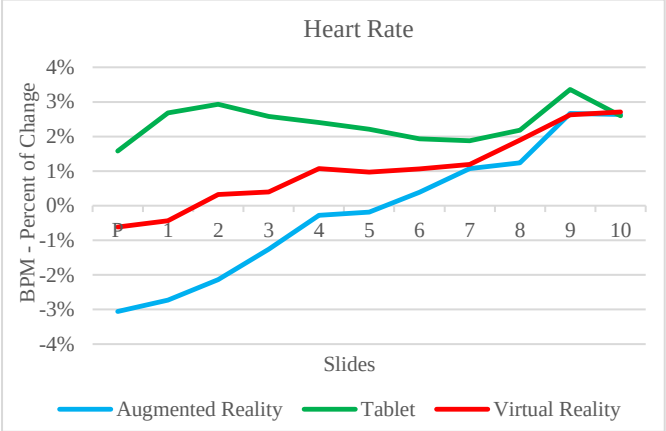



*Figure 5: A comparison of the* average increase *in heart rate, beats per minute (BPM), a value of 0% represents the group baseline.*

## V. Conclusion

In this article, the utility of three immersive multimedia experiences was evaluated and compared based on QoE assessment approaches. The QoE evaluation was performed to solicit explicit and implicit responses from the users. User ratings were collected using a post-test questionnaire. Implicit measurements in the form of heart rate and electrodermal activity were observed as an insightful measure of physiological and, what is perceived to be, psychological arousal.

In terms of physiological response, similarities were found between the AR and tablet groups. This finding is interesting

as both groups experience multimedia content whilst remaining in the context of a real-world setting. Analysis of the VR group reveals a unique response, which is a gradual elevation in EDA as participants progress through the virtual speech language therapy assessment. Existing research suggests that this slow and steady increase can be correlated with a change in cognitive activity. Typically, influencing factors associated with positive trends in EDA can be attributed to an increase in mental workload or stress.

Overall, the results presented for both explicit and implicit measurements reveal that the AR and tablet groups experience a higher QoE. Although the implicit capture of EDA reveals an increase in physiological arousal for the VR group, this may be attributed to limitations associated with the technology. For instance, unlike the VR hardware, both the AR and tablet environments support native gesture controls. In addition, the display technology differs between each of the immersive multimedia devices. As such, it is believed that the future generation of VR HMDs may alleviate some of these shortfalls, therefore bringing the overall QoE of the VR group in line with the AR and tablet groups.

The results reveal the suitability of all platforms as potential methods for speech and language assessment. Future work will further explore the utility of EDA as a measure of cognitive load or stress within the context of a virtual assessment. In particular, the work will integrate and evaluate the utility of already established measures of cognitive activity, such as pupillary response. The objective of this is to further validate the findings of this article.

## ACKNOWLEDGMENT

This research was supported by the Irish Research Council (grant number: GOIPG/2018/1314).

## REFERENCES

[1] 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.

[2] 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.

[3] 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.

[4] Qualinet, "Qualinet White Paper on Definitions of Quality of Experience," White Paper, Novi Sad, 2013.

[5] A. Raake and S. Egger, "Quality and Quality of Experience," in *Quality of Experience: Advanced Concepts, Applications and Methods*, Springer, 2014, pp. 11-33.

[6] M. Schmitt, J. Redi, D. Bulterman and P. S. Cesar, "Towards Individual QoE for Multiparty Videoconferencing," *IEEE Transactions on Multimedia*, vol. 20, no. 7, pp. 1781 - 1795, 2017.

[7] N. Murray, G.-M. Muntean , Y. Qiao , S. Brennan and B. Lee, "Modeling User Quality of Experience of Olfaction-Enhanced Multimedia," *IEEE Transactions on Broadcasting*, vol. 64, no. 2, pp. 539-551, 2018.

[8] V. R. Gaddam, M. Riegler, R. Eg, C. Griwodz and P. Halvorsen , "Tiling in Interactive Panoramic Video: Approaches and Evaluation," *IEEE Transactions on Multimedia*, vol. 18, no. 9, pp. 1819 - 1831, 2016.

[9] J. Cubelos, P. Carballeira , J. Gutiérrez and N. García , "QoE Analysis of Dense Multiview Video With Head-Mounted Devices," *IEEE Transactions on Multimedia*, vol. 22, no. 1, pp. 69-81, 2020.

[10] 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.

[11] N. Murray, Y. Qiao, C. Keighrey, D. Egan, D. Pereira Salgado, G. Miro Muntean, C. Timmerer, O. A. Ademoye, G. Ghinea and B. Lee, "Evaluating QoE of Immersive Multisensory Experiences," *IEEE MMTC Communications Frontiers*, vol. 13, no. 1, pp. 6-13, 2018.

[12] ITU-T, "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," [Online]. Available: https://www.itu.int/rec/T-REC-P.913/en. [Accessed 14 03 2019].

[13] T. Hoßfeld, R. Schatz and S. Egger, "SOS: The MOS is Not Enough!," in *3rd International Workshop on Quality of Multimedia Experience*, Mechelen, Belgium, 2011.

[14] T. Hoßfeld, P. E. Heegaard, M. Varela and S. Möller, "QoE beyond the MOS: an in-depth look at QoE via better metrics and their relation to MOS," *Quality and User Experience*, vol. 1, no. 1, 2016.

[15] U. Engelke, D. P. Darcy, G. H. Mulliken, S. Bosse, M. G. Martini, S. Arndt, J.-N. Antons, K. Y. Chan, N. Ramzan and K. Brunnstrom, "Psychophysiology-Based QoE Assessment: A Survey," *IEEE Journal of Selected Topics in Signal Processing*, vol. 11, no. 1, pp. 6 - 21, 2016.

[16] 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.

[17] L. Janowski and M. Pinson, "The Accuracy of Subjects in a Quality Experiment: A Theoretical Subject Model," *IEEE Transactions on Multimedia* , vol. 17, no. 12, pp. 2210 - 2224, 2015.

[18] M. Narwaria , L. Krasula and P. Le Callet, "Data Analysis in Multimedia Quality Assessment: Revisiting the Statistical Tests," *IEEE Transactions on Multimedia*, vol. 20, no. 8, pp. 2063 - 2072, 2018.

[19] T. Kim , J. Kang, S. Lee and A. C. Bovik, "Multimodal Interactive Continuous Scoring of Subjective 3D Video Quality of Experience," *IEEE Transactions on Multimedia*, vol. 16, no. 2, pp. 387-402, 2014.

[20] C. Keighrey, R. Flynn, S. Murray and N. Murray, "A QoE Evaluation of Immersive Augmented and Virtual Reality Speech & Language Assessment Applications," in *QoMEX 2017 – 9th International Conference on Quality of Multimedia*, Erfurt, Germany, 2017.

[21] 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.

[22] 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 *5th ACM SIGGRAPH Symposium on Video Games* , Los Angeles, California, 2010.

[23] 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.

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

[25] C. Keighrey, "Post-Test Questionnaire," [Online]. Available: http://www.conorkeighrey.info/. [Accessed 20 02 2020].

[26] 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," 01 03 2016. [Online]. Available: https://www.itu.int/rec/T-REC-P.913/en. [Accessed 14 03 2019].

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

[28] C. C. Brown, "A Proposed Standard Nomenclature for Psychophysiologic Measures," *Psychophysiology*, vol. 2, no. 2, pp. 260-264, 1967.

[29] PIP, "Home |Pip | Stress management device," PIP, [Online]. Available:

https://thepip.com/en-eu/. [Accessed 17 07 2018].

[30] 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.

[31] 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.

[32] "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 04 2017].

[33] 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.

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

[35] "IBM SPSS - IBM Analytics," IBM, [Online]. Available: https://www.ibm.com/analytics/us/en/technology/spss/. [Accessed 14 03 2019].

[36] 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.

**Conor Keighrey** is a PhD research candidate in Athlone Institute of Technology (AIT), Ireland. He received his BSc. in Computer Network Management and Cloud Infrastructure in 2016 and is currently in pursuit of his PhD. His current research work focuses on understanding the key influencing factors that affect quality of experience of emerging immersive multimedia experiences (augmented reality and virtual reality).

**Ronan Flynn** received the BE degree in Electronic Engineering from University College Dublin, the MEng degree in Computer Systems from University of Limerick and the PhD degree from National University of Ireland, Galway. He has industrial experience in telecommunication product design and development for international markets, having previously worked with Siemens AG in Munich and Northern Telecom (Irl) Ltd. in Galway. He is currently a lecturer and researcher in the Faculty of Engineering & Informatics in Athlone Institute of Technology. His research interests include speech recognition, speech enhancement, emotion recognition in speech and multi-modal affective computing.

**Siobhan Murray** is a Senior Speech and Language Pathologist with the Health Service Executive, Ireland. She has a BSc in Speech and Language Pathology from the University of Strathclyde, Glasgow, Scotland. Her research interests are with respect to the application of technology in the diagnosis and intervention of speech and language disorders.

**Niall Murray** is a lecturer with the Faculty of Engineering and Informatics, in Athlone Institute of Technology (AIT), Ireland. He is founder (in 2014) and principal investigator in the truly Immersive and Interactive Multimedia Experiences (tIIMEx) research group in AIT. He is a Science Foundation Ireland Funded Investigator in the Adapt Centre for Digital Content Technology and the Confirm Centre for Smart Manufacturing. His current research interests include immersive and multisensory multimedia communication and applications, multimedia signal processing, quality of experience, and wearable sensor systems.