A Quality of Experience Evaluation of Text and 3D Instruction Formats in Augmented Reality Applications

Author

Eoghan Hynes

Supervised by:

Dr Niall Murray and Dr Ronan Flynn

Technological University of the Shannon
Midlands Midwest

Ollscoil Teicneolaiocha don Sionainn
Lar na Tire Lar Thiar

Submitted to TUS: Midlands in partial fulfilment of the requirements for the degree of

Doctor of Philosophy

September 2023

DECLARATION

I hereby declare that the thesis entitled “A Quality of Experience Evaluation of Text and 3D Instruction Formats in Augmented Reality Applications”, submitted for the award of the degree of Doctor of Philosophy to the Technological University of the Shannon: Midlands (TUS: Midlands), is a record of bona-fide work carried out by Eoghan Hynes under the supervision of Dr. Niall Murray and Dr. Ronan Flynn, Lecturers in the Department of Computer and Software Engineering, Faculty of Engineering and Informatics, TUS: Midlands.

I further declare that the work reported in this thesis has not been submitted, either in part or in full, for the award of any other degree or diploma in this technological university or any other institute or university.

Additionally, I provide the details of all publications resulting from this work in the publications section at the beginning of this thesis, from which the methodologies and results are duplicated verbatim in this thesis.

Date: 20/10/2023

Signature of the Candidate

EOGHAN HYNES

“Pay no attention to that man behind the curtain”.

- The Wizard of Oz

ACKNOWLEDGEMENTS

With immense pleasure and a deep sense of gratitude, I wish to express my sincere thanks to my supervisors **Dr. Niall Murray** and **Dr. Ronan Flynn** of the Faculty of Engineering and Informatics, Department of Computer and Software Engineering, TUS: Midlands. Without their motivation and continuous encouragement, this research would not have been successfully completed.

I am grateful to the President of TUS, **Prof. Vincent Cunnane** for providing me with infrastructural facilities, funding and many other resources needed for my research. I express my sincere thanks to **Dr. Enda Fallon,** the Head of Department of Computer and Software Engineering, TUS: Midlands and **Dr. Brian Lee**, Director of the Software Research Institute, TUS: Midlands for their support and encouragement.

It would be remiss of me not to acknowledge the support of **my colleagues** in several ways throughout my PhD programme. Your friendship has made my time in TUS: Midlands some of the most enjoyable and fulfilling of my life. I cherish the memories of our time together.

I wish to extend my profound sense of gratitude to **my parents** for all the sacrifices they made during my research and providing me with moral support and encouragement as required.

This thesis has emanated from research conducted with the financial support of Science Foundation Ireland (SFI) under grant number SFI/16/RC/3918, the AIT President's Seed Fund under number P221-233-PSF2017, and the Horizon Europe TRANSMIXR project (grant number: 101070109) for which I am very grateful.

Date: 20/10/2023 **EOGHAN HYNES**

ABSTRACT

Augmented reality (AR) is an emerging technology that has significant potential as a solution for novel procedure assistance and repeatable procedure training. Instructions are a method to communicate how to perform a procedure for different reasons and pedagogical goals. This can range from assistance with once-off product assembly to long term learning. The main barrier to mass adoption of optical see-through AR headsets for these roles arises when AR instruction fails to fulfil the user's pragmatic and hedonic needs and expectations due to human, system and context influencing factors. User quality of experience (QoE) considers this fulfilment to be reflected in the user's degree of delight or annoyance. The ability to directly measure emotional response using modern psychophysiological instruments is shifting the focus of quality assessment towards evaluation of fulfilment of user needs and expectations. In this context, the work presented in this thesis focuses on understanding the influence of instruction formats considering AR as a potential platform for procedure assistance and training. Instruction format was evaluated over two distinct studies specific to the procedure assistance and training roles. In Study 1, the influence of paper-based and AR-based text instruction formats on user QoE for procedure assistance was evaluated using a Rubik's Cube® proof of concept. In Study 2, a combined text and interactive animated 3D model instruction format was compared against a text-only instruction format within AR using a GoCube™ proof of concept for training. Two separate AR applications were developed. Physiological ratings, facial expressions and eye gaze metrics were recorded. Subjective experience was reported using Likert scale, self-assessment manikin and NASA task load questionnaires. Statistical analysis was employed to identify statistically significant differences between usage of the different instruction formats. Correlation and regression analysis were undertaken to identify novel implicit metrics of QoE. The results from Study 1 show that the AR instruction format yielded objective performance benefits over the paper-based instruction format for procedure assistance while participants reported higher acceptability of AR. Heart rate features indicated increasing stress in both test groups, which corelated to mental load in both groups. Study 2 results show that the text-only instruction format yielded faster instruction response times in procedure training compared to a combined text and model instruction format. Female trainees using the combined instruction format were significantly slower in training and recall than females that used the text-only instruction format but reported requiring less cognitive effort than male participants during training and recall. An absence of statistically significant correlations between physiological ratings, facial expression and emotion terms used by the participants, calls into question the utility of such emotion terms as measures of emotional state. Facial expressions of action unit 20 correlated to task duration in both studies.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ACRONYMS

2D: Two Dimensional

3D: Three Dimensional

6DoF: Six Degrees Of Freedom

AR: Augmented Reality

AU: Action Unit

AV: Augmented Virtuality

BVP: Blood Volume Pulse

CAD: Computer Assisted Design

CG: Control Group

CIF: Context Influencing Factor

CLNF: Conditional Logical Neural Fields

ECG: Electrocardiogram

EDA: Electrodermal Activity

FOV: Field Of View

FPS: Frames Per Second

HIF: Human Influencing Factor

HL2: Hololens 2

HR: Heart Rate

HRV: Heart Rate Variability

IBI: Interbeat Interval

ITU-T: International Telecommunications Union

LTM: Long-term Memory

MFE: Micro Facial Expression

MOS: Mean Opinion Score

NFE: Normal Facial Expression

PPG: Photoplethysmography

QoE: Quality of Experience

QoS: Quality Of Service

RAM: Random Access Memory

SAM: Self Assesment Manikin

SD: Standard Deviation

SDK: Software Development Kit

SDM: Semantic Differential Method

SIF: System Influencing Factor

SM: Sensory Memory

SPSS: Statistical Package for Social Sciences

TG: Test Group

VR: Virtual Reality

Wi-Fi: Wireless Fidelity

WM: Working Memory

CHAPTER 1

Introduction

1.1 Background and motivation

Instructions are commonly used to communicate information about how to perform a procedure. A procedure is a suite of steps that typically needs to be executed in a specific order for successful completion. Instructions can be used during training to teach a novice how to perform a repeatable procedure. Instructions can also be used to directly assist a person while performing a procedure that changes so frequently or is so rarely encountered that learning how to repeat it does not fit the user’s pragmatic needs. Optimal warehouse distribution and mass customisation are examples of such frequently changing procedures. Instructions can also be used to assist a person with repeatable procedures, such as to ensure that strict safety protocols are adhered to. The distinguishing feature of training is the trainee’s pragmatic need for learning, which may not necessarily be the case during assistance.

Users of procedure assistance instruction often rely on paper-based media (e.g. for furniture assembly). However, the highly variable procedures of mass customisation provide the motivation for the evaluation of more adaptive assistance technologies such as augmented reality (AR). AR is an emerging technology that has significant potential as a procedure assistance and training platform. AR technology fundamentally consists of a combination of input sensory information and output media that are mediated by software. The software combines or adds additional information to the sensory input for presentation to the user in the output media. The distinguishing feature of AR is that this is done primarily to enhance the user’s interaction with their physical environment. AR-based procedure assistance has the potential to increase worker utility in the current climate of ever-increasing levels of automation of repetitive procedures. The literature calls for the evaluation of AR applications to assist and strengthen human roles in performing frequently changing procedures [1]. This provides the motivation for a comparative evaluation of AR-based and paper-based textual procedure assistance instruction formats.

AR is also regarded as a promising training platform. Context-aware AR applications allow for interactive training that enables corrective feedback to ensure correct learning. This can reduce trainee cognitive load. However, cognitive load can also be impacted by instruction format. The different extraneous cognitive loads inherent in text (procedural) and graphical (example) training instruction formats can influence learning in different ways. It is believed that a graphical representation of information can reduce extraneous cognitive load by allowing the trainee to better conceptualise a task. It is also believed that the reduced cognitive effort required of such graphical instruction formats can in turn negatively influence learning and transfer due to the development of over-dependence. Conversely, the cognitive effort required to carry out text instructions may benefit learning and transfer. Research is required to evaluate the influence of training instruction formats on the AR trainee. This provides the motivation for an evaluation of procedural and example training instruction formats within AR.

Mass adoption of AR head mounted displays (HMDs) is dependent upon the realisation of applications of utility in the context of multiple human, system and context influencing factors. Human influencing factors specific to binocular optical see-through AR HMDs include double vision (diplopia). This results from the user trying to focus on multiple depth planes at once (e.g., a real object held at arm’s length, and its augmentation presented close to the eye). System factors include object tracking-based procedural flow-control. This system-level tracking factor can be influenced by environmental contexts (see Fig. 1.1) including lighting, reflection, target pose and target occlusion, which can fall outside of the control of the AR developer. AR HMDs are primed for adoption in distinct procedure assistance and training roles across multiple disciplines. One of the main barriers to mass adoption of AR HMDs for these roles arises where AR instruction formats fail to fulfil the user’s pragmatic (i.e. utility) and hedonic (i.e. enjoyment) needs and expectations. An understanding of procedure assistance and training instruction formats considering human, system and context influencing factors, including those mentioned here, provides the context of and motivation for this research. This is required to realise the potential of the optical see-through AR HMD as a procedure assistance and training platform.

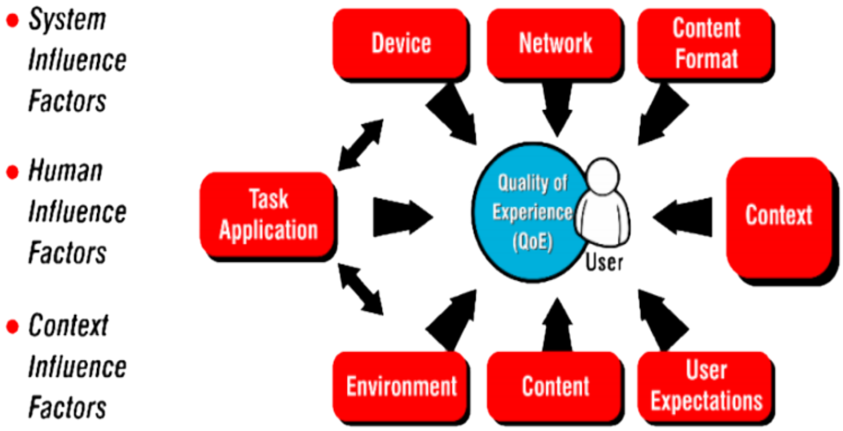


Fig. 1.1. The system, human and context factors that influence user QoE [2]**.**

The literature highlights an absence of such an evaluation of instruction formats for the current generation of state-of-the-art optical see-through AR HMDs [3]. Balanced research encompassing instruction format is needed to demonstrate the benefits of AR for procedure assistance and training roles [4]. This can be achieved by means of quality of experience (QoE) evaluations. QoE considers the degree of fulfilment of a user's pragmatic and hedonic needs and expectations that an application, system or service provides.

People use instructions for procedure assistance with the goal of efficient and accurate procedure completion. They also use instructions for procedure training with the goal of learning how to perform the procedure, and/or transfer of that knowledge to similar procedures. The user's hedonic needs and expectations for these roles are less well documented but are believed to be influenced by aesthetic, usability and interaction quality factors [5]–[7]. The degree of fulfilment of these pragmatic and hedonic needs and expectations is reflected in part in the user's degree of delight or annoyance in response to use of the instruction formats. The ability to directly measure this emotional response using modern psychophysiological instruments is shifting the focus of quality assessment towards the evaluation of fulfilment of user needs and expectations in the form of QoE evaluations.

This research consists of two distinct QoE evaluations of instruction formats for procedure assistance and training, considering AR as a potential platform for these roles:

1. Study 1 evaluated the influence of AR-based and paper-based "text" instruction formats for procedure assistance using a Rubik's Cube® solving proof of concept.

2. Study 2 was a within-AR QoE evaluation of a combined "text and interactive animated 3D model" instruction format compared to a "text-only" instruction format using a GoCube™ (an electronic version of the Rubik's Cube®) training procedure.

This was accomplished by formulating the research questions of this work from the perspective of QoE evaluations, which are detailed in the following section.

1.2 Research Questions

The overarching research questions of this work are:

1. How does text instruction in AR influence user QoE for procedure assistance compared to a paper-based control?

2. How does a combined text and interactive animated 3D model instruction format influence user QoE for procedure training compared to a text-only instruction format?

These research questions are answered by conducting two distinct studies (Study 1 and Study 2) for the procedure assistance and training use cases. The research questions are broken down into a set of five research sub-questions that are answered across Study 1 and Study 2.

RSQ1: How do the instruction formats influence the user's pragmatic needs and expectations?

RSQ2: What do users self-report in terms of the degree of fulfilment of their hedonic needs and expectations when experiencing the instruction formats?

RSQ3: Can physiological measurements and facial expressions support a better understanding of user responses in the context of a QoE evaluation of the different instruction formats?

RSQ4: What is the influence of gender on the degree of fulfilment of pragmatic needs of the user for the different instruction formats?

RSQ5: How do different cognitive loads inherent in the different instruction formats influence user QoE?

RSQ1 and RSQ4 are answered in Study 1 in Chapter 4 on page 72 and are summarised for Study 2 in Chapter 5 on page 102. The answers to RSQ2 and RSQ5 are summarised for Study 1 in Chapter 4 on page 71 and for Study 2 in Chapter 5 across pages 95 to 97. RSQ3's answer is summarised for Study 1 in Chapter 4 on pages 75 and 76. For Study 2 RSQ3 is answered in Chapter 5 on pages 104 for physiological results, page 105 for eye gaze results and 111 for facial expression results.

1.3 Contributions

The primary output from of this research is the design, development and implementation of test methodologies to evaluate the influence of instruction format on user QoE for procedure assistance and training roles. The following contributions reflect the impact of this work:

1. This work informs the development of an experimental methodology and protocol that incorporates a comprehensive set of metrics for user evaluations. This holistic approach can be used to derive an understanding of how physiological responses, facial expressions and eye gaze relate to subjective experience in terms of task-load, cognitive load and QoE.

2. The results of this work inform AR procedure assistance and training application design. A list of optical see-through AR HMD augmentation design recommendations is given in the final chapter of this thesis.

3. The results of this work identify statistically significant correlations between novel implicit metrics and subjective experience. The implicit metrics that cross-correlate to multiple subjective and performance metrics are good candidates for reproducibility in future research and as real time indicators of AR-user QoE.

4. A deep critique of the literature conducted as part of this research summarises the state-of-the-art in QoE evaluation of AR procedure assistance and training applications. The data sets captured during this research have the potential for use at a future time to aid in emotion and QoE classification projects.

The following peer-reviewed publications have resulted from this work.

E. Hynes, R. Flynn, B. Lee, and N. Murray, “A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance,” in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, 2019, pp. 1–6. (Full international conference paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, “A QoE Evaluation of an Augmented Reality Procedure Assistance Application” in IEEE 12th International Conference on Quality of Multimedia Experience (QoMEX), 2020. (Demo paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, “An Evaluation of Lower Facial Micro Expressions as an Implicit QoE Metric for an Augmented Reality Procedure Assistance Application” in IEEE 31st Irish Signals and Systems Conference (ISSC), 2020. (Full national conference paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, “A Quality of Experience Evaluation of Instruction Formats for Procedure Training in Augmented Reality” in the doctoral consortium, ACM 7th International conference on Immersive Multimedia Experience (IMX), 2021. (Extended abstract, doctoral consortium track)

E. Hynes, R. Flynn, B. Lee, and N. Murray, “A QoE evaluation of procedural and example instruction formats for procedure training in augmented reality” in Proceedings of the 13th ACM Multimedia Systems Conference (MMSys), 2022. (Nominated for best demo paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, ‘A QoE evaluation of augmented reality for the informational phase of procedure assistance’, Qual. User Exp., vol. 8, no. 1, p. 1, Feb. 2023, doi: 10.1007/s41233-023-00054-7. (QUEx journal paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, “Towards a symmetrical definition of QoE: An Evaluation of Emotion Semantics in Augmented Reality Training” in IEEE 15th International Conference on Quality of Multimedia Experience (QoMEX), 2023. (Full international conference paper, awarded best paper)

E. Hynes, R. Flynn, B. Lee, and N. Murray, ‘A Quality of Experience Evaluation of Procedural and Example Instruction Formats for Training in Augmented Reality’, Int. J. Hum. Comput. Stud, Aug 2023. (Journal paper in press)

1.4 Structure of the thesis

Chapter 2 offers a critique of the relevant literature related to this thesis in terms of: instruction formats; distinction between procedural and example instruction formats; uses of instructions and a distinction between procedure assistance and training. This leads to a discussion of the influence of instruction formats on cognitive load, which has the potential to limit the benefit of AR for these roles. Furthermore, a discussion of the potential benefits and challenges involved in using AR for procedure assistance and training roles is presented. The field of QoE is introduced as are instruments used to evaluate QoE (post-experience questionnaires, facial expressions and physiological ratings). A discussion on how QoE can be used as a vehicle to evaluate the influence of instruction formats and the relevance of emotional and cognitive components is presented.

Chapter 3 presents the research methodology employed in this research. This includes an overview of the seven-phase experimental protocol common to both studies (Study 1 and Study 2). It also includes a description of the data and statistical analysis performed in both studies.

Chapter 4 provides a detailed description of Study 1. It begins with a discussion of the study aims, which were to evaluate the influence of paper-based and AR-based text instructions on user QoE for procedure assistance. The specifics of the methodology of Study 1 are detailed. It includes a discussion of the specifics of the seven-phase protocol relevant to Study 1. The paper-based and AR-based text instruction formats are described, and the development of the AR procedure assistance application is detailed. This is followed by a discussion of data analysis and results. Chapter 4 concludes with a summary of the key findings and how it partially informed the methodology of Study 1.

Chapter 5 presents Study 2. It outlines the aims of the second experimental study i.e. to evaluate the influence on user QoE of a text-only instruction format compared to a combined text and animated interactive 3D model instruction format within AR for procedure training. The elements of the methodology that are unique to Study 2 are detailed, including a discussion of the protocol. The text-only and combined instruction formats are discussed, followed by a description of the data analysis that was carried out as part of Study 2. Chapter 5 is concluded by a summary of the main findings from Study 2.

Chapter 6 concludes the thesis by revisiting the results of Study 1 and Study 2 and how they answered the research questions and sub research questions. Future research opportunities arising from this work are identified. Guidance for the methodologies of such future work is given in the form of a cost/value analysis of the instruments used during this research. AR design recommendations are given arising from lessons learned in this work to aid in future research of AR applications. Finally, the limitations of this research and how they might influence the results presented in this thesis are acknowledged.

CHAPTER 2

Literature Review

This chapter presents a review of text and graphical instruction formats used in procedure assistance and training literature. It includes the influence of instruction format on learning via cognitive load during training. It also critiques the relevant literature that has considered AR as a potential platform for delivering procedure assistance and training instructions. AR is presented in terms of hardware and software solutions, and its potential and the challenges involved in using it as an assistance and training platform. QoE is introduced in terms of its definition, its origins, the roles of emotion and cognitive process in QoE. The instruments used to measure QoE are discussed. Finally, relevant research on the impact of human, system and context-level QoE influencing factors of optical see-through AR applications for assistance and training roles is presented.

2.1 Procedure assistance and training instruction formats

Instructions provide a means to communicate how to perform a procedure; they can be presented in procedural and example formats [8]. Procedural instructions tell the user how to perform a specific task. Precision is crucial to the utility of procedural instructions (e.g., mathematical formulae or detailed text). Examples show the user how to perform a specific task; they resemble the task and provide users with an opportunity to better conceptualise what they should expect when doing the task themselves. Example instructions can be passive or interactive [9]. Passive examples can include images and video [10], [11]. Interactive examples can include dynamically changing models of the workpiece [10]. Interactive instructions are essential for interactive activities [10], such as feedback during training to ensure correct learning.

Instructions serve various purposes ranging from singular assembly tasks [3], [12] to procedure training [13]. The pedagogical objectives of training encompass learning and the application of acquired knowledge, commonly referred to as transfer [14]. Learning is not necessarily the objective during direct assistance with procedures that are unlikely to be encountered in future. During training, the trainee's objective is to learn how to perform a procedure [14]. Transfer involves generalising that knowledge to similar procedures in the same domain. The degree of similarity of the learned procedure to previously unseen instances within the same domain is referred to as near or far transfer [14].

Training is a common use case for instructions where automation of repeatable procedures is not practical. This may be the case where humans are more dextrous and more adaptive in certain production value chains [1]. Learning and transfer are influenced by cognitive load. Cognitive load refers to the total amount of mental activity imposed on memory at an instant in time [15]. Cognitive load is in turn influenced by instruction format. The cognitive model of human learning provides a model of mental processes of the human memory system consisting of a series of three discrete memory subcomponents, as shown in Fig. 2.1 [15]. These are sensory memory (SM), working memory (WM) and long-term memory (LTM). These memory components perform stepwise mental processes to acquire, process, store and retrieve information [15], [16]. The SM component acquires a continual stream of new information from the sensory systems. Selective attention and perception initially process the acquired information to extract relevant elements (about 1%) and to discard nonimportant elements [15]. New information that has been attended to and perceived is transmitted to WM. WM receives new information transmitted from SM and prior knowledge retrieved from LTM [15]. WM is

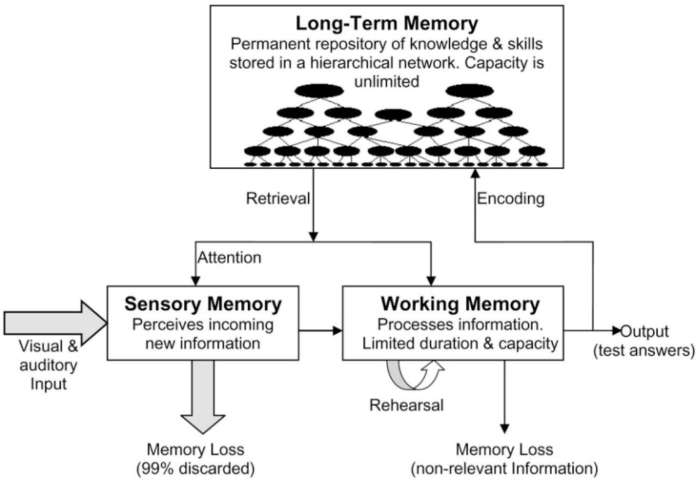

Long-Term Memory
Permanent repository of knowledge & skills stored in a hierarchical network. Capacity is unlimited
Retrieval
Encoding
Attention
Sensory Memory
Perceives incoming new information
Working Memory
Processes information. Limited duration & capacity
Visual & auditory Input
Output (test answers)
Rehearsal
Memory Loss (99% discarded)
Memory Loss (non-relevant Information)

Fig. 2.1. The architecture of the human memory system based on information processing theory [15].

the embodiment of human consciousness and the only memory component where the awareness of information exists. Selected information is organised, processed and encoded for storage in LTM, or used to generate cognitive output. Examples of such output are problem-solving or generating answers to test questions. Storage capacity and duration of LTM are theoretically unlimited, although retrieval over time can become increasingly difficult [15].The fundamental principles of cognitive load theory assume that WM is limited in capacity when processing new, unfamiliar information. WM is actively engaged in comprehension and processing activities when learning. Learning will be ineffective if the cognitive resources of WM are exceeded.

Cognitive load theory makes a distinction between intrinsic, germane and extraneous components of the overall cognitive load that arises during learning [17]. Intrinsic cognitive load depends on the relational complexity and the trainee's degree of prior knowledge of the content. This will reduce as the trainee progresses from novice to expert. Germane cognitive load refers to the cognitive resources involved in encoding the information into LTM. Extraneous cognitive load is influenced by the format that the information is presented in. Extraneous cognitive load can impede learning by using cognitive resources that could otherwise be used for intrinsic and germane cognitive resources required in learning [18].

The information processing literature suggests that the extraneous cognitive load caused by training instruction formats will influence trainee QoE because it can influence the user's pragmatic need of learning. The different extraneous cognitive loads caused by procedural and example instruction formats provide the motivation for the evaluation of these instruction formats for training where learning and transfer are the user's goals. The rapidly changing instructions for mass customisation provide the motivation for the evaluation of AR as an adaptive instruction media. As an emerging technology with potential as a solution for instruction delivery in procedure assistance and training roles, this research examines AR instruction formats considering human, system and context influencing factors. The following section describes AR's potential for procedure assistance and training in the context of these influencing factors as detailed in the following section.

2.2 The potential and challenge of AR for procedure assistance and training

This section introduces AR as a potential platform for the two distinct roles of procedure assistance and training. A description of various hardware and software technologies used to deliver AR experiences is given. The description of AR software solutions includes a discussion on object tracking-based AR application control and augmentation formats.

2.2.1 An introduction to Augmented Reality

In contrast to virtual reality (VR) [19], the fundamental purpose of AR is to enhance the user's interaction with their physical environment. This is achieved by presenting the user with additional information about their environment that is not naturally available to them [20]. This can be realised using a variety of input sensors. The additional information is presented to the user by means of output media. The sensor input and the output media are moderated by software that either combines the input sensory information, or adds additional information to it, for presentation to the user using output devices [20]. AR's place on the virtuality spectrum [20] is shown in Fig. 2.2. AR overlays the user's real-world view with virtual objects, whereas in VR, the user is fully immersed in a virtual environment. In augmented virtuality (AV), real physical objects are controlled by virtual interfaces.

There are several different hardware solutions to AR. These include PC, mobile tablet/phone, spatial projection and HMD technologies [20]. Common to these different solutions is the presence of input sensors (typically a video camera), tracking and graphics software, and output

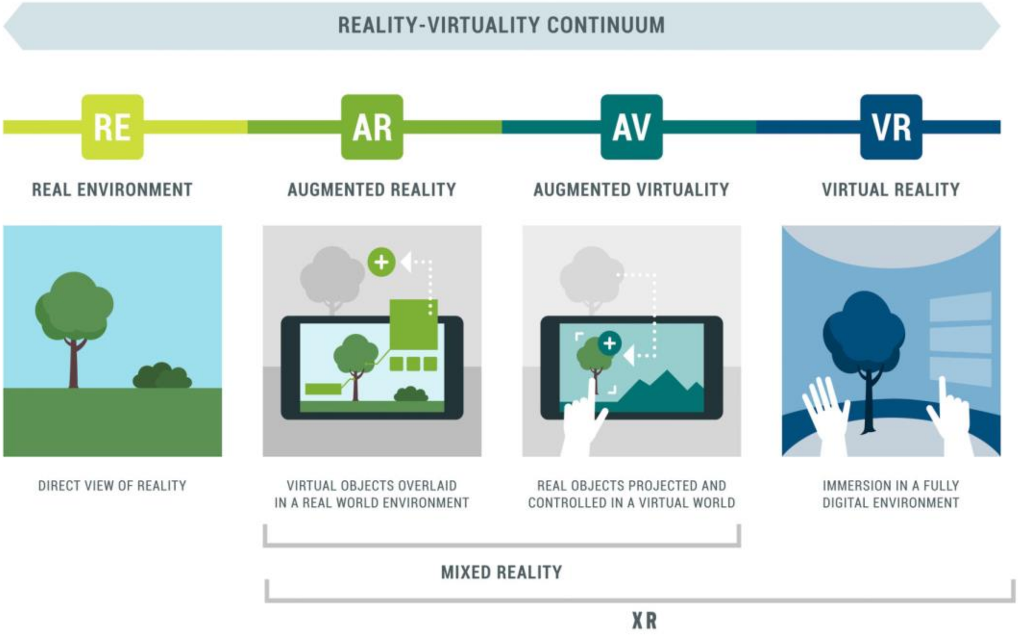

REALITY-VIRTUALITY CONTINUUM
RE
REAL ENVIRONMENT
AR
AUGMENTED REALITY
AV
AUGMENTED VIRTUALITY
VR
VIRTUAL REALITY
DIRECT VIEW OF REALITY
VIRTUAL OBJECTS OVERLAID IN A REAL WORLD ENVIRONMENT
REAL OBJECTS PROJECTED AND CONTROLLED IN A VIRTUAL WORLD
IMMERSION IN A FULLY DIGITAL ENVIRONMENT
MIXED REALITY
XR

Fig. 2.2. Milgram's reality-virtuality continuum [20], [21].

devices (typically a screen) [20]. AR solutions can feature a combination of sensors, including Wi-Fi sensors [22]–[24] . Information gathered and encoded by these sensors from the different frequencies of the electromagnetic spectrum can be combined with those within the range of human perception for presentation to the AR user as audio-visual content [25]. In this way, research has showcased how AR can be used to assist people with perceptual impairments [26], [27] and endow users with superhuman perception [9], [28]. However, AR must be evaluated from the perspective of human, system and context-level QoE influencing factors to prove its potential as a procedure assistance and training instruction platform [1]. The following section discusses various AR systems and their QoE influencing factors.

2.2.2 System-level QoE influencing factors of AR

This section discusses AR system factors that influence user QoE. These system-level influencing factors are divided into AR hardware and software components. Numerous hardware and software approaches to AR exist for various contexts depending on the use-case, environment, lighting, target object-type and augmentation requirements. The discussion of system-level influencing factors of AR software is divided into object tracking-based application control and instruction formats.

2.2.2.1 AR hardware

PC-based AR typically has the benefit of large amounts of computer memory, graphics processing and power, however it has the disadvantage of being immobile. Mobile AR affords the user a full six degrees of freedom (6DoF) in their movement [29]. Mobile AR can take the form of handheld devices and HMDs. Handheld devices such as smart phones and tablets can be used for mobile AR applications. A limitation of handheld devices is that the user typically holds the device with their hands. AR HMDs are more practical for the procedures requiring bimanual manipulation common to many disciplines [30]–[32], as they free up both of the user’s hands to perform the given procedure on the workpiece [13], [19], [25], [33]–[38]. This provides a motivation for evaluation of AR HMDs for the variety of disciplines in which they are expected to be adopted for bimanual procedures in the coming years [31], [32], [39]. Handheld devices provide video pass-through AR functionality. The user views the real environment through the lens of the device’s camera. The two main AR HMD solutions are optical see-through and video pass-through functionality. With video pass-through HMDs, video cameras are positioned directly in front of the user's eyes. The video pass-through AR HMD user does not view their physical environment directly.

Optical see-through HMDs afford the wearer a direct view of their environment, which is augmented with virtual content. Optical see-through HMDs use semi-reflective / semi-transparent screens to allow the user to directly see their environment and the virtual augmentations at the same time. On AR HMDs, the input video camera is typically in the forward-facing position located in proximity to the wearer's eyes. In this way, the camera sees the environment from close to the wearer's perspective. The main challenge with this approach is positioning the camera(s) close to the user's eyes without obstructing their view. This can result in varying degrees of eye-offset, which can cause displacement artefacts in video output on some AR headsets if not corrected for by the AR application. This was experienced in the early stage of this research using the Epson Moverio glasses, where the camera is positioned to the side of the glasses (see Table 2.1, which also shows the hardware specifications of state-of-the-art and market leading see-through AR HMDs.).

Tethered HMDs (e.g., the META 2™) boast more processing power and longer usage durations than mobile headsets (e.g., the Microsoft™ HoloLens 2™ (HL2)) [29]. Wireless HMDs tend to be heavier because they contain onboard batteries as well as a system on a chip for mobile processing requirements, including graphics rendering [23]. One of the main challenges with current mobile AR HMDs is narrow field of view (FOV) [25], [29]. Narrow FOV truncates the augmentations and negatively affects the perception of immersion. Narrow FOV can be overcome in spatial projection AR solutions [39], [40]. This is where projectors are used to project augmentations onto surfaces in the real environment. Projection has the benefit of not requiring head-worn or handheld devices and can display augmentations over a wide area. Legibility of augmentations in this AR solution suffers from uneven surfaces [39] and it works best in indoor environments with low lighting [20]. The META 2™ and HL2 AR HMDs were chosen for their state-of-the-art specifications in processing power and FOV.

Table 2.1. Specifications of state-of-the-art and market leading optical see-through AR HMDs, adapted from [41]. RAM: random access memory. FOV: field of view.

Product	Weight	Display	Hardware	Power	Image
Epson Moverio BT300™	~69 g not inc. pack	24-bit HD colour 23° FOV	Intel Atomx5 16 GB RAM	6 hrs	
Magic Leap 2™	260 g inc. pack	1440x1660 per eye 40 ° FOV	AMD Quad-core x86 128 GB RAM	3.5 hrs	
Google Enterprise 2™	51 g	640x360 25 ° FOV	Intel Atomx5-Z8550 3 GB RAM	8 hrs	
META 2™	420 g	2560x1440 90 ° FOV	Intel Core i5 4 GB RAM	USB	
Microsoft™ HoloLens 2™	~566 g	2k per eye 52 ° FOV	Qualcomm Snapdragon 850 4 GB RAM	3 hrs	

2.2.2.2 AR software

2.2.2.2.1 Object tracking and augmentation

Interaction with and augmentation of real physical objects in the user’s environment is one of the main advantages offered by AR over VR [20], [25]. The term template matching describes the standard approach to determining object state in AR applications [42]. This is where a description of the target object (the template) in its current state is provided to the AR application. This can be achieved by 2D graphical information, 3D models (including computer-aided design (CAD) models), software specifications (which can include colour and edge detection algorithms), machine learning models, or by means of other sensory information. When the AR application detects the template in the input sensor feed (e.g. video), a corresponding output feed can then be augmented with the desired information for presentation to the user [29].

2D graphical representations of the desired target can include fiducial markers such as quick response codes or 2D images of the target object itself. Pre-applied fiducial markers (including lights as used with game controllers [42]) can be placed in the real environment to trigger augmentations at the same location on screen where the fiducial marker is detected in an input video feed. A fiducial marker can be applied to an area of the real environment as seen in Fig. 2.3, or be attached to a specific object that is targeted for augmentation. Fiducial markers are a relatively robust control for unknown environment lighting conditions [42]. However, many AR applications, such as in military or medical applications, do not facilitate pre-application of fiducial markers. Dependency on fiducial marker detection as an object tracking solution is also

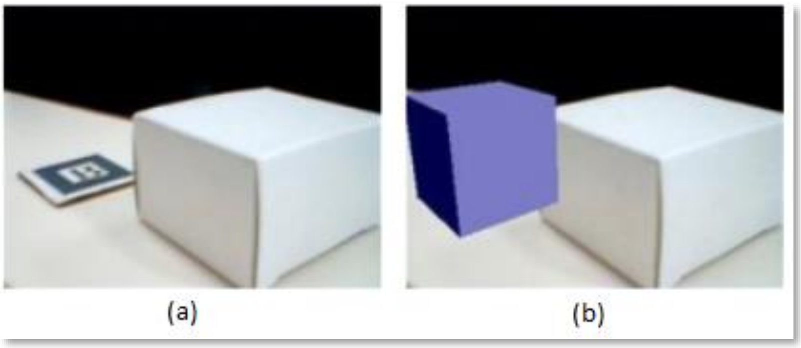


Fig. 2.3. An augmentation of a blue cube is rendered in Fig. 2.3 (b) corresponding to the location of detection of a fiducial marker in the input video stream in Fig 2.3 (a) [43].

vulnerable to marker occlusion [3], [29]. Image-based template matching can eliminate fiducial marker dependency when images of the objects themselves are used as markers. In this way, when the AR application detects the same edge patterns in the target object (e.g. Vuforia uses the scale invariant and feature transform edge detection algorithm) as in the template image, it can track the target object and action the relevant augmentations accordingly. An example of this is shown in Fig. 2.4. This approach has been demonstrated as a robust solution for non-rigid deformable surfaces such as articles of clothing [20].

Consider the case of objects with highly configurable surfaces such as the Rubik's Cube®; if information about the Cube state is required, a template of each desired state is also required [42]. To register each surface configuration of a standard 3x3 Rubik's Cube®, a template for each of the Cube's 10^{19} possible configurations would be required. In any case, this approach only works for objects whose surfaces are rich in texture [29], which is not the case for the standard 3x3 Rubik's Cube®. During the development of the AR application in Study 1 of this research, it was found that image template matching using the Vuforia™ AR SDK did not work for registration and tracking of the standard 3x3 Rubik's Cube® for this reason. This is because Vuforia™ uses grey scale and edge detection filters to detect patterns and does not consider colour; the simple grid pattern on a Cube face was not sufficient for tracking. Vuforia™ also offers functionality for 3D reconstructed textured modelling for target creation. This is achieved by using a standard camera to capture images of real-world objects from 6 DoF. Vuforia's 3D

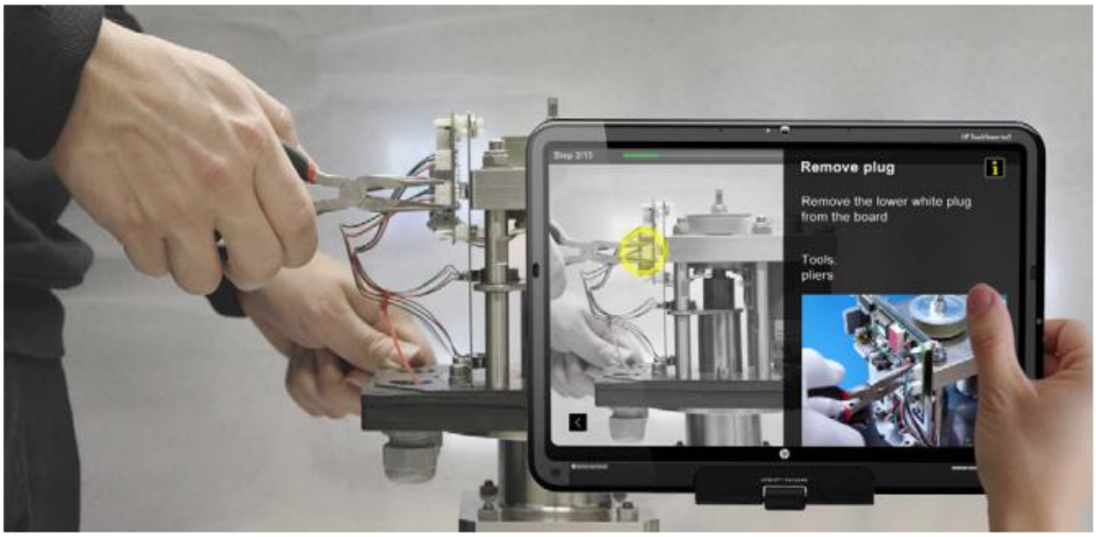


Fig. 2.4. An example of target object pattern recognition in the absence of fiducial markers [44].

reconstructed textured modelling also did not work for the Rubik’s Cube® due to the lack of textures required for its edge detection algorithms. CAD modelling provides a robust solution to 6DoF object pose detection [29] but requires expertise in CAD and is substantially more complex to integrate into an AR application than an AR software development kit (SDK) solution such as Vuforia™. It also does not overcome the requirement for 10^{19} model configurations that would be required to track every possible state of the Rubik’s Cube®.

Machine learning models provide a robust solution to object tracking. This can be achieved using datasets such as the “Common Objects in Context” dataset [45]. The main drawback to this approach is a requirement to train object detection models on large datasets of images of the target object in all possible configurations. This could be achieved using synthetic data generation but does not overcome the object state requirement of training the model on multiple target configurations.

In instances where image or model template matching is not a practical solution due to the quantity of templates required, a dynamic software specification of the object in all its configurations provides a more efficient and robust solution. The main drawback of this approach are the technical skills required for authoring these customised tracking algorithms [46], [47]. It is a relatively complex approach requiring multiple lines of code to describe simple geometric shapes. This is partially because dependency on individual 2D input video frames requires that a geometric shape be best described in terms or ratios of height-to-width to allow for object detection at varying distances and angles from the input video camera. Furthermore, hard-coded colour detection algorithms are susceptible to small changes in lighting intensity and reflection on target surfaces [46]. Development of customised AR applications for tracking specific target objects is complex and requires a range of diverse expertise and evolving tools [29], [46], [47]. Technical software development and animation design skillset asymmetry presents a barrier to deployment of customised AR applications of utility [46].

2.2.2.2.2 AR instruction formats

AR instructions can be presented to the user in many formats. Each format poses its own technical and perceptual challenges with the potential to influence the user’s QoE. Therefore it requires careful evaluation to inform instruction format design [4], [15], [18]. The literature calls for research into these instruction formats to benefit the presentation of information in procedure assistance and training roles [4]. There is a long and accepted tradition of retrieving procedure assistance instruction from detached paper-based media [12]. Research has shown that this can account for up to 50% of procedure completion durations of particular tasks [30]. Procedural instructions usually occur in written format (e.g., detailed text or mathematical formulae) [8].

Text instructions have many characteristics that need careful consideration during their design. These include the colour and size of the instructions. The colour and size of instructions can influence the legibility of the instruction. Instruction legibility is particularly critical for applications where the AR user's attention cannot be taken from the workpiece [48]. In AR, instruction legibility can be impacted by colours and textures in the background of the user's environment [49], [50]. Instruction colours presented in AR can be made to dynamically contrast the background environment colours and lighting to improve legibility [51]. However, the authors of [52] found that participants performed a text identification task quicker, with fewer errors, using static text colours. Billboarding is now a widely used practice to improve instruction legibility [51], [52], which is where instructions are placed within a border with a static solid background colour to minimise the influence of the environment on legibility.

Despite the risk of trainee dependency [8], graphical instruction formats can reduce extrinsic cognitive load [53] and can improve QoE [54]. Graphical instructions can include pictures [55], video [54] or interactive animated 3D models of the workpiece [9]. Interactive instructions have been shown to yield improved task performance in comparison with non-interactive instructions [56]. Skillset asymmetry in graphic design, animation and software development often pose a barrier of entry to realistic and truly immersive AR experiences [46]. Animated asset creation for use as instructions can be developed from technologies ranging from those that require software development skills (e.g., OpenGL™, DirectX™), to those that require graphic design skills (e.g., Blender™, Maya™), to those with drag and drop functionality of pre-existing assets requiring little to no development skills (e.g., Unity™, or Unreal™ asset stores). Dr. Klaus Bengler et al. [48] recommend avoiding the use of animated augmentations for critical AR applications where the AR user cannot be distracted from the task at hand for safety reasons. The authors of [48] stated that user performance is the metric that should define augmentation properties, such as resolution and frame rate, for the given application. AR user distraction is a recurring issue in the literature [57]–[59]. Attention tunnelling [20], [48], [49] (augmentation over-reliance) is a documented phenomenon where users focus too much, or become dependent on, the instructions at the cost of ignoring problems or warnings in their environment.

Examples using precise replicas of the workpiece can reduce extraneous cognitive load [8], [60] of the user by better allowing them to conceptualise an abstract task [53]. However, it can create dependencies that can negatively impact transfer [8], [44], [61], [62]. Van Krevelen et al. [20] recommended guidelines for augmentation design to ensure that the AR user is not overloaded with information during critical applications. This includes by obstruction, clutter, rate of new information (cognitive overload), confusing contradictions or emotional content (cognitive capture). In [49] it is recommend that the AR developer avoids positioning certain augmentations central to the user's FOV to reduce obstruction of the physical workpiece. Multimedia instruction formats can consume more HMD resources than static media augmentations such as text, which is a concern for mobile AR application design where such resources are limited [22]. Lack of graphic processing power can induce augmentation position-lag due to tracking latency. Augmentation lag is the delay in augmentation position in relation to the corresponding object [63]. With HMDs, movement of the target object is generally caused by movement of either the target object itself or by movement of the user's head. Klinker at al. [64] recommend prioritising the reduction of augmentation position latency over augmentation quality trade-offs in resource-intensive applications. This is critical where latency impacts task success. For example, where incorrect augmentation alignment causes mistakes and must be reduced at the cost of augmentation quality. Position of augmentations in relation to target objects can influence user perception of both the target object and the augmentation [65], [66].

2.2.3 Human influencing factors specific to optical see-through AR HMD usage

Optical see-through AR HMD users are susceptible to vergence accommodation conflict and binocular disparity [65], [66]. Humans can only focus on one depth plane at a time. Vergence accommodation conflict occurs where focus of eye gaze on an object at one distance causes divergence of eye gaze from another object at a different distance. Binocular disparity is where augmentations appear to be offset from the object they are intended to overlay. This can be caused by an ill-fitted (tilted) HMD. Vergence accommodation is arguably the more challenging phenomenon in AR development. It is less of a problem for large, static or relatively featureless target objects in the real environment that do not require continued attention by the user. However, consider the relatively small, coloured grid pattern of Rubik's Cube® faces. The user must focus on the Cube tiles to ensure that they are in the correct configuration. Attempting to overlay the Cube tiles directly with semi-transparent augmentations that also require user focus causes vergence-accommodation conflicts. One solution to this challenge is to locate augmentations in proximity to the target object instead of overlaying them directly. This affords the user the opportunity to shift focus from one to the other without straining to try to focus on both at the same time. Video pass-through extended-reality HMDs are emerging as commercial solutions to circumvent these human-level QoE-influencing factors. In video pass-through HMDs, the user sees their real environment means of a video feed recorded by cameras positioned directly in front of their eyes. This video feed can be supplemented with digital content presented to the user on the same depth plane as their environment. The degree to which this video feed is supplemented with computer generated content defines the experience as AR or VR on the reality-virtuality continuum in Fig. 2.2.

2.2.4 Context influencing factors specific to procedural AR application control

Tracking and augmentation of specific real-world objects remain the biggest challenges facing AR since its inception [29], [67], [68]. This can be due to context-level influencing factors such as environment lighting, target object pose, and target object occlusion. These context influencing factors can often fall outside of the control of the AR developer [46], posing a specific challenge during interactive activities such as procedure assistance and training. This is because a procedure is a suite of steps that must be performed in a rigorous order for successful completion. For interactive procedure assistance or training AR applications, this means that a suite of augmentations must be displayed to the user in a precise order. This in turn means that the AR application must accurately determine the target object's state at each step of the procedure [34]. For target objects whose state can change, this may involve changes to their internal structure or to their surfaces. A network-enabled target object can relay its state to the AR HMD via a network. Sensor-based AR solutions may enable detection of changes in target object state [9]. Alternatively, the object's state must be visually determined from its surface configurations using video input [68].

Erroneous object tracking due to environmental influences can lead to unintended delivery of instruction augmentations when used to control procedural instruction delivery [34]. Conversely, an inability to register or track the target object can result in blocking of further instruction delivery [34]. D.E. Qeshmy et al. [69] stated that AR is not an appropriate tool to manage human errors because the technology is not mature enough, citing computer vision challenges as one of the main reasons for this. The challenges posed by accurate target object state tracking coupled with the requirement for repeatable controlled experimentation in procedural AR human trials often gives rise to use of the Wizard-of-Oz approach [70], [71]. This is where incomplete or nonfunctioning parts of the system are simulated by either the researcher or the participant. In procedural AR research, this involves simulating automatic procedural state change of the AR application by means of user input [34]. For example, this can include hand tracking SDKs [63], to approximate the logical position of a handheld object. User input/object tracking hybrid approaches can provide a solution to the requirement for repeatable trials by integrating hand gesture recognition, voice commands, arbitrary template matching, user interfaces or other inputs to control the delivery of procedural instructions [19], [22], [34], [35], [37], [49], [72], [73]. Each of these approaches have their own challenges. Hand gesture recognition uses arbitrary gestures to reduce false positives in natural hand movements [25], but accidental arbitrary hand gestures can also result in mis-cues. Hands-free AR applications are preferable in applications where bimanual procedures are common [25], [33] but voice commands may be susceptible to noisy industrial, medical or military environments. This can be overcome by integrating noise cancellation technologies to ensure reliable speech recognition in noisy environments, even up to environment noise levels of 90 dB [57]. Finally, many industrial, medical or military applications do not facilitate the pre-application of arbitrary template targets such as fiducial markers.

2.2.5 Summary of the potential and challenge of AR for procedure assistance and training

As an emerging technology, AR offers potential benefits for procedure assistance and training. However, the literature shows how this potential may not be realised if the AR instruction format is not designed with consideration of the human, system and context factors that influence QoE. Consequently, the literature repeatedly calls for research into instruction formats to benefit the presentation of information in procedure assistance and training roles [3], [4]. Optical see-though AR HMDs and the Rubik’s Cube® workpiece were used in this work specifically to evaluate these persistent influencing factors. These must be understood to realise the utility of optical see-through AR HMDs for procedure assistance and training roles. This can be achieved by means of QoE evaluations which considers both pragmatic and hedonic user needs. An introduction to QoE is therefore given in the following section to provide an understanding of the evaluation of instruction formats in consideration of these influencing factors from a QoE perspective.

2.3 Quality of experience

This section discusses QoE in terms of its definition, origins and influencing factors. It also discusses the roles of user perception, emotion, and cognitive process in QoE.

2.3.1 The definition of QoE.

QoE is defined as "*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*" [16]. QoE reflects a person's quality judgment of experiences of applications, services or systems. It is an evaluation of the extent that the application, service or system fulfils the user's pragmatic and hedonic needs and expectations considering their context, personality and current state. The user's personality refers to consistency in their behaviour and thinking, while their current state refers to the temporal nature of changes in their thinking and emotion [16]. The user's current state can influence their experience and QoE judgment [74]. The degree of delight or annoyance cited in the beginning of the definition of QoE references two opposing emotions from a spectrum of emotions that can result from the perception, reflection and description of the experience of an application, service or system. This occurs in relation to need and expectation fulfilment.

2.3.2 The origins of QoE.

Telephony gave rise to the requirement for quality of service (QoS) as a measure to determine system performance from a business perspective [16], [75], [76]. QoS uses objective system performance metrics such as throughput, delay, jitter, packet loss, service dependability and customer complaint resolution times [16]. It has been found that telephony signal digitisation and improvements in digital codecs, packet routing redundancy and buffering solutions alone are not the best way to improve user experience [77]. System functionality, user agency and the context of use are given as system-level QoS influencing factors in [76]. The end user's interaction and perceptual acuity are given as human factors that influence the user perceived quality of a particular service [76]. End-to-end networked applications, services and systems have become so interactive and immersive that end user perceptions and interactions are better evaluated from a QoE perspective [75], [76]. This applies to immersive experiences such as AR and VR. This is because the user's interaction is influenced by their perception acuity and the quality of the application interface. Acceptability is influenced by the end-user's subjective experience [78]. Thanks to our common ancestry, although influenced by myriad contexts, the individual human experience is underpinned by a common biochemical response to stimuli (i.e., emotions). Such emotions, in conjunction with additional cognitive processes, can be recorded during a QoE evaluation to shed light a user's QoE.

2.3.3 QoE influencing factors

There are three broad categories of factors that may influence QoE, namely human, system and context influencing factors, as shown in Fig. 1.1. These are summarised in Table 2.2. The human influence factors (HIFs) are categorised into static and dynamic types. The system influencing factors (SIFs) refer to properties and characteristics that determine the technically produced quality of an application, system or service. The context influencing factors (CIFs) embrace any situational property to describe the user's environment.

HIFs, SIFs and CIFs can affect how a user perceives the features of an application, service, or system under evaluation. The user's formation of quality can only be accurately considered in the context within which it was derived [16]. QoE evaluations can take place in controlled laboratory environments to mitigate extraneous CIFs such as variable lighting or noise pollution. It would be practically impossible to set about controlling all the dynamic HIFs that could influence a participant's QoE. It is best practice during a QoE evaluation to take a baseline rating of a participant's state prior to the application of the technological stimulus under evaluation. Therefore, any changes in the participant's state will likely be due to the SIFs that

Table 2.2. Human, system and context QoE influencing factors.

Human		System		Context	
Static	Demographics	Network	Packet loss	Temporal	Time of day
	Ethnography		Packet delay		Time of week
	Physical constitution		Latency	Interaction	Activity
	Mental constitution		Bandwidth		Mobility
	Cognition		Throughput		Language
	Perception acuity		Availability	Physical	Indoor
	Age		Jitter		Outdoor
	Sex	Quality	Audio quality		Noise
Dynamic	Pre-experience		Visual quality		Lighting
	Expectations		Resolution		Weather
	Feelings		Frame rate	Social	Other people
	Moods		Content format		Economics
	Emotions		Error handling		Regulatory

are under evaluation, regardless of the participant's prior state and its myriad causes. The focus of this research is the system-level QoE influencing factor of content format as per Fig. 1.1.

2.3.4 The role of emotion in QoE

QoE refers to the quality of a user's experience of applications, services and systems, experienced through the senses, which can evoke an emotional response in the user. It is not settled amongst researchers if emotions are best represented in dimensions, spectrums or categories [16]. For the purposes of an introduction to the role of emotions in QoE, a user's QoE is visualised in Fig. 2.5 as a value on an emotion spectrum from good, labelled as delight, to bad, labelled as annoyance, as per the QoE definition. This is done as a step towards developing a more in-depth description of how to evaluate a user's emotional state as an indication of their QoE and the resulting quality judgment that they are likely to make.

In his seminal work on emotion classification [79], J. Russell evaluated the semantics of emotion terms, including delighted and annoyed. Each discrete emotion has a unique

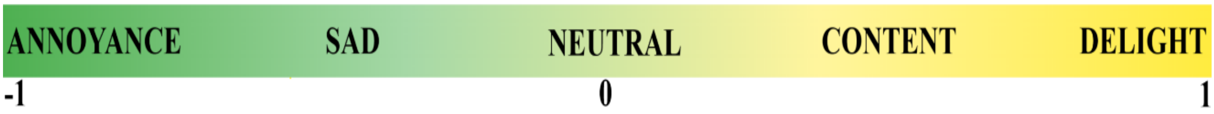


Fig. 2.5. The degree of delight or annoyance depicted as a bipolar spectrum of emotions ranging from negative to positive, including the additional discrete emotions of sad, content, and neutral.

semantic label. For example, feeling somewhat delighted would be better classified as feeling content. Therefore, a user's post experience degree of delight and annoyance should be identifiable using a discrete emotion label from a set of emotions ranging from negativity to positivity. The polarity of pleasure or displeasure of emotion is commonly referred to as its valence [80]. This set, or spectrum, of emotion valence ranging from annoyance to delight including some additional discrete emotions is visualised in Fig. 2.5.

Emotions are described throughout the literature as consisting of more than the valence dimension [79], [81]–[83]. In addition to valence, the authors of [16] identify arousal as the amount of energy in the emotion. To give due consideration to the arousal component of emotion, a two dimensional (2D) visual conceptualisation of the degree of delight or annoyance is shown in Fig. 2.6. It can be seen that both delight and annoyance have a positive arousal component. The addition of two other emotions to this spectrum (sad and content), over and above those given in the 2012 definition of QoE demonstrates that the component that differentiates the spectrum of emotions is not purely one of valence. The difference between feeling content and delighted is one of energy (arousal). It is important to note that Fig. 2.6 was created here solely to illustrate a 2D view of emotions. Annoyance is commonly depicted in the

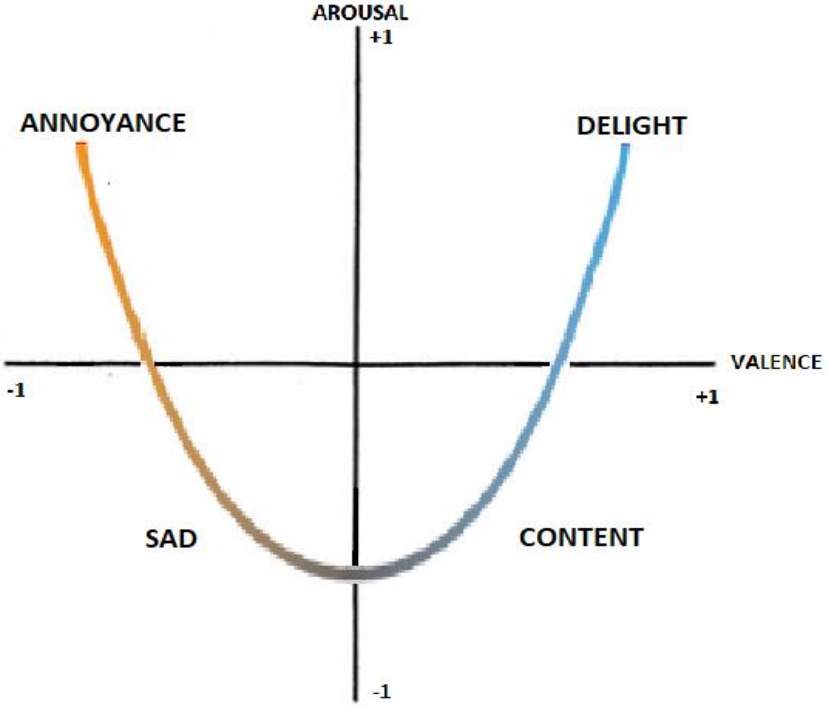


Fig. 2.6. A 2D visualisation of the degree of delight or annoyance.

2D emotion space as having more positive arousal than delight [79]. Delight and annoyance are also rarely if ever identified as having equally opposing amounts of valence [79], [80], [82]–[86].

It was in 1980 that James Russell investigated the amount of arousal and valence in emotion semantics by general consensus [79]. He demonstrated that emotion semantics were consistently positioned in a 2D circle in terms of positive and negative arousal and valence dimensions by a sample of 36 participants. The participants were instructed to position 28 emotion labels within a 2D space. They positioned the labels where they understood such emotions should reside in terms of positive or negative valence and arousal as in Fig. 2.6. This included the terms delighted and annoyed. The labels were sorted so that words at opposite sides of the circle described opposing emotions and those positioned close together were similar. A distance metric had a median correlation of $r = 0.80$ across the 36 participants and correlated to previously theorised positions with $r > 0.90$. Russell concluded in [79] that the resulting 2D circumplex model of emotion provides a convenient means for self-reporting the cognitive conceptualisation of emotion. Consideration can also be given to the dominance dimension of emotion. The dominance dimension of emotion relates to the person's sense of agency in relation to the stimulus. Research into effect has shown that dominance accounts for the least amount of variance in affective judgments [81]. Even so, it is important to record the dominance dimension of emotion in post-test questionnaires to prevent the misattribution of dominance to the valence or arousal dimensions [16].

Under definition of QoE in [16] the user's emotional state is influenced by the fulfilment of both hedonic and pragmatic needs and expectations. S. Möller et al. [16] state that it is far from settled how emotion influences QoE and vice versa. However, it stands to reason that if the experience of using an application, system or service fulfils the user's pragmatic and hedonic needs and expectations, then the user's emotional state will be positive. This is likely to lead to a perception of good quality, and a positive quality judgment will ensue. In theory, the user's emotional state provides a strong indication of the QoE judgment that they are likely to make. In this way, a QoE evaluation that considers the user's emotional state can inform a user-centred design approach towards quality design. What remains is a requirement for scientific studies that empirically correlate physiological and physical manifestations of emotion to subjectively reported QoE. Hedonic features of QoE include aesthetics, usability and interaction quality

[16]. However, high QoE will only be achieved if the application, service or system is perceived as useful in the given context. The user's emotional state will reflect the level of fulfilment of both pragmatic and hedonic needs and expectations.

2.3.5 The role of cognitive processes in QoE

Quality judgements are considered to be the result of cognitive processes in which the delight or annoyance related to the experience needs to be evaluated by the user to come to a QoE judgement [16]. The resulting quality judgement is linked to the identification of sensory, conceptual or actional quality features of the experience. This is influenced by prior experience and expectations. The authors of [16] describe QoE features as characteristics of perceptual events that occur in a multidimensional space (context and time). QoE features are grouped into five layers: (i) direct perception, (ii) action, (iii) interaction, (iv) usage-instances and (v) service-related. The service level includes acceptability as a QoE feature. Although QoE is not based on acceptability, acceptability is the outcome of a decision that is partially based on QoE [16].

While implicit metrics are useful for continuous real-time estimation of QoE during the experience, they do not provide information about the quality features that influenced the user's quality formation process. A comprehensive QoE evaluation cannot rely on physiological or physical responses to the stimulus alone but should also allow the user to subjectively relate their experience [87]. Wechsung et al. [5] recommended using such subjective reports in combination with objective performance metrics to holistically evaluate user QoE. Perkis et al [87] described a multi-method approach, combining subjective (conscious introspection) methods with ecologically viable physical and physiological methods, for QoE evaluation as a viable way to assess an immersive mixed reality experience (IMEx) in all its facets [87]. In this way, each method compensates for any disadvantages of the others. These various methods and instruments for comprehensively evaluating user QoE are described in the following sections.

2.3.6 Instruments used to evaluate QoE

By definition, every QoE evaluation should consider the user's emotional state [88]. In QoE evaluations, the user is effectively an emotional barometer as their emotions portray their affective state [74]. Affective state is directly affected by QoE features of the application, system or service. Affective state is also directly affected by human, system and context factors.

This occurs in a cascading process and the resulting affective state informs the user’s quality judgment [74]. Affective state encompasses moods, feelings and emotions. The field of psychology considers moods as longer-lasting affective states, less likely triggered by particular external events [85]. Moods are oftentimes solely described as good or bad [83]. Scherer [83] describes an emotion as a coordinated episode between neurophysiological systems, reserving the term “feelings” to describe the subjective experience of emotion. Emotions are generally elicited by stimulus events [83]. S. Möller et al. [16] described feelings similarly as the subjective description of the perception of an emotion (e.g. I feel delighted), whereas animals simply experience emotions as a physiological change in preparation for a reaction to a stimulus as an evolutionary survival mechanism. The trunk of the word emotion itself (motion minus the root "e") implies that a tendency to act is inherent in every emotion [89]. In a QoE evaluation, the experiencing of technology impacts the signals presented to the person [16]. Therefore, delight and annoyance, as they are applied in the definition of QoE, are intended as emotional responses to an application, system or service-related experience.

The relationship between QoE and user behaviour is considered to be both direct and compound (i.e., user behaviour as a result of QoE and visa-versa) [74]. This suggests that QoE can be inferred from user state and user behaviour. The user’s emotional response results from all of the usage-instance and service-related QoE features (usability, utility and appeal) of the application, service or system under test, as well as the context (environment) and human factors (including affective state). The authors of [16], [81], [83] stipulated that emotional state evaluation can be achieved in three main ways. These are, (i) explicitly by affective reports, or (ii) implicitly by overt behavioural acts including facial expressions or (iii) physiological reactivity. The following sections outline these methods of emotional evaluation and the instruments that were used to evaluate user QoE of AR for procedure assistance and training in this research.

2.3.6.1 Recording experience using explicit approaches

Questionnaires are commonly used in QoE evaluations to allow the participant to relate their experience [2], [90], [91]. In the following sections, standardised questionnaires for the self-reporting of user experience, affect, task load and cognitive load are described.

2.3.6.1.1 The Likert Scale questionnaire

In 1932, Dr. Rensis Likert published “A Technique for the Measurement of Attitudes” [92], outlining a method to formalise the conversion of a sample’s opinions to ordinal values for statistical analysis. Dr. Likert’s goal was to develop a means of measuring psychological attitudes in a scientific way. Specifically, he sought a method that would produce attitude measures that could reasonably be interpreted as measurements on a proper metric scale.

Likert scales are a non-comparative scaling technique and are unidimensional (only measure a single trait) in nature. Respondents are asked to indicate their level of agreement with a given statement by way of an ordinal scale. This is most commonly seen as a 5-point scale ranging from Strongly Disagree on one end to Strongly Agree on the other with Neither Agree nor Disagree in the middle. Each level on the scale is assigned a numeric value or coding, usually starting at 1 and incremented by one for each level.

In [5], a comprehensive list of QoE aspects was defined. The authors stated that the aim of a system developer is user satisfaction, where users evaluate a system through pragmatic and hedonic quality aspects. These quality aspects must be evaluated by users providing judgments on what they perceive. Such judgments are direct (explicit) QoE measurements, while physiological ratings are indirect (implicit) QoE measurements. Although performance indices are objective measures of system performance, they are indirect measurements of QoE itself. The aspects of QoE are given in [5] as (i) interaction quality, (ii) efficiency, (iii) usability, (iv) aesthetics, (v) utility and (vi) acceptability. These are shown in Table 2.3 in terms of pragmatic and hedonic constituents. The authors of [5] concluded that the use of this taxonomy of QoE aspects for QoE evaluations will help to determine the systematic effect of quality factors on quality aspects, which can then be identified for a given application and weighted accordingly. This method of explicit QoE reporting has a weakness in that it relies on retrospective recall, which can be subject to recency bias in the stimulus [93] and may be subject to unconscientious responses caused by misunderstanding or questionnaire fatigue. For this reason, the Likert scale questionnaire was used to compliment a suite of behavioural and psychophysiological metrics captured in real-time during the experience [87].

ITU-T recommendations include calculating the mean opinion score (MOS) from the questionnaire responses to determine any statistically significant differences between the

Table 2.3. QoE aspects with their pragmatic and hedonic constituents [5].

QoE aspect	Pragmatic	Hedonic
Interaction	Input/Output speed. Naturalness of the interface.	Familiarity. Emotional appeal. Motivation.
Efficiency	Effectiveness in task completion. System learnability.	Effort. Control. Predictability. Transparency.
Usability	Ease-of-use. Effectiveness.	Joy-of-use and satisfaction
Aesthetics	Accessibility. Consistency.	Sensory experience elicited by the system. Personal preferences.
Utility	Requirement satisfaction given interaction effort.	Requirement satisfaction given joy-of-use
Acceptability	How readily a user will use the system.	Enjoyment. Satisfaction. Engagement.

mutually exclusive test groups [94], [95]. As the name suggest, the MOS is the average of the ratings given by the test subjects after they have all experienced usage of the technology under evaluation in the given context. Concern over the loss of information by using the MOS alone and neglecting standard deviations in the ratings has been expressed [96]. In QoE evaluations, the questionnaires are typically completed after the experience.

2.3.6.1.2 The Self-Assessment Manikin (SAM) questionnaire

The SAM questionnaire (see Appendix A) is an affect report, designed to explicitly capture the arousal, valence and dominance components of emotional state [81]. The SAM questionnaire was proposed by Peter J. Lang in 1985 for measuring emotion to simplify the complexities of Russell’s Semantic Differential Method (SDM). The SDM was the previous state-of-the-art for recording explicit affect. The SDM consists of 18 bipolar adjective pairs, each rated on a 9-point scale. Factor analysis of the scores on the valence, arousal and dominance dimensions results in a cumbersome database that requires statistical expertise to resolve. The use of a verbal rating system also restricts use to test subjects who are literate in the given language. The SAM questionnaire is a direct and simple method of affect reporting, overcoming these difficulties associated with the SDM.

The SAM questionnaire depicts the arousal, valence and dominance components of emotion with a graphic character (manikin) along a continuous nine-point scale. For valence, SAM ranges from frowning (negative) to smiling (positive). For arousal, SAM ranges from sleepy to exploding with energy. For dominance, SAM ranges from small, for submissive, to large, for powerful and in-control in relation to the stimulus. Dominance accounts for the least amount of variance in affective judgements [81]. Valence is defined as the dimension of experience that refers to the hedonic note; arousal describes the level of energy in the hedonic note [85]. The participants complete the SAM questionnaire by circling one manikin on each scale, representing the level of the dimension of affect that they experienced. The benefits of using a manakin style questionnaire are that it is non-verbal and hence transcends age, language, culture and cognitive ability; it is quick to complete and can be used in many contexts.

The SAM questionnaire was promisingly evaluated against the SDM in [81]. The authors of [81] demonstrated that paper-based SAM questionnaires correlated with the SDM for valence, arousal and dominance with $r = 0.97$, $r = 0.94$ and $r = 0.23$ respectively. For the two major affect dimensions (arousal and valence), SAM showed almost complete agreement with the far more complicated SDM.

2.3.6.1.3 The NASA-TLX task load questionnaire

In 1988, NASA developed the NASA-TLX questionnaire [97], which proposed a multi-dimensional rating scale in which information about the magnitude and sources of six workload-related factors are combined to derive a sensitive and reliable estimate of workload (see Appendix B). The motivation for this was to create a means to define task overload thresholds for their pilots and astronauts. The rating scale created was the result of 16 task experiments carried out over three years. The aim was to identify variations in subjective workload within and between different types of tasks to define the determinants of workload. The experimental tasks included simple and complex cognitive and manual control tasks. Objective performance data was correlated against variants in subjectively reported perceptions of task load. This resulted in the identification of six determinants of task load: mental, physical, temporal, performance, effort and frustration. Use of the resulting weighted rating scale reduces subjective variations between reported task loads.

The NASA-TLX questionnaire explains the meaning of the determinants to the respondee in the following way. Mental and physical determinants relate to mental and physical demands perceived by the participant during the task. The temporal determinant relates to perceived time pressure during the task. The performance determinant relates to the participant's perceived performance satisfaction in task success. Effort relates to mental and physical exertion required during the task. Frustration is a scale from annoyance to gratification experienced during the task. The NASA-TLX questionnaire consists of two parts as seen in Appendix B. In the first part, the respondent indicates their perceived importance (raw weight) of each determinant on a scale from 0 to 100 in increments of 5. In the second part, the user must choose the more influential between pairs of the six determinants on fifteen occasions (weight). Overall task load is then calculated as:

$$TLX = \sum_{i=1}^{6}(Weighti * Rawi) \tag{1}$$

This provides a composite tailored to the individual's task load definition, where the weighting increases determinant sensitivity and reduces intra-rater variability. Use of the resulting weighted rating scale reduces subjective variations between reported task loads, for the formal assessment of perceived task load of a given task.

The efficacy of NASA-TLX over the twenty-year period since its introduction is outlined in [98]. In this work, Sandra Hart, a NASA employee, conducted a survey of 550 studies that used NASA-TLX. Typically, the NASA-TLX questionnaire is carried out immediately after task completion. The sample of 550 articles was reviewed as a reasonable cross section of the 2,870 joint Google Scholar results for "NASA-TLX" & "NASA TLX" with time constraints as the limiting factor. Audio visual displays and augmented reality-based experiments accounted for 37% of the use cases of the reviewed papers. Most of these studies included measures of performance and many also included measures of physiological function. NASA-TLX has achieved a venerability in the field of, and is used as a benchmark for, task load assessment. The effort and mental load components of the NASA-TLX have been shown to correlate to cognitive load [99]. This has been used for the development of questions specific to capturing cognitive load rather than task load as described in the following sections.

2.3.6.1.4 The Paas cognitive load questionnaire

Mental demand on the NASA-TLX questionnaire was shown to correlate strongly to cognitive load by Prof. F. Paas, with $r = 0.80$ [100]. While the NASA-TLX records task load, the Paas questionnaire was designed specifically to record cognitive load only. The Paas questionnaire considers cognitive load as a construct consisting of mental load, mental effort and performance. Mental load arises at the intersection of task complexity and user ability. Performance is how well the user performs at the given task. Mental effort is how much cognitive effort the user is volunteering to task success. If mental load increases due to increased task complexity, performance can remain constant if the user focuses more. Therefore, task performance is not sufficient on its own to give an insight into cognitive load and mental load and effort should also be considered.

2.3.6.1.5 The Leppink cognitive load questionnaire

J. Leppink et al. [101] used confirmatory factor analysis to develop a cognitive load questionnaire that can allow the respondee to explicitly report a task's influence on intrinsic, extrinsic and germane elements of cognitive load independently. They state that this is a helpful questionnaire to determine the influence of different instruction formats on extrinsic cognitive load because it gives more insight into how different instruction formats can aid some users while hindering others in task performance (e.g., males versus females). The questions they designed are not restricted to a particular knowledge domain. With minor adjustments, these items can be used for research in other complex knowledge domains.

Leppink et al. [101] also cite work where a convergence of new 'biological' metrics compliment these subjective metrics of cognitive load. They state that if both biological and subjective metrics measure the same constructs, one would expect high and positive correlations between them. With that in mind, the following section describes the implicit metrics of cognitive load used in this work, which include physiological ratings, facial expressions and eye gaze features. These features, when used in conjunction with the questionnaires described, allowed for greater insight into the influence of cognitive load on user QoE.

2.3.6.2 Recording experience using implicit approaches

The two means of evaluating experience using implicit approaches cited earlier in Section 2.3.6 were (i) overt behavioural acts and (ii) physiological reactivity [16], [81], [83]. These two methods are described in the following sections.

2.3.6.2.1 Overt behavioural acts.

Posture and head pose can indicate emotional state [83]. The literature has reported that emotion can be expressed in the frequency of head rotation [102], [103]. The head rotates around three axes. These three types of rotation are called pitch, yaw and roll. Fig. 2.7 shows that pitch refers to rotating the head up and down (nodding, Fig. 2.7.a), yaw refers to turning the head from side to side (shaking, Fig. 2.7.b) and roll refers to rotating the head to the side (tilting, Fig. 2.7.c).

Y. Ding et al. [102] used head rotation frequencies and static head position as low-level features for emotion recognition. They used a head motion dataset consisting of an actress reading a script to define the natural range of head rotation frequency as 0-12 Hz, with unnatural head rotations above 14 Hz described as occurring very rarely. They defined standardised head rotation frequency ranges as low (0-5Hz), intermediate (5-10 Hz) and high (10-15 Hz) and evaluated the influence of head rotation frequency on the expression of happiness, sadness, anger, and neutral emotions. They concluded that low frequency head rotation expresses happiness, sadness, and anger together; intermediate frequencies express happiness and anger together; and high frequencies express anger exclusively. While it would be difficult to discern the component emotions in low and moderate head rotation frequencies, the high frequency range exclusively expresses anger emotion.

Head pose includes facial expressions, which can encode and transmit signals of emotion [104]. In 1978 Paul Ekman and Wallace Friesen invented the Facial Action Coding System [104], for automatic detection of facial activity. They defined 46 facial Action Units (AUs), where an AU corresponds to each independent motion of the face. Twenty-one of these AUs concerned movement of the lips, tongue and cheeks. In [104], Ekman et al. assessed the best

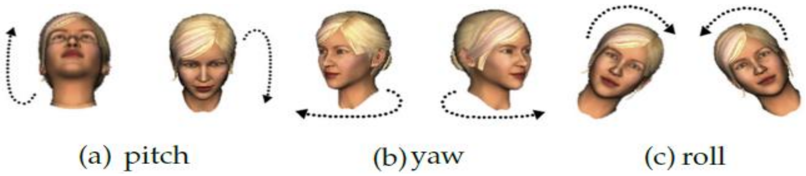


Fig. 2.7. The directions of pitch, yaw and roll rotation of the head [102].

image analysis method for automatic AU detection. The authors showed that the best performance of the automatic classification system was jointly obtained by a local Gabor filter representation and an Independent Component representation. These methods obtained 96% correct classification equalling expert human rater accuracy. Gabor filters are obtained by modulating a 2D sine wave with a Gaussian envelope. Such filters remove most of the variability in images due to variation in lighting and contrast. Independent component analysis learns kernels from high-order dependencies in addition to second-order dependencies amongst pixels. In this way, the work of [104] showed that automatic AU detection is a good mechanism for evaluating affective state.

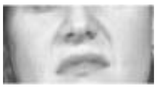
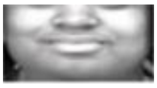
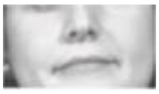
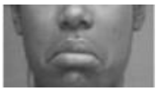
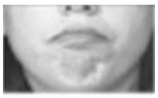
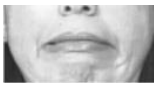
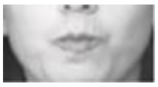
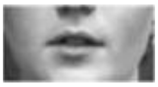
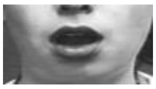
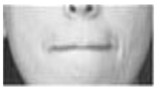
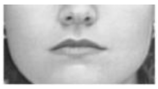
Where normal facial expressions (NFEs) are consciously used for inter-human communication, categorisation of distinct emotions is highly dependent upon context [105]. Consider sarcasm for example, where an individual might smile to consciously express anger, or laugh nervously with fear. Micro facial expressions (MFEs) are described as spontaneous subconscious facial movements [106] that occur when a person experiences emotion [107]. MFEs reveal true and potential expression [106] and are more accurate indicators of a train of thought, or even subtle, passive or involuntary thoughts [108], particularly when the person is trying to conceal or repress that emotion [109]. MFEs are distinguished from NFEs by their sub half-second duration [107]. The subconscious nature of MFEs and their duration, which is so brief as to be imperceptible to the untrained eye, are not intended for inter-human communication. MFEs are more spontaneous (authentic) indications of affect than NFEs [110], [111]. Personality, beliefs, culture, values and socialisation (politeness etc) condition our conscious facial expression of emotion [89], from which MFEs can reveal true intent. Facial expressions have played an important role in inter-human communication but they can now also play an important role in human-computer interaction [111].

MFEs have an upper duration threshold of 502.78 ms and a lower duration threshold of 169.07 ms [112]. W-J Yan et al. [112] demonstrated this by recording the facial expressions of 20 test subjects as they watched a randomised set of 17 videos containing positively or negatively valenced content. The test subjects were instructed not to show any facial expressions for a high stakes financial reward. From 1,000 leaked facial expressions, 245 lasted up to 1 s and 109 lasted up to 0.5 s. The distribution of durations fitted a Gamma model with a Kolmogorov-Smirnov static deviation from a normal distribution of 0.082, with lower and upper bounds of 170 ms and 500 ms (rounded), respectively. MFEs were also shown to feature rapid onset with an upper onset duration threshold of 260 ms [110]. Pfister et al. [113] demonstrated that it is possible to automatically detect micro facial expressions of this duration using a standard video camera recording at 25 frames per second i.e. between 4–12 frames.

In [114], the authors identified consistent (>70% of occurrences) and exclusively expressed facial AUs [104], [115]. While categorising a new set of 15 compound expressions, by combining 1,610 images of 230 individuals' basic facial expressions, the consistent and exclusive occurrence of AU10, AU12, AU15, AU20 and AU26 was observed (see Table 2.4). The authors of [116] independently identified the same set of exclusively occurring AUs. This was done as part of a multi-step selection process towards extending the existing Cohn-Kanade dataset [117] for automatic detection of facial expressions by analysing the expression sequences of a further 26 test participants.

T. Baltrusaitis et al. presented their project named OpenFace in [118]. It is an open-source software tool for automatic facial behaviour analysis, freely available for use in affective computing applications. It is the first open-source tool demonstrating state-of-the art real-time facial landmark detection, head pose estimation, facial action unit recognition and eye-gaze estimation. OpenFace uses conditional local neural field (CLNF) models for facial landmark detection, head pose and eye gaze estimation. CLNF is a recent method of facial recognition and as such, was not evaluated by Ekman et al. in [104]. OpenFace outperformed the competition in predictions for landmark detection, head pose and eye gaze estimation, and AU detection when evaluated against annotated datasets. In the Facial Expression Recognition and Analysis 2015 challenge [119], OpenFace outperformed the competition in 66.67% of classifications. OpenFace can recognise a subset of 18 of the 46 AUs defined by Ekman in 1978. The majority of absent AUs from OpenFace relate to head rotation and eye direction. The full set of lower facial AUs detectable from OpenFace is shown in Table 2.4.

Table 2.4. The lower facial AUs available from OpenFace [118].

AU	Description	Image
10	Upper lip raiser	
12	Lip corners puller	
14	Dimpler	
15	Lip corner depressor	
17	Chin raiser	
20	Lips stretcher	
23	Lip tightener	
25	Lips apart	
26	Jaw drop	
28	Lip suck	
Neutral	Lips relaxed and closed	

2.3.6.2.2 Physiological Reactivity.

The arousal component of a person’s emotional state is exhibited in their physiological signals via the sympathetic nervous system [120], [121]. In a state of heightened arousal resulting from mental, physical or emotional activation [120], [121], heart rate, skin temperature and conductance, and blood volume pulse will all increase together as part of the fight-or-flight response in reaction to a stimulus or situation. Physiological changes are sometimes considered the more objective way to measure emotions as they are difficult to manipulate voluntarily in contrast to self-assessment [16]. Emotions are controlled by older brain structures and are difficult to influence voluntarily [16]. As such, physiological relativity is arguably a more accurate metric of the user's arousal state than subjective reports, which can be subconsciously skewed by primacy, recency and peak stimuli [93], or even disingenuous responses, perhaps due to misunderstanding or questionnaire fatigue. Post-experience questionnaires remain the predominant QoE evaluation instrument used to gain insights into the subjective perception and quality formation process of a user. In light of the limitations with subjective reporting mentioned in this section, a recently published whitepaper [87] calls for the identification of implicit metrics of immersive experiences, such as AR, that can complement the usage of post-experience questionnaires. Accordingly, many QoE evaluations have been undertaken to investigate the value of physiological metrics such as skin temperature, heart rate, electrodermal activity [2], [90], [103], [122]–[125] and electroencephalogram [126].

An emphasis on capture and analysis techniques of physiological processes and experimental design considerations for QoE evaluations of multimedia technologies are given in the comprehensive survey of psychophysiological-based QoE assessment technologies reported in [127]. Psychophysiology is concerned with psychological and physiological correlates. This involves correlating physiological ratings to psychological states. The psychological component of these psychophysiological methodologies typically takes the form of self-reporting questionnaires. In these questionnaires, the test subjects can report their subjective experiences. These questionnaire results can then be correlated to the physiological and objective performance data to gain an insight into how a user's subjective experience is reflected in their physiology. The authors of [127] define electrocardiography (ECG) as a time-varying measure reflecting contraction and relaxing of cardiac fibres, that can be used to measure heart rate variability (HRV), which as a response of the sympathetic nervous systems, is indicative of stress levels. The analysis of the physiological data in [127] includes spectral analysis of electroencephalogram data to study cognitive states that covary with power modulations in different frequency bands. These bands are 4-8Hz for attention and 8-13Hz for alertness. Near infrared spectroscopy data is analysed to extract peak time and amplitude of oxygenated and deoxygenated blood flow, which is prohibitive in the context of QoE due to the technology required [127]. HRV is measured as non-uniform interbeat intervals (IBI), where more uniform IBI indicates higher stress levels. Electrodermal activity (EDA) or galvanic skin response, are measures of skin resistance, and as a response of the sympathetic nervous system to stimuli, can also be used as a measure of stress levels.

Eye gaze is typically analysed in post-processing to create fixation density maps, which include location, duration and order information. Total scanning time and gaze shift rate [32] have been demonstrated to be good implicit metrics of cognitive load [32], [128]. Gaze shift rates are the number of fixations divided by total scan time [32]. A fixation is stationary gaze above a given duration threshold. Pupillometry data is analysed based on the task being performed as tonic (windowed) or phasic (time-varying) pupil diameter. This provides information about the time of pupil dilation, which can correspond to an event. Blink rate and duration can be summed as a measure of fatigue. The authors of [127] highlight the signal analysis methodologies, including filters and the frequency ranges indicative of stress and cognitive load.

It is clear from [81], [127] that physiological reactivity plays an important role in emotional state assessment for QoE evaluations of multimedia technologies. A review of the literature was undertaken to identify the non-prohibitive, least intrusive instruments for evaluating physiological reactivity. In [121], Empatica E4™ photoplethysmogram (PPG) and EDA readings were compared against stationary electrocardiogram and finger skin conductivity electrodes. Twenty-two test subjects with a mean age of 22 undertook a within-group Trier social stress test. This test simulated a five-minute job interview in front of a panel. Reading a five-minute passage aloud and alone was taken as a baseline. Both sensors were used at the same time in testing. A Kubios NRV 2.2 frequency domain analysis of the heart rate (HR) features was undertaken. Features included mean HR, standard deviation of HR and root mean square of successive differences of HR. The authors concluded that the E4 blood volume pulse (BVP) yielded a significant loss of interbeat interval (IBI) data. However, E4 mean heart rate was well estimated allowing good stress discrimination. The skin conductance signal from the E4 was better than the stationary finger electrodes for stress discrimination despite a reduced number of eccrine sudoriferous glands (that regulate EDA) at the wrist compared to fingertips. C. Mc Carthy et al. [129] verified the quality of the PPG data produced by the Empatica E4 by testing it against a clinical standard General Electric SEER Light Extend Recorder Holter portable electrocardiogram. The Holter device required the attachment of circa 10 electrodes to the chest area. The test was carried out on seven healthy test subjects between 21-30 years old. Most of the tests were carried out over 24 hours, two of the tests were carried out over 48 hours. Due to data corruption, only the data from the two 48-hour trials were used. ECG/PPG frequencies are expected to have a repeating pattern; non-repeating patterns indicate noise and were filtered out. Two peer reviewers in the biomedical engineering field reviewed the signals independently to assess subjectivity in the qualitative data quality classification. They concluded that the two devices measured the same quality of data 85% of the time, where the Holter outperformed the E4 5% of the time. The loss of quality from the E4 was because it was worn on the wrist for 48 hours, where movement introduces noise. This is not such an influential factor in controlled laboratory environment testing during relatively short periods where the participant remains seated [94]. The authors of [130] used a 0.05 - 0.4 Hz band-pass filter and second derivative tests during post processing of the E4's EDA signal in the frequency domain for a QoE evaluation of gait correcting media (AR and haptic feedback). This confirmed that the E4's time domain EDA signal contained true EDA signals where the participants were moving.

2.3.6.3 Implicit cognitive load evaluation

As stated previously, physiological reactivity occurs in response to increased emotion activation and physical or cognitive exertion [120], [121]. Attempts to include a cognitive component into definitions of emotion have been resisted by the research community. Theorists argue that emotion and cognition as two independent but interacting systems [83]. Cognitive load interacts with emotion and although considered separately, are linked by physiological reactivity. Cognitive load can be discriminated from stress in physiological ratings [131]. In [131] the authors used mean EDA as a promising feature to distinguish between stress and mild cognitive load with 82.8% accuracy. This was achieved using a methodology consisting of a stress phase (timed solving of arithmetic problems with social evaluation) followed by a cognitive load phase (non-timed solving of arithmetic problems with no social evaluation), rather than the stress-rest phases common to other stress related studies. EDA was used because it is innervated by sympathetic nerves only, thus ideal for recording stress reaction compared to heart rate, for example [131].

Facial expressions can be used to identify mental diversion during times of cognitive load [132]. The authors of [132] used correlations of AUs during cognitive capture phases to create features to classify distraction with 68% accuracy. Many of the cognitive load related AUs that they identified relate to narrowing of the eyes, but also include some lower facial AUs including

AU17, AU23 and AU25 (see Table 2.4). AU17 combined with AU 23 were used as a feature for pensiveness or coping potential, depending on activation of certain upper facial AUs. Some of these cognitive load AUs were independently identified in [133] with the inclusion of AU20, AU26 and AU28 (see Table 2.4).

The authors of [32] used eye tracking as an implicit metric to evaluate cognitive load by correlation to NASA-TLX and Paas cognitive load questionnaires. They evaluated total scanning time and rate of gaze shifts as indicators of cognitive load. Total scan time was measured as the duration from task-beginning to task-end, measured in seconds to three decimal places. The number of gaze shifts is the number of times the participant's gaze fixated on a given target (i.e., an ultrasound machine screen in [32]). The per minute gaze shift rate is calculated as:

$$s = \frac{f}{et - st} * 60 \tag{2}$$

Where s is the gaze shift rate and f is the number of gaze fixations. The number of fixations is divided by the test start time (st) subtracted from the test end time (et) which are in seconds, to arrive at fixations per total scan time (in seconds). This is multiplied by 60 (seconds) to give gaze shift rate per minute. The authors showed that total scan time and gaze shift rate were significant predictors of Paas scale cognitive load ratings. They saw a negative correlation between gaze shift rate and cognitive load, showing each gaze shift was associated with decreased Paas scale rating. This is likely due to increased dwell times as total scan time was a positive correlate of total cognitive load.

An increase in the occurrence of micro-facial expressions has been shown to be an indication of cognitive load [134]. The expression of certain AUs has been strongly correlated to cognitive load [132]. A correlation between cognitive load and QoE has been seen in an education context in [54]. Certain facial expressions have been shown to correlate to physiological indicators of stress [135], [136]. Stress has been shown to influence QoE [2], [58], [126]. Lower facial AUs have been successfully used in stress and cognitive load evaluations [132], [135]. These can be used where AR HMDs occlude upper facial AUs. There is potential to evaluate the influence and relationship between QoE, affect, stress and cognitive load using NFEs and MFEs [111], [132], [136], [137].

2.4 Summary

The literature identifies AR as an emerging technology with potential in distinct procedure assistance and training roles. AR has the potential to replace paper-based instructions for procedure assistance. This provided the motivation for a comparison of text instructions on AR-based and paper-based media. Instruction formats can influence learning by means of cognitive load. This provided the motivation for an evaluation of text-only and combined text and interactive animated 3D model instruction formats for AR-based training, where learning is the goal of training. This includes the influence of context and human factors, such as user perception of the instructions in a given context. QoE considers the degree of fulfilment of a user's pragmatic and hedonic needs and expectations considering system, human and context influencing factors. This provided the framework within which to evaluate the influence of these instruction formats. This thesis contributes to the state-of-the art in QoE evaluation of these instruction formats using state-of-the-art optical see-through AR HMDs by integrating a comprehensive suite of QoE metrics into the methodologies of two studies. This includes task performance, physiological ratings, facial expressions, eye gaze, task-load, cognitive load and QoE.

ITU-T recommendations [94], [95] informed the protocols of the studies undertaken as part of this research, including sample sizes, informed consent, controlled testing environment, and gender balance in the sample. Both studies undertaken as part of this research were conducted in a test laboratory that controlled for environmental background colour [95]. Magenta text is a commonly encountered colour in AR experiences due to contrasting with natural backgrounds to increase legibility of augmentations. Magenta text instructions was used in both Studies in this research.

Favourable reviews of the medical grade Empatica E4 in the literature compared to prior state-of-the-art medical grade devices motivated its use to capture the participants physiological ratings in both studies (Study 1 and Study 2) of this research. The E4 has a comfortable and compact wearable form making experimental setup easy. The E4 boasts advances in clean data acquisition with built in artefact removal in the PPG sensor. OpenFace performance compared to alternative facial feature estimation projects reported in the literature inspired its usage in both studies in this research. The open-source nature of the OpenFace project facilitated integration in these methodologies. OpenFace output contains presence of facial AUs. This is convenient for facial expression detection. Of the 18 AUs detected by OpenFace, a subset of 10 lower facial AUs, plus neutral, (see Table 2.4) were used in this work. These lower facial AUs were used where the AR HMD could potentially occlude upper facial AUs. Post-processing scripts were written to automate the categorisation of AUs as NFEs and MFEs by presence duration thresholds in line with the literature. Contiguous AU presence durations less than 0.5 s were classified as MFEs and contiguous AR presence durations greater than 0.5s were classified as NFEs.

A summary of the QoE evaluation instruments used in this work is given in Table 2.5. The SAM and NASA-TLX questionnaires were used in the methodologies of both studies due to their proven efficacy, simplicity and widespread usage throughout the literature. The SAM questionnaire allowed the participants to report their post-experience affective state. The

Table 2.5. Summary of the QoE evaluation instruments used in this research.

Category	Instrument	Features	Metric
Explicit (questionnaires)	Likert scale	Interaction, usability, aesthetics, utility, acceptabillity	QoE aspects
	Self-Assessment Manikin	Valence, arousal, domiance	Affect
	NASA-TLX	Mental, pyshical, temporal, performance, effort, frustration	Task load and cognitive load
	Paas	Overall cognitive load	Cognitive load
	Leppink	Implicit, extraneous and germane cognitive load	Cognitive load
Implicit	OpenFace and video camera	Head roation frequency Mirco and normal facial expressions	Affect Affect / cognitive load
	Empatica E4™	Physiological reactivity: EDA, BVP, HR, IBI, skin temperature	Arousal / stress
	HL2 eye tracking sensors	Gaze dwell, gaze shift rate	Cognitive load

NASA-TLX questionnaire provided a convenient means to allow the participants to report their perceived task load. Correlation analysis across Likert scale, SAM and NASA-TLX questionnaires allowed for the corroboration of consistent subjective reporting of affective state, cognitive load and elements of QoE. Correlates between objective, implicit and subjective metrics contributed to the understanding of their influence on user QoE. The common elements of the methodologies of Study 1 and Study 2 undertaken during this research are described in detail in the following chapter.

CHAPTER 3

Methodology

This chapter gives an overview of the commonalities between the methodologies and protocols used in both Study 1 and Study 2 which were undertaken as part of this research. This includes a description of the statistical analysis methods applied in both studies. The unique elements specific to each study are discussed in detail in Chapter 4 and Chapter 5.

3.1 Experimental methodology

This research is based on mixed methods experimentation with between-subjects study designs, relying heavily on both quantitative and qualitative research methods. Each study had a core experimental design that was applied to capture explicit (questionnaire responses), implicit (physiological, facial expression and head rotation) and objective task performance data. ITU-T recommendations P.913 [94] and P.919 [95] together with existing QoE research in immersive experiences [2] [122] informed the protocols used in this research. This included recommendations for self-paced between-groups study designs conducted in a controlled testing environment after appropriate informed consent and instruction. The research undertaken was approved by the Ethics Committee of the Athlone Institute of Technology, which became the Technological Institute of the Shannon: Midlands and Midwest by commencement of Study 2. Ethics approval was granted for an adult human trial (17 - 65 years of age, i.e. non-paediatric and non-geriatric). Participant consent was obtained in written format and securely stored. All data collected were anonymized and securely stored under lock and key. Common to both studies was the use of optical see-through AR HMDs, although the makes and models differed in each case. Also common to both studies was a task that involved the use of a Rubik's Cube® based workpiece, although the make of Rubik's Cube® differed in each case. The specifics of these differences are given in Chapter 4 and Chapter 5. The justification for using Rubik's Cube® tasks in this research is given in the following section.

3.2 The Rubik’s Cube®-based tasks

Accurate target object tracking remains the biggest challenge in AR since its inception [29], [67], [68]. This has implications for AR applications of a procedural nature. This is because timely and accurate delivery of procedural AR instructions requires accurate tracking of the workpiece. To evaluate AR’s utility as a procedure assistance and training medium, a procedural task is called for. Optimally solving the Rubik’s Cube® is a procedural task. The instructions to optimally solve the Rubik’s Cube® must be carried out in the exact order as delivered to achieve success. This is not necessarily the case for other workpieces commonly used throughout the AR literature, including LEGO™ [35] and the Towers of Hanoi [138]. With LEGO™ or the Towers of Hanoi tasks, the final goal can be achieved using sub-optimal and unordered sequences. The Rubik’s Cube® controls for the un-ordered execution of instructions by resulting in a failed task. This aids in the evaluation of AR as a procedure assistance and training platform. Successful completion of the Rubik’s Cube® is highly unlikely in any fashion other than the one specifically instructed.

The role of AR for assistance and training roles specific to assembly tasks are very common in the literature [22], [25], [33], [36], [37], [59], [61], [69], [73], [139]–[143]. Werrlich et al. [72] have criticised academic AR research for its overdependence on Lego™ [33], [35], [36], [59], [144] in such research, stating that the results are not “applicable to real industrial use cases” and are not accepted by certain manufacturers. They believe this is one of the main reasons stifling mass adoption of AR in industry. On the other hand, it is believed that training with the Rubik’s Cube® increases general mental rotation abilities [145], [146]. Transfer of mental rotation abilities from Rubik’s Cube® training was evaluated in Study 2 of this research. Furthermore, solving the Rubik’s Cube® involves a multitude of fine motor and visuo-spatial aptitudes. These include alignment, adjustment, and orientation, combined with visual identification, inspection, comparison, and verification. These skills are shared in common across multiple industry, medical and military applications where AR is anticipated to be adopted [13], [31], [147], and are not restricted to product assembly.

The literature calls for the evaluation of AR applications to assist and strengthen worker utility while performing frequently changing procedures [1]. The AR application used in Study 1 of this research was capable of optimally solving the Rubik’s Cube® from one of its 10^{19}

possible states. A 1 in 10^{19} chance of encountering a given Rubik's Cube® configuration provides a robust proof-of-concept for the frequently changing or rarely encountered procedures in mass customisation. Solving the Rubik's Cube optimally is a task that people cannot complete without assistance [148]. Therefore, assistance with solving the Rubik's Cube in a least moves optimal fashion ensures a robust proof-of-concept of procedure assistance.

The use of the Rubik's Cube® across Study 1 and Study 2 allowed for longitudinal analysis of the influence of the instruction formats on the visuomotor Rubik's Cube® interactions common to the assistance and training use cases. The use of an electronic networked enabled version of the Rubik's Cube® in Study 2 ensured the repeatable workpiece state tracking required of the scientific method in human trials.

3.3 Experimental Protocol

The experimental protocols used in Study 1 and Study 2 consisted of seven main phases. The structure of these phases was informed by ITU-T P.913 [94], P.919 [95] and research in QoE of immersive experiences [2], [122]. These phases were:

1. Sampling and information sharing phase.
2. Screening phase.
3. Baseline phase.
4. Instruction phase.
5. Practice phase.
6. Testing phase.
7. Questionnaire phase.

Each one of these phases is described at a high level in the following sub-sections. The distinctions of the phases between the two studies are described in Chapter 4 and Chapter 5.

3.3.1 Sampling and information sharing phase

Participants were recruited by convenience sampling, drawn largely from the post graduate student body of the Technological University of the Shannon: Midlands and Midwest. This was achieved through a combination of scheduling, social media contact and approaching volunteers in person. Details of the sample demographics of Study 1 and Study 2 are provided in section 3.4. Each participant was greeted and thanked for their participation in the testing room. They were then provided with a test information sheet explaining the evaluation in full. After reading this, and upon giving written consent, the participant proceeded to the screening phase.

3.3.2 Screening phase

The same screening protocol was applied during each study. Each participant was first screened for visual acuity using a standard Snellen eye test [149] as seen in Fig 3.1 (a). Then they were screened for colour perception using a digital version [150] of the standard Ishihara colour blind plates (see Fig. 3.1 (b)) [151]. Following this, an interactive digital version [152] of the Vandenberg mental rotation test [153] was administered. The goal of the Vandenberg test is to select the one correct shape, from a choice of four, that matches the given shape on top, as seen in Fig. 3.2. In this interactive version of the test, the participant was able to rotate each shape in three dimensions using mouse input. The participant had one minute to get as many of these correct as possible. This test provides a direct and convenient baseline measurement for mental rotation abilities [35]. Mental rotation involves the ability to rapidly and accurately rotate 2D or 3D objects [154], which is what is involved in rotating the faces of the Rubik’s Cube® [145]. No participants were excluded from testing during the screening phase as recommended in ITU-T P.913. The Vandenberg test provided a dual purpose of screening but also provided a baseline of participants’ mental rotation abilities.

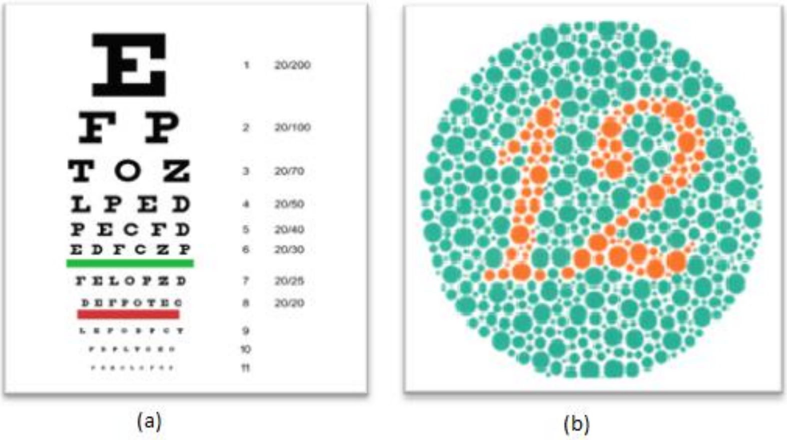


Fig. 3.1. (a) the standard Snellen eye chart and (b) a standard Ishihara colour plate, as administered during screening in Study 1 and Study 2.

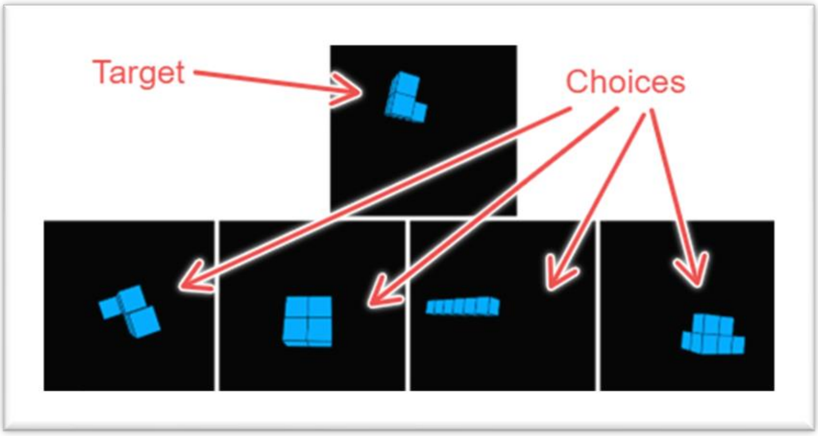


Fig. 3.2. An example of the Vandenberg mental rotation tests used during screening.

3.3.3 Baseline phase

Each study included a five-minute baseline phase. This phase was intended to establish each participant’s state prior to application of the technological stimulus under evaluation. In this way, changes to the participant’s state could be confidently described as having been influenced by the stimulus in question. Common to both studies, the data captured during the baseline phases were used to extract minimum, mean and maximum physiological ratings as described in more detail in Section 3.6. NFEs and MFEs were calculated for this period. The deviation of these features from baseline during the task was used to create deviation features to indicate the influence of the instruction format on the participants.

3.3.4 Instruction phase

The purpose of the instruction phase was to inform the participant how to perform the task that was required of them during the testing phase. Each participant was provided with written instructions describing what they would have to do for the given testing environment that they were assigned to. An opportunity was afforded to each participant to ask any questions prior to continuing to the practice phase.

3.3.5 Practice phase

The purpose of the practice phase was to allow the participants to demonstrate that they understood the instructions provided to them during the previous instruction phase. In both studies, this involved the manipulation of a version of the Rubik's Cube® puzzle. The specifics of the practice phases are described in detail in Chapter 4 and Chapter 5.

3.3.6 Test phase

Common to both Study 1 and Study 2, the participants were required to perform a task as part of the evaluations. Both studies used a version of the Rubik's Cube® as a workpiece. Common to Study 1 and Study 2, each participant's task performance was recorded in terms of error rates and durations in manipulating the Cubes as instructed. The specifics of the tasks involved in Study 1 and Study 2 are described in detail in Chapter 4 and Chapter 5, respectively.

3.3.7 Questionnaire phase

Common to both studies, the participants completed a Likert scale questionnaire, the SAM questionnaire, and the NASA-TLX questionnaire. The Likert scale questionnaires were designed to allow the participants to report their quality judgments on the interaction, efficiency, usability, aesthetics, utility and acceptability aspects of the instruction formats under consideration for the given study. This was achieved by employing the adjectives linked to these QoE aspect in [76] which include usefulness, interest, frustration, distraction, intuitiveness, naturalness, learnability, confidence, stress, joy-of-use, ease-of-use and comfort. In addition to this, the recommendations of ITU-T P. 851 [71] provided guidelines for design the questionnaire statements on interaction quality. Usability, efficiency and utility statements were inspired by IBM's Post System Usability Questionnaire and Computer System Usability Questionnaire [155]. The Questionnaire for User Interface Satisfaction [156] also helped to inform statements on efficiency and interaction. Finally, the Technology Acceptance Model [157], [158] inspired the statement in relation to user acceptability. The participants completed a nine-point SAM questionnaire to report their post-experience affective state. Common to both studies, the participants finally completed the digital NASA-TLX questionnaire to report their subjective task load.

3.4 Sample demographics

In Study 1, a sample size of 48 participants was used. These were assigned to one of two independent groups of 24 participants as calculated in ITU-T P. 913 recommendations [94]. In Study 2, a larger sample size of 60 participants was used. These were divided into two independent groups of 30 participants with equal gender representation, which resulted from an oversampling of 2 participants per group as per the minimum required sample size calculated in ITU-T P. 919 Appendix II [95]. Participants were assigned to the independent test groups of Study 1 and Study 2 based on gender. Study 1 consisted of two independent groups with 12 males and 12 females in each group. Study 2 consisted of two independent groups with 15 males and 15 females in each group. The participants in Study 1 ranged in age from 20 to 64 years old with a mean age of 32 years old. In Study 2, participants ranged in age from 19 to 62 years old with a mean age of 32. Sixteen nationalities were represented in the Study 1 sample including Poland, Ireland, Brazil, India, Egypt, Lithuania, the Philippines, Malaysia, Pakistan, China, Germany, Nigeria, South Africa, Portugal, Latvia and Canada. Twelve nationalities were represented in the Study 2 sample including France, Venezuela and Mexico in addition to many of the same nationalities as Study 1. The participants of different nationalities were randomly assigned to the test groups.

3.5 Statistical analysis methods

Statistical analysis of the data involved five steps including distribution analysis, outlier removal, statistically significant difference analysis, correlation analysis and regression analysis. IBM's Statistical Package for the Social Sciences™ (SPSS) was the main tool used for performing these steps because it facilitates all the statistical analysis listed above with a range of data visualisation options. The most appropriate statistical methods for each of these steps were used during this research are described below.

Tukey's method of outlier removal was performed for a given variable. This is involved in the creation of box plots in SPSS by default. This allows for visual identification of outliers. The Tukey test is a pairwise comparison of all pairs of means in the variable as shown in Table 3.1. If the comparison is greater than a threshold obtained from the distribution, the data point is deemed to be an outlier.

The Student's t-test and the Mann-Whitney U-test are used by default in SPSS to perform independent samples tests to identify statistically significant differences between two independent test groups (see Table 3.1). The t-test makes three assumptions (parameters):

Table 3.1. The statistical methods used to test for normality, homogeneity and statistically significance differences. This includes the test name, purpose and calculation.

Method	Purpose	Calculation
Tukey	Identify outliers	$Q_s = \frac{Y_A - Y_B}{SE}$ Where $Y_A$is the larger of the two means and SE is the standard error. Outlier if Q_s > 1.5.
Shapiro-Wilk	Normality test	$W = \frac{\left(\sum_{i=1}^{n} a_i x_{(i)}\right)^2}{\sum_{i=1}^{n}(x_{i-}\bar{x})^2}$ Where n is the number of observations, $x_{(i)}$ is the ith element of the ordered data, $\bar{x}$ is the mean and a_i are the coefficients where: $a_i = MV^{-1}[(M'V^{-1})(V^{-1}M)]^{-\frac{1}{2}}$ Where M denotes the expected values of standard normal order statistics for a sample of size n and V is the corresponding covariance matrix.
Levene's	Variance test	$W = \frac{(N-k)}{(k-1)} \times \frac{\sum_{i=1}^{k} N_i (Z_{i.} - Z_{..})^2}{\sum_{i=1}^{k}\sum_{j=1}^{N_i}(Z_{ij} - Z_{i.})^2}$ where k is the number of groups, N_i is the number of cases in the ith group, N is the total number of cases in all groups, $Y_{ij}$is the value of the measured variable for the ith case from the jth group and $Z_{ij=} \begin{cases} \lvert Y_{ij} - \bar{Y}_{\imath.} \rvert \; \bar{Y}_{\imath} \; \textit{is a mean of the ith group} \\ \lvert Y_{ij} - \tilde{Y}_{\imath.} \rvert \; \tilde{Y}_{\imath} \; \textit{is a median of the ith group} \end{cases}$
Two independent samples Student's t-test	Statistically significant differences	$t = \frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{\frac{2}{n}}} \; where \; s_p = \sqrt{\frac{{}^{2}_{s}x_1 + {}^{2}_{s}x_2}{2}}$ where the denominator of t is the standard error of the difference between two means, $s^2{}_{X1}$ and $s^2{}_{X2}$ represent sample variance.

Mann-Whitney U-test	Statistically significant differences	$U = \sum_{i=1}^{n}\sum_{j=1}^{m} S(X_i, Y_i)$ $with\, S(X,Y) = \begin{cases} 1, if\ X > Y, \\ \frac{1}{2}, if\ X = Y, \\ 0, if\ X < Y. \end{cases}$ Where n and m are the number of observations in sample X and sample Y respectively.

1. The data sampled is from two independent groups.
2. The means of the data follow a normal distribution (bell curve).
3. The data from the two groups have equal variance (homogeneity).

These assumptions need to be tested to inform the correct usage of the parametric Student's t-test or the non-parametric Mann-Whitney U-test to look for statistically significant differences between the groups. The between-subjects study designs of both Study 1 and Study 2 ensured that the test groups experienced independent stimuli, satisfying presumption 1. To satisfy assumption 2, the Shapiro-Wilk test was used to evaluate the data for normal distribution. The null hypothesis of the Shapiro-Wilk test is that the data are normally distributed. This null hypothesis is rejected if the probability (denoted by *p*, see Table 3.2) of incorrectly rejecting the alternative hypothesis (non-normal distribution) is less than 5% by default. This confidence threshold is called the alpha value, denoted by α (see Table 3.2). A *p* value of < 0.05 allows one to conclude that there is enough evidence (95% confidence) to suggest that the given variable is not normally distributed throughout the sample. Otherwise, the null hypothesis is not rejected.

The variance in assumption 3 is a measure of how far the data is spread out from the mean. It is calculated by taking the differences between each number from the mean, then squaring the differences to make them positive and dividing the sum of the squares by the number of values in the data set (see Table 3.1). The two groups' variances are deemed to be unequal if the Levene's test of variances results in a *p* value of < 0.05 by default. However, if the sample sizes in the two groups being compared are equal, the Student's t-test is highly robust to the presence of unequal variances. In this research, the two independent groups consisted of an exactly equal number of participants.

Table 3.2. The standard statistical symbology used in reporting the statistical results in this thesis.

Symbol	Name	Description
p	significance	This is the probability that a given test's null hypothesis is accepted in error.
α	alpha	Confidence threshold. This is confidence that a type I error is not being made, i.e., rejecting the null hypothesis when it is true. For statistically significant differences in this thesis, $\alpha = 0.05$. For correlation and regression analysis, $\alpha = 0.01$.
r	Pearson's correlation coefficient	Used to report the strength of correlations between normally distributed data.
ρ	Spearman's rho	Used to report the strength of correlations between non-normally distributed data.
R^2	R-squared	Used to report the influence of an independent variable on a dependent variable. Linear regression is used for continuous data. Ordinal regression is used where at least the dependent variable is ordinal, the results of which are distinguished with †. Binary logistic regression is used where at least the dependent variable is binary, the results of which are distinguished by ††.
df	Degrees of freedom	This is added to the statistical results of this thesis to provide information about the size of the sample in question for the given result. *df* is used in calculating the statistical significance of various statistics. Typically: Sample size – number of variables

In SPSS, the result of the Shapiro-Wilk normality test can be accompanied by a Q-Q plot (quantile-quantile plot), that shows the variance of the variable between the two groups as a scatter plot, where low variance is depicted as adhering more closely to a trend line. If the data points vary wildly from the trend line, a Levene's variance test can be carried out if necessary. If the data of both groups have unequal variance, a U-test can be used to evaluate statically significant differences between the groups.

If the variable was found to be not normally distributed, a Mann-Whitney U-test was used to evaluate statistically significant differences between the two independent groups. Otherwise, the Student's t-test was used. These tests were conducted with a 95% level of confidence ($\alpha = 0.05$) by default in SPSS.

A statistically significant difference between the groups was deemed to warrant correlation analysis. This would help to shed light on the influence of correlates on the significant difference. For normally distributed variables, a Pearson's correlation was used; for non-normally distributed data, a Spearman's correlation was used (see Table 3.3). The Pearson's correlation coefficients (*r*) and Spearman's rho (ρ) are used to report the strength of these correlations (see Table 3.2). Only correlations stronger than 0.50 (a moderate correlation) were deemed strong enough to warrant further investigation using regression analysis. Correlation analysis is undertaken to 99% confidence ($\alpha = 0.01$) by default in SPSS.

Regression analysis was performed to allow the reporting of the influence of a dependent correlate on an independent correlate. This was reported as the percentage of variation in the dependent variable that was accounted for by variation in the independent variable, denoted by R^2 (see Table 3.2). This result is accompanied by a *p* significance value. Regression analysis is undertaken to 99% confidence ($\alpha = 0.01$) by default in SPSS. Linear regression was used to analyse continuous data. Ordinal regression was used where at least the dependent variable was ordinal. Binary logistic regression was used where at least the dependent variable was binary (e.g., gender as 1 or 0). Ordinal and binary regression produce pseudo R^2 values that don't account for changes in the dependent variable as accurately as for continuous data and are distinguished with additional symbology in this thesis as described in Table 3.3.

The statistical results are accompanied by the degrees of freedom (*df*) for the given variables in question (see Table 3.2.). This is the number of values in the calculation that are free to vary without violating the presumptions of the statistical test. This is the number of instances in the samples minus the number of variables in the statistical test.

Table 3.3. The correlation and regression analysis methods used in this research, including method name and how the results are calculated.

Method	Calculation
Pearson's correlation coefficient	$$r_{xy} = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2}\sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}}$$ where n is the sample size, x_i and $y_i$are individual sample points, $\bar{x}$ and $\bar{y}$ are sample means.
Spearman's rank correlation coefficient	$$r_s = \rho_R(x), R(y) = \frac{cov(R(X), R(Y))}{\sigma R(X) \sigma R(Y)}$$ where ρ denotes Pearson correlation coefficient, $cov\big(R(X), R(Y)\big)$ is the covariance of the rank variables, $\sigma R(X)\sigma R(Y)$ are the standard deviations of the rank variables.
Linear regression	$$y = \beta_0 + \beta_1 X + \varepsilon$$ where y is the predicted value of the dependent variable for X, X is the independent variable, β_0 is the intercept value of y when X = 0 and ε is variation.
Ordinal regression (logit)	$$P_r(y \le i \mid x) = \frac{1}{1 + e^{-(\theta_i - \boldsymbol{w.x})}}$$ Where x is the number of observations, y are the ordinal responses and w is a set of thresholds $\theta_{1,\dots,}\theta_{k-1}$, Results in Mc Fadden's Pseudo R^2 distinguished in this thesis with the † symbol.
Binary logistic regression	$$p(x) = \frac{1}{1 + e^{-(\beta_o + \beta_1 x)}}$$ Where $\beta_o$is the intercept and $\beta_1$is the rate parameter. Results in Nagelkerke's pseudo R^2 distinguished in this thesis with † † symbols.

3.6 Data analysis

Post-evaluation data analysis involved time domain feature extraction from the physiological, facial expression and eye gaze data. Thirteen-digit UNIX™ timestamps were recorded for each of these implicit data. These timestamps were used to synchronise the implicit data to millisecond precision. Minimum, mean and maximum features were extracted from the baseline and testing phases. In Study 2, this included practice, training and recall sub-phases. The difference from baseline to these phases was used to create an additional deviation feature for each of these metrics; Fig. 3.3 depicts these physiological features. In Study 1, systolic (peak) and diastolic (trough) BVP amplitudes were considered, inspired by [120]. In Study 2, mean BVP (peak – trough) was also considered.

A C930s Logitech™ 1080p video camera [159] was used in conjunction with OpenFace estimation of AU presence was used to classify facial expressions as NFEs or MFEs depending on their duration. Contiguous AU presence durations less than 0.5 s were classified as MFEs and contiguous AR presence durations greater than 0.5 s were classified as NFEs as per [112]. These NFE and MFE features were normalised on a per minute [32], [160] and percentage [161] basis. In Study 1, per-minute AU features during baseline and task were considered, while

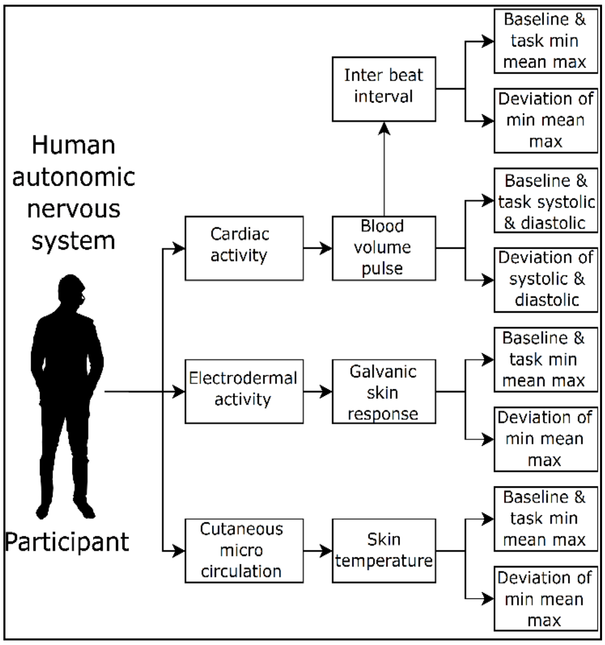

Human autonomic nervous system
Participant
Cardiac activity
Electrodermal activity
Cutaneous micro circulation
Blood volume pulse
Inter beat interval
Galvanic skin response
Skin temperature
Baseline & task min mean max
Deviation of min mean max
Baseline & task systolic & diastolic
Deviation of systolic & diastolic
Baseline & task min mean max
Deviation of min mean max
Baseline & task min mean max
Deviation of min mean max

Fig. 3.3. Acquisition and creation of physiological features, including baseline and deviation from baseline features as inspired by [120], [162].

deviation from baseline of per-minute and percentage AU features were also calculated. In Study 2, MFEs and NFEs were normalised on a per minute basis to create features that change independently of one another (as opposed to percentage of expression features). This helped with interpretability of results.

OpenFace also captured the participants’ head rotation in radians around yaw pitch and roll axes. In post-experience analysis of Study 1, frequency domain analysis was performed on an eight-second duration [163] of post-task head rotation for correlation to evaluate the participants’ emotional state at task completion. Eight seconds allows for the onset-apex-offset perception cycle of affect in head pose. A sample rate of 27 FPS allowed for the detection of frequencies up to 13.5 Hz in line with the Nyquist-Shannon sampling theorem [164]. This allowed for analysis of the full range of natural head rotation frequencies defined in [102]. Linear interpolation [165] of a maximum of 2 FPS was used for up/down sampling of the time domain signal. This time domain signal was passed through a fast Fourier transform in MATLAB™ for frequency domain analysis of head rotation. The high frequency head rotations were divided into 22 bins from 10 Hz to 13.5 Hz. These were used to evaluate the degree of annoyance experienced by the test subjects [102].

Consideration of eye gaze as an implicit metric of participant QoE was unique to Study 2 given the availability of eye tracking sensors on the HL2. Gaze shift rate and instruction dwell features were extracted from the eye gaze data using ray tracing-based instruction hits. A fixation is a stationary gaze of more than 200ms [166]. Eye gaze below this threshold is likely a natural rapid eye movement such as a saccade or an ocular micro-tremor. Gaze shift rate was calculated as the number of fixations normalised on a per minute basis [32]. Specific to Study 2, participants’ open-ended and 2D space emotion terms were assigned ordinal values for statistical analysis.

3.7 Summary

This chapter described how this research is based on mixed methods experimentation with between-subjects study designs over the course of two Studies (Study 1 and Study 2). Each study captured explicit, implicit (including eye gaze in Study 2) and objective task performance data. A description of the methodologies and protocols used in both Study 1 and Study 2 was given. The seven-phase experimental protocol common to both studies was described as consisting of sampling and information sharing, screening, baseline, instruction, practice, testing and questionnaire phases.

The statistical methods and data analysis common to both studies was described. All statistical results presented in this thesis are formatted and presented in accordance with [167]. As such, statistical results are reported throughout the thesis using standard statistical notations as detailed in this chapter. A description of the sample demographics of both studies was given in terms of age, nationality and that the samples of both studies were controlled for equal gender distribution. More specific details of Study 1 and Study 2 are given in Chapter 4 and Chapter 5.

CHAPTER 4

A QoE Evaluation of Paper-based and AR-based Textual Procedure Assistance Instruction Formats

This chapter describes the QoE evaluation of a text-based procedure assistance instruction format. The instruction format was compared after being presented in both AR and paper-based media. This QoE evaluation is referred to hereafter as Study 1. In the remainder of this chapter, the motivation and aims for undertaking this study are given. The methodology is described in detail including the AR and paper-based media, the QoE recoding instruments, the Rubik's Cube® solving procedure and the experimental protocol used.

4.1 Motivation

Assistive instructions are required to aid a person to complete an unfamiliar procedure. Users of procedure assistance instructions often rely on a paper-based format [3], [12]. Assistance with highly variable procedures require more adaptive assistance formats [3], [25] such as AR. The literature calls for the evaluation of AR applications to assist and strengthen human roles in a climate of increasing automation of repeatable procedures [1]. The optimal Rubik's Cube solving procedure was used as a proof of concept of this as described in section 3.2.2.

The aim of Study 1 was to evaluate the influence a text-based procedure assistance instruction format presented in AR and a paper-based medium on user QoE. Text instructions were employed to control for clarity, precision and user comprehension. AR, as an emerging medium with potential for adaptive, hands free procedure assistance, was evaluated against a paper-based control, as the most common procedure assistance medium [13], [140], [142]. This study was facilitated by the development of a test framework that incorporated the capture of a set of QoE metrics, including the user's physiological ratings, facial expression features and self-reported measures in terms of affect, task load, cognitive load and QoE. This allowed for the identification of novel implicit metrics of QoE by means of correlation analysis between these metrics. This study gives due consideration to the user's hedonic needs and expectations by allowing participants to self-report on QoE aspects that influenced their joy of experience and satisfaction.

4.2 Methodology

This section describes the methodology that was used in Study 1. This includes the task that the participants undertook, the paper-based and AR-based procedure assistance instruction media, the AR HMD, and the seven-phase experimental protocol seen in Section 3.3. The Rubik's Cube® proof-of-concept procedure used in this study is described in the following subsection.

In Study 1, a between-groups experiment with a sample of 48 participants was used to evaluate the influence of text instruction formats for procedure assistance presented in paper-based and AR-based media on user QoE. An optimal Rubik's Cube® solving procedure was used in this study. The constraint of solving the Cube optimally is one of solving it in the least number of moves. T. Rokicki, the lead author on the proof of the diameter of the Rubik's Cube® [168], states in [148] that this is not something that people can do unassisted, from any non-trivial Cube state [168]. The sample was divided into two independent groups with equal gender representation of 12 males and 12 females in each group. The participants in the AR group exclusively experienced the AR-based instruction format; the participants in the control group Control (CG) exclusively experienced the paper-based instruction medium. The various elements of the methodology are described in the following subsections.

4.2.1 The evaluation task

Many people learn to solve the Rubik's Cube® unassisted in a suboptimal fashion by repeating a memorised suite of algorithms. These algorithms are generally followed by the solver without knowing if a given Cube manipulation takes the Cube one step nearer to or further from the solved state [168]. Conversely, an algorithm that can detect the Rubik's Cube's® state, such as with AR, can search through the numerous moves from the Cube's current state to the solved state for presentation to the solver [168]. In this way, the solver can know for sure that each Cube manipulation takes the Cube closer to the solved state. In this way, context-aware AR applications can assist humans in performing a procedure optimally. The optimal Rubik's Cube® procedure provides a robust proof of this.

Due to the requirement for standardised and repeatable experimentation of the scientific method, a single Rubik's Cube® state, the superflip position [169], was used for each test in this study. The superflip position has the furthest distance from the Cube's solved state, requiring at least 20 Cube face manipulations to solve using the optimal algorithm [168]. The standard 3x3 Rubik's Cube® has six faces. In Rubik's Cube® nomenclature, a Cube face is referenced by the tile at its centre. This is because each centre tile is bound to one face. In Fig. 4.1 we see how the centre tile of the yellow Cube face matches the colour of the yellow face on the solved Cube. The standard Rubik's Cube® faces are coloured blue, green, white, yellow, orange and red. The participants were instructed to rotate the given Cube face in three ways by reference to Cube face colour. These were (i) a 90^o clockwise rotation, (ii) a 180^0 degree clockwise rotation and (iii) a 90^o anti-clockwise rotation. If the participant correctly followed each of the instructions, they ended the test with a solved Rubik's Cube®.

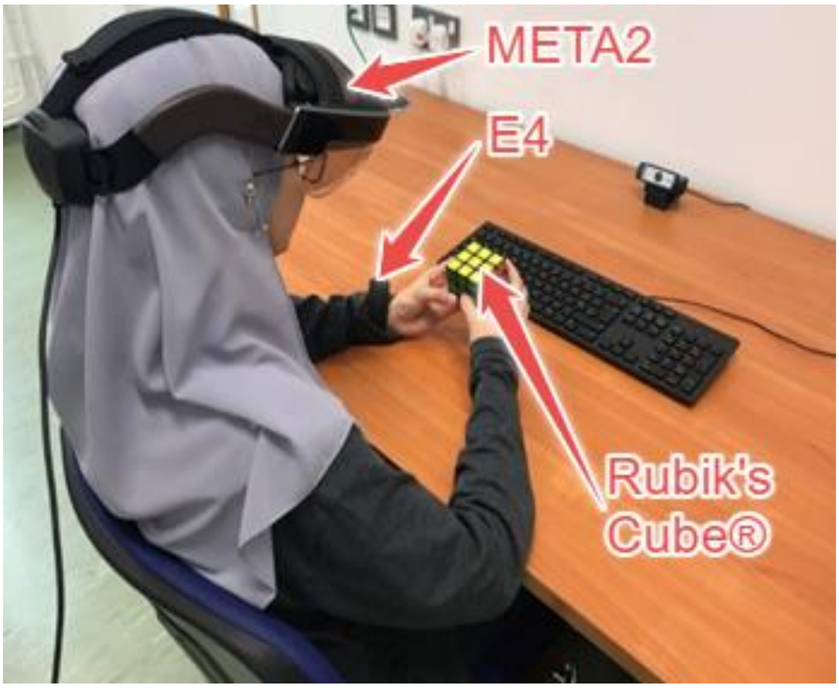


Fig. 4.1. The test set-up for Study 1 including the META 2™ AR HMD, video camera, Empatica E4 sensor, keyboard and the standard 3x3 Rubik's Cube®.

4.2.2 The procedure assistance instruction formats

This section describes the paper-based and AR-based instruction formats that the participants of the two independent test groups experienced; CG participants used the paper-based format and AR group participants used the AR-based instruction format.

For the QoE evaluation of a text instruction format presented in AR as a potential procedure assistance medium, an AR application running on the Meta 2™ AR HMD was designed and developed (see Fig. 4.2 and Table 2.1). This AR application was translated from a Java-based repository [170] into C# for use with the Meta 2™ AR HMD. The AR application used the Kociemba algorithm [168] for optimal Rubik's Cube® solving. At the beginning of each test, the front-facing camera on the Meta 2™ was first used to scan all faces of the scrambled cube. Once the Cube was successfully scanned, the AR application then proceeded to heuristically step through a decision tree of possible moves towards the solved state. For efficiency, the algorithm was configured to consider a two-layer deep decision tree. The steps to solving the cube were displayed in the user's FOV, one after the other. As with the paper-based format, each instruction consisted solely of a line of text, describing the angle, direction and amount of rotation for the given instruction. An example of such an instruction is shown in Fig. 4.2. In both instruction formats, the Cube face names in the instruction were colour

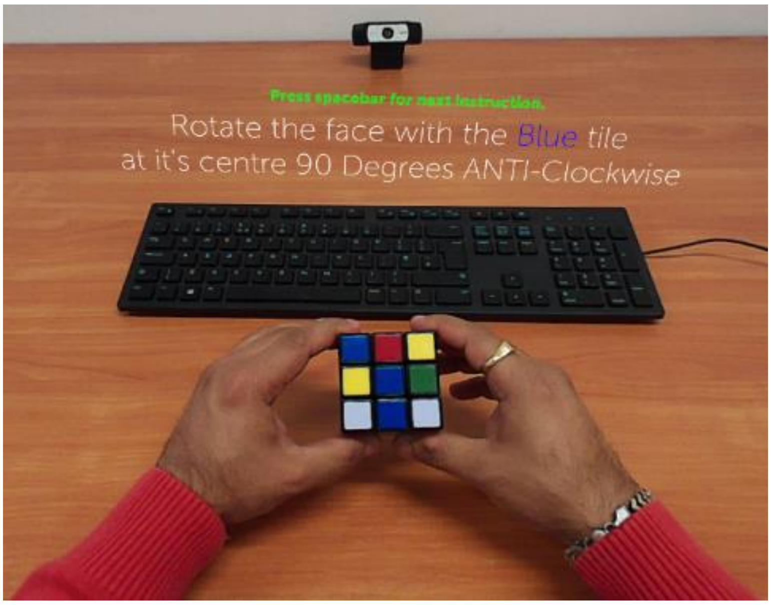

Press spacebar for next instruction.
Rotate the face with the Blue tile
at it's centre 90 Degrees ANTI-Clockwise

Fig. 4.2. An AR participant's view showing the AR text instruction, the keyboard for progression control, and the desk-mounted video camera used to capture facial AUs.

coded. To standardise instruction progression control across both test groups, the AR participant used keyboard input to progress through each instruction.

As a control medium, a 23-page A4 instruction manual (see Appendix C) was created using the same suite of instructions as used by the AR group. This is shown in Fig. 4.3. Each page consisted of one text instruction. Each instruction described the Cube face to rotate, the direction to rotate it and the amount of rotation required. The participant turned each page in turn to progress through the instructions. In both groups, the final instruction simply stated that "*The Cube should now be solved*". The following subsection describes the aspects of the protocol phases outlined in Section 3.3 that are unique to Study 1 and builds upon the information already provided in Chapter 3. This same protocol was applied to both AR and paper-based participants.

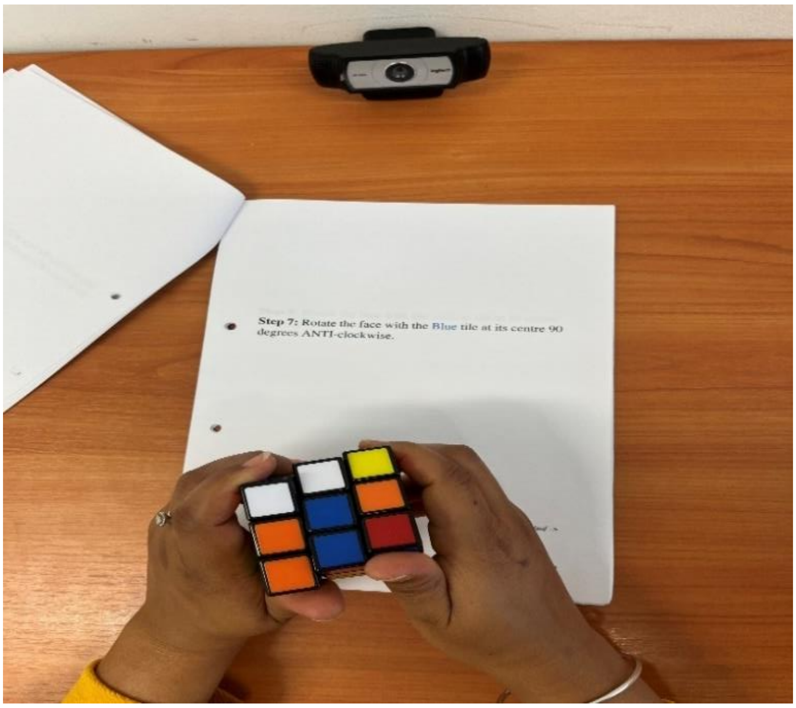

Step 7: Rotate the face with the Blue tile at its centre 90 degrees ANTI-clockwise.

Fig. 4.3. A control group participant's view, showing the paper-based instruction manual (see Appendix C) and the video camera used to capture facial AUs.

4.2.2.1 Phase 1: Sampling and Information Sharing.

None of the participants had prior experience of the Rubik's Cube® assistance instructions. Each participant was provided with the test information sheet in Appendix D. After reading this, each participant completed the consent form in Appendix E. The sampling and information phase lasted on average 3 minutes.

4.2.2.2 Phase 2: Screening.

Details of the screening phase are given in Chapter 3. Thirteen of the participants that were assigned to the AR test group did not have 20/20 vision compared to 11 in the control group. Four participants assigned to each group indicated varying degrees of red-green colour blindness by failing to correctly identify the numbers and shapes in some of the Ishihara colour plates. No participants were excluded during screening in line with ITU-T P.913 recommendations [94]. The participants were not prohibited from wearing prescription glasses during the test. The screening phase lasted for 6 minutes on average.

4.2.2.3 Phase 3: Baseline.

When the participant was fitted with the Empatica E4 the physiological data acquisition began. The beginning of the recording of facial AUs (as per Fig. 4.4) marked the start of the baseline phase.

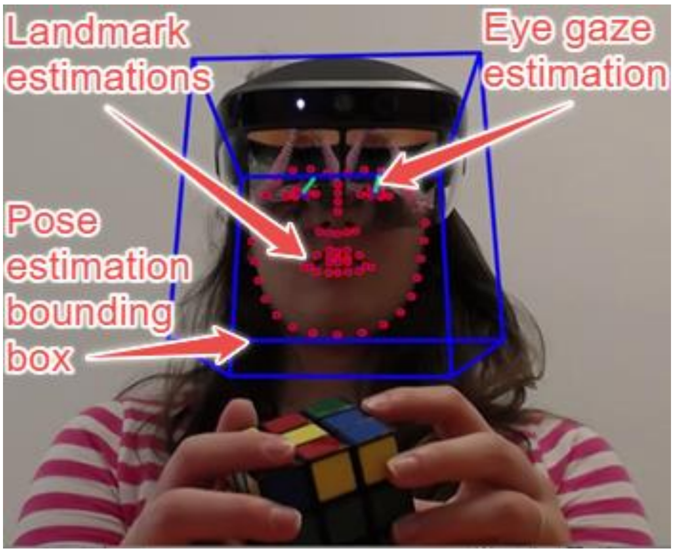


Fig. 4.4. Real-time OpenFace head pose estimation in the AR environment showing the facial landmarks and bounding box estimations.

4.2.2.4 Phase 4: Instruction

Written instructions were provided for each participant, describing the assistance medium they would use (AR or paper). These instructions outlined the requirements of the Rubik's

Cube® solving procedure. The participant was provided with a randomly scrambled Rubik’s Cube® and asked to manipulate it in the manner described by the instructions. In this way, the participant’s understanding of the terminology used in the Rubik’s Cube® manipulation instructions was verified prior to proceeding to the practice phase. The average instruction phase took under 2 minutes.

4.2.2.5 Phase 5: Practice

Each practice run consisted of a fixed set of six instructions. There was one instruction for each of the Cube’s faces. This set included two of each of the three instruction types (90^0 clockwise rotation, 180^o degree clockwise rotation and 90^o anti-clockwise rotation). In the AR-based instruction environment, the participant was fitted with the Meta 2™ AR HMD and presented with a randomly shuffled Rubik’s Cube®. The AR participant progressed through the instructions by pressing the space bar on the keyboard positioned on the table in front of them as per Fig. 4.2.

In the paper-based CG, the participant was presented with an instruction manual containing the same set of six instructions as the AR group; the instruction manual consisted of one instruction per page. The CG participant was also presented with a randomly shuffled Rubik’s Cube®. The CG participant progressed through each instruction by turning each page of the instruction manual in turn, which was recorded by the assessor. Test participants attempted to follow each instruction in turn by manipulating the Rubik’s Cube® as instructed.

Practice run durations, number of required practice runs, total errors and the index of incorrectly followed instructions were recorded to assess the learning curve of both paper-based and AR-based instruction formats. If a test participant made an erroneous Cube manipulation, they were afforded a further practice run. The maximum number of practice runs required by each group to follow all instructions successfully was 2. The average practice phase took just over 2 minutes across both test groups. After the practice phase, the participant proceeded to the testing phase.

4.2.2.6 Phase 6: Testing

At the beginning of the testing phase, the participant was presented with a Rubik’s Cube® in the superflip position. Participants of the AR group continued to wear the AR HMD from the practice phase to test phase. Participants of the CG were presented with the 23-page test instruction manual (see Appendix C). If the participant followed each step correctly, they ended the test with a correctly solved Rubik’s Cube® as seen in Fig. 4.1. The test ended after recording the final instruction, which simply stated that “*The Cube should now be solved*”. The assessor recorded task completion success rates in each test condition. The average test phase duration was approximately 3 minutes. Recording of physiological and facial expression metrics continued until the end of this testing phase. After the test phase, the participant proceeded to the questionnaire phase.

4.2.2.7 Phase 7: Questionnaires

Participants first completed the five-point 14-statement Likert scale questionnaire in Appendix F. The 14 statements were designed to cover interaction, efficiency, usability, aesthetics, utility and acceptability QoE aspects [5]. Aspects of interaction were included in Likert Statements 3, 4, 5, 9 and 14 in terms of comfort, frustration, confidence and naturalness of the instruction media. Efficiency was covered in Likert Statement 6 and 13 in terms of learnability and intuitiveness. Aspects of usability were included in Likert Statements 7 and 12 in terms of joy-of-experience and ease-of-use. Aesthetics were covered in Likert Statement 2 in terms of user interest. Aspects of utility were included in Likert Statements 1 and 11 in terms of usefulness and joy-of-use. Finally, Likert Statement 10 was designed to capture acceptability.

The participants then completed the SAM questionnaire in Appendix A. Finally, they completed a digital version of the NASA-TLX questionnaire in Appendix B. By the end of the questionnaire phase, a further average of 7 minutes had elapsed, which included the time taken to remove the E4 sensor from the participant. This made the total average evaluation duration 38.5 minutes.

4.3 Results and discussion

In this section, the subjective questionnaire results, objective task performance results, and implicit physiological and facial expression results are presented. This includes the significance of statistical differences and the strength and significance of correlation and regression analysis that was performed on the data. The explicit questionnaire results include the participant’s subjective QoE and quality judgments of their respective instruction medium (AR-based or paper-based). Although the questionnaires were completed last, they are presented here first to facilitate the discussion of correlation results throughout the section.

4.3.1 Subjective results from Study 1

This section discusses the results of subjectively reported QoE, affect and task-load by the participants in post-experience questionnaires.

4.3.1.1 Likert scale questionnaire results

Fig. 4.5 shows the main adjective associated with the Likert scale questionnaire statements, including the Mann-Whitney U-Test statistical significance values. The full table of Likert scale responses is provided in Appendix G for the interested reader. There were statistically significant differences between the groups for Likert Statements 1, 3 and 8. The CG's response to Likert Statement 1 (instruction usefulness) was significantly higher than that of the AR group

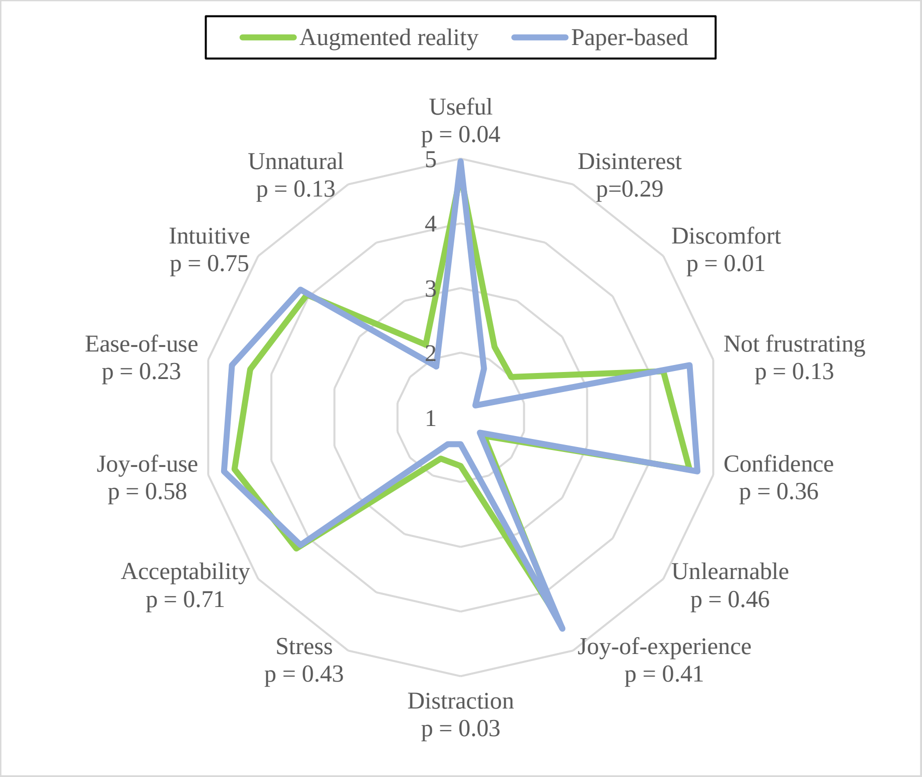


Fig. 4.5. A radar graph of the statistical significance of Likert scale adjectives between the augmented reality and paper-based groups.

with *p* = 0.04. For this evaluation, the instructions were text-only to control for instruction clarity. This result may reflect a lack of hedonic expectation fulfilment in the AR environment, while pragmatic needs were fulfilled as reflected in better task performance results for the AR group as reported in Section 4.3.2. The CG's response to Likert statement 3 (discomfort) was significantly lower than the AR groups with *p* = 0.01. The difference in reported comfort was partially influenced by wearers of reading glasses who reported less comfort with the HMD in the AR group; 6 CG participants and 5 AR participants wore glasses. The Meta 2™ AR HMD was designed for use with spectacles but caused some pressure at the sides of the head. This left a temporarily visible mark on some glass wearers of the AR group after use.

The AR group's response to Likert Statement 8 (distraction) was significantly higher than the CG's with p = 0.03. This statement was posed negatively, signifying that the AR group reported that the AR environment was significantly more distracting than the control environment. In the AR environment, the AR instructions remained in the AR user's FOV throughout the experience. In the control environment, the CG test participants were free to focus their full attention on the Rubik's Cube® once they had read each instruction from the instruction manual. This suggests that if AR augmentations were not carefully designed, they could result in increased distraction from the workpiece. The AR HMD alone may have caused distraction in its own right [57].

Likert Statement 10 was posed as a proxy for acceptability [158]. This was the only aspect ranked in favour of AR (see Appendix G). The participants ranked every other aspect in favour of paper-based instruction. This highlights the complex relationship between QoE and the acceptability of novel technologies [171], [172].

These results suggest that AR applications and HMD design should consider user comfort and user distraction. Augmentations should be designed to afford the user an unencumbered view to minimise distraction from the workpiece. These recommendations may aid mass adoption of AR for applications, where AR HMDs are intended to be worn throughout the working day, and to improve social acceptance in the general population [20].

4.3.1.2 SAM questionnaire results

The distribution of SAM responses is shown in Fig. 4.6. This shows that dominance accounted for the least variance in affective judgments in line with the literature

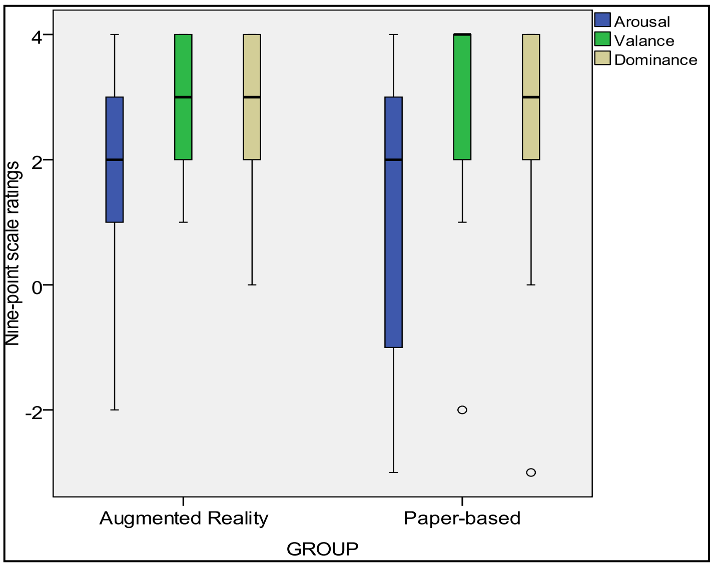


Fig. 4.6. Distribution of SAM questionnaire responses for arousal, valence and dominance from the AR test group and paper-based control group.

[173]. The paper-based CG reported more positive valence than the AR group ($p = 0.16$). In the CG, subjective valence correlated significantly to subjective interest (Likert Statement 2) with $\rho = -0.57$, $p < 0.01$ ($R^2 = 0.32$, $p < 0.01$) and ease-of-use (Likert Statement 12) with $\rho = 0.80$, $p < 0.01$ ($R^2 = 0.\ 89^{\dagger}$, $p < 0.01$) $df = 22$. Fig. 4.6 shows that the paper-based instruction format elicited more negative arousal than AR ($p = 0.36$).

When valence and arousal were combined [79], the difference between the groups was statistically significant with $p = 0.01$. Table 4.1 shows significant correlations common to both groups for SAM affect and Likert scale joy-of-use responses for Likert Statement 7 and Likert Statement 11. These significant moderate to strong correlations demonstrate a positive relationship between positive affective state and positive joy-of-use.

Table 4.1. Significant correlations between subjective affect and responses to Likert scale statements regarding joy-of-use common to both test groups.

		Likert scale questionnaire	
Group	**SAM**	**No. 7: Joy-of-use**	**No. 11: Joy-of-use**
AR	**Arousal**	$\rho = 0.63$, $p = 0.01$	$\rho = 0.62$, $p = 0.01$
	Valence	$\rho = 0.58$, $p = 0.02$	$\rho = 0.58$, $p < 0.02$
CG	**Arousal**	$\rho = 0.50$, $p < 0.01$	$\rho = 0.38$, $p = 0.08$
	Valence	$\rho = 0.89$, $p = 0.01$	$p = 0.82$, $p < 0.01$

4.3.1.3 NASA-TLX questionnaire results

Both groups gave similar total task load scores (AR: 773, paper-based CG: 765 summed, $p = 0.94$) on the NASA-TLX questionnaire. Fig. 4.7 presents the percentage contribution of the weighted determinants to total task load including the statistical significance of these results. This demonstrates that each group perceived total task load in different ways. The paper-based CG's perceived performance was significantly higher with $p < 0.01$, suggesting that despite objective task performance results in favour of AR, as reported in Section 4.3.2 which follows, the paper-based CG felt more confident in their task performance.

In summary of the subjective questionnaire results, the AR group perceived AR instruction to be significantly less useful, more distracting and more uncomfortable. The higher discomfort in AR was largely reported by wearers of reading glasses. The AR group reported more frustration in both the NASA-TLX and Likert scale questionnaires. Positive arousal and valence on the SAM questionnaire correlated significantly to positive joy-of-use on the Likert scale questionnaires across both test groups. Higher valence in the CG also correlated moderately to interest and ease ease-of-use.

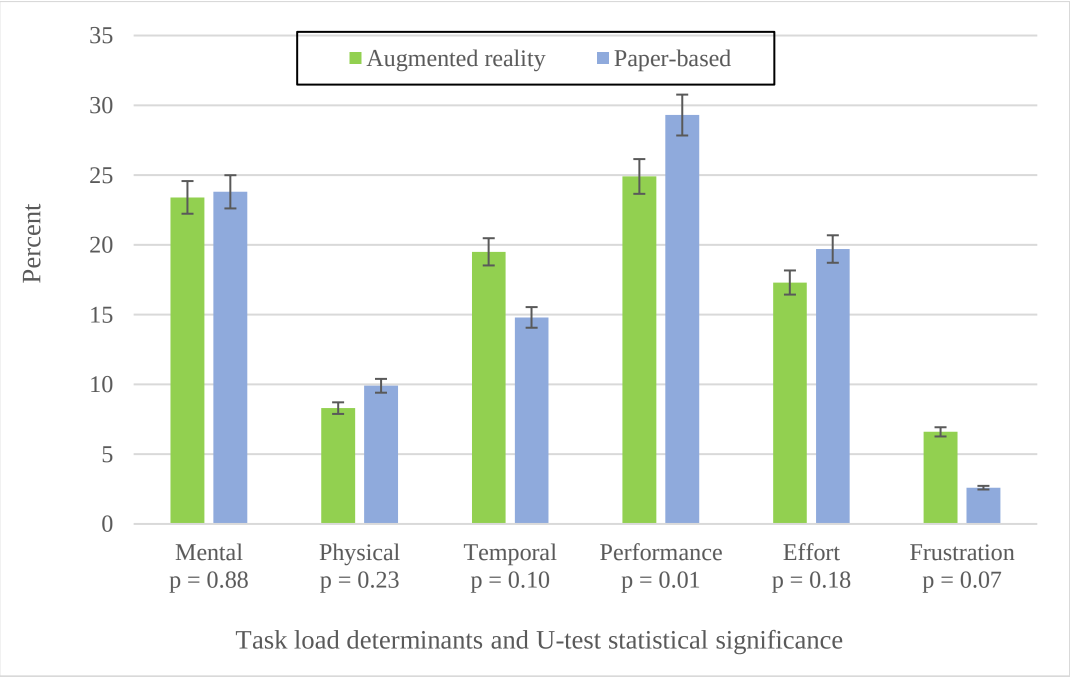


Fig. 4.7. Percentage contribution of the six task load determinants to overall task load on the NASA-TLX questionnaire for the AR group and paper-based control group, including U-test statistical significance.

4.3.2 Objective and implicit results from Study 1

This section reports the objective and implicit results of Study 1, including task performance, physiological and physical metrics. The physiological metrics are BVP, HR, IBI, EDA and skin temperature. The physical metrics are facial expressions and head rotation frequencies.

4.3.2.1 Task performance results

The AR group took significantly longer to complete the practice phase than the CG (AR: 2.28 mins, CG: 2.15 mins) with $p = 0.01$, $df = 47$. This suggests that newcomers to AR may require more time to familiarise themselves with this novel medium. CG females completed the practice phase significantly faster than CG males (females 2.1 mins, males: 2.7 mins) with $p = 0.02$, $df = 23$.

During the testing phase, the AR group produced a correctly solved Rubik's Cube® 96% of the time compared to 94% for the CG with $p = 0.56$. When this is broken down by gender, 100% of male participants completed the task with a solved Cube, compared to 87.5% of female participants, with $p = 0.08$, $df = 47$. A t-test showed that mean task completion times (AR: 2.4 mins, CG: 2.7 mins) were statistically significant with $p = 0.04$, $df = 47$. These findings compliment the results of [35], [37], [39], [171], showing that AR offers efficiency and productivity gains by a 12% reduction for procedure completion durations for the AR group compared to instruction retrieval from detached paper-based media. Overall, the female participants were quicker to perform the task at 149 s compared to 156 s for males with $p = 0.47$, $df = 47$.

4.3.2.2 Implicit results

This section reports the physiological ratings, facial expression and head rotation results, reporting on the statistically significant differences and correlations that were seen between physiological metrics, task performance and subjective experience. The following subsection starts with the differences between the groups' physiological features, followed by how they correlated to task performance and subjective experience metrics. The relationships between the significant correlates are reported by means of regression analysis. The correlation strength, p value significance, R^2 strength and degrees of freedom (df) of these relationships are reported.

4.3.2.2.1 Physiological results

Table 4.2 shows the mean values of, and the statistical differences between, the groups' physiological features: BVP, IBI, EDA and skin temperature. Similar baseline ratings suggest that the samples were well balanced in terms of physiological ratings. Deviation from baseline is how much the rating increased or decreased during the testing phase relative to what was recorded during the resting baseline phase. Interestingly Table 4.2 shows typically larger deviations from baseline in the minimum features, perhaps demonstrating greater utility of minimum ratings than the standard use of average values alone. Usage of minimum, mean and maximum values combined gives a more complete picture of physiological deviations in Table 4.2. Reducing IBI ratings and a reduction in the difference between systolic and diastolic BVP ratings show an increase in stress levels in both groups [162]. This increasing stress was marginally higher in the AR group with their reduction of systolic BVP being significantly greater than that of the CG (highlighted in grey in Table 4.2).

In the CG, deviation of systolic BVP correlated negatively to its increase of percentage of AU26 NFEs (AR: -1, CG: 2) with $r = -0.56$, $p = 0.01$, df = 22. Thirty one percent of the variance seen in the CG's increase of AU26 was accounted for by the variance seen in their lesser decrease in systolic BVP ratings with $R^2 = 0.31$, p < 0.01, df = 22. The CG's combined subjective arousal, valence and dominance accounted for 78% of the variance seen in their higher minimum skin temperature during the task with $R^2 = 0.78$†, p = 0.02, df = 22. This negative correlation of ρ = -0.59, p < 0.01, *df* = 22, is due to the CG's 85% more negative subjective arousal as seen in Fig. 4.6. The CG's greater increase of minimum skin temperature correlated moderately to its greater perception of joy-of-use (Likert Statement 7) with ρ = 0.62, *p* < 0.01, *df* = 22 and ease-of-use (Likert Statement 12) with ρ = 0.58, *p* < 0.01, *df* = 22. Higher joy-of-use in the CG (Likert Statement 11) accounted for 28% of the variance seen in their maximum IBI ratings during the task with $R^2 = 0.28^{\dagger}$, *p* = 0.01, *df* = 22. This higher joy-of-use correlated negatively with its higher physiological stress (shorter IBIs) with ρ = -0.54, *p* = 0.01, *df* = 22. This suggests that the arousal implied in the CG's higher skin temperature was positively valenced as per Fig. 4.6.

Longer task durations in the control environment accounted for 29% of the variance in the CG's reduction of maximum IBIs (increasing stress) with $R^2 = 0.29$, *p* = 0.01, *df* = 22. The raw

Table 4.2. Statistical significance and mean values of the physiological ratings for the AR group and paper-based CG.

Physiological feature	AR	CG	Result
Baseline minimum skin temp.	31.2 ^{o}C	31.6 ^{o}C	*0.39***
Test minimum skin temp.	32.2 ^{o}C	33.0 ^{o}C	*0.12***
Baseline mean skin temp.	31.7 ^{o}C	32.3 ^{o}C	*0.19***
Test mean skin temp.	32.3 ^{o}C	33.1 ^{o}C	*0.11***
Baseline maximum skin temp.	32.2 ^{o}C	32.8 ^{o}C	*0.18***
Test maximum skin temp.	32.4 ^{o}C	33.1 ^{o}C	*0.11***
Baseline minimum EDA	1.4 µS	0.9 µS	*0.73**
Test minimum EDA	2.3 µS	2.1 µS	*0.63**
Baseline mean EDA	3.2 µS	2.2 µS	*0.42**
Test mean EDA	2.7 µS	2.6 µS	*0.53**
Baseline maximum EDA	4.5 µS	3.5 µS	*0.44**
Test maximum EDA	3.2 µS	3.3 µS	*0.28**
Baseline diastolic BVP	-447.6 nW	-421.8 nW	*0.66***
Test diastolic BVP	-308.6 nW	-334.9 nW	*0.63***
Baseline systolic BVP	464.6 nW	426.2 nW	*0.52***
Test systolic BVP	282.8 nW	357.9 nW	*0.89***
Baseline minimum IBI	0.58 s	0.61 s	*0.16***
Test minimum IBI	0.58 s	0.60 s	*0.41**
Baseline mean IBI	0.786 s	0.792 s	*0.76**
Test mean IBI	0.79 s	0.77 s	*0.67**
Baseline maximum IBI	1.01 s	0.97 s	*0.33***
Test maximum IBI	1.01 s	0.97 s	*0.48***
Deviation of minimum skin temp.	1.0 ^{o}C	1.2 ^{o}C	*0.31**
Deviation of mean skin temp.	0.6 ^{o}C	0.7 ^{o}C	*0.38***
Deviation of maximum skin temp.	0.3 ^{o}C	0.4 ^{o}C	*0.10***
Deviation of minimum EDA	0.9 µS	1.2 µS	*0.71**
Deviation of mean EDA	-0.6 µS	0.4 µS	*0.78**
Deviation of maximum EDA	-1.3 µS	-0.2 µS	*0.90**
Deviation of diastolic BVP	126.5 nW	50.5 nW	*0.29***
Deviation of systolic BVP	-177.4 nW	-32.5 nW	*0.02***
Deviation of minimum IBI	-0.03 s	-0.01 s	*0.54***
Deviation of mean IBI	-0.01 s	-0.02 s	*0.57**
Deviation of maximum IBI	-0.04 s	-0.01 s	*0.28**

*: U-test, **: t-test.

weight given to mental task load accounted for 99% of the variance in the CG's deviation of diastolic BVP with $R^2 = 0.99^\dagger$, $p < 0.01$, $df = 22$, and 61% of the variance in the AR group's deviation of mean IBI with $R^2 = 0.61^\dagger$, $p = 0.03$, $df = 22$.

In summary, minimum physiological features typically deviated more from baseline than mean and maximum ratings. IBI and BVP deviations showed increasing stress in both groups with this being marginally higher in the AR group. A combination of correlation and regression analysis showed how higher skin temperature was partially accounted for by higher subjective joy-of-use in the CG. This suggested the arousal implied in this physiological signal was positively valenced in line with subjective reports (see section 4.3.1.2). Longer task durations in the control environment partially accounted for the CG's increasing stress. The weight given to mental task load influenced shorter mean IBI ratings in the AR group and shallower diastolic BVP in the CG.

4.3.2.2.2 Facial expression results

This section reports the statistically significant differences of facial expressions between the groups. The statistically significant correlations between the groups' facial expressions, task performance and subjective experience are also discussed. As described in Section 3.6, facial expressions lasting less than half a second were classified as MFEs. Facial expressions lasting longer than this threshold were classified as NFEs. Graphical depictions of the AUs discussed are shown in Table 2.4. Table 4.3 shows the statistically significant differences between the groups' facial expression features. The weight that the CG gave to NASA-TLX effort (see Fig. 4.7) accounted for 99% of the variance in their deviation of percentage of AU15 MFEs (AR: 0.05 %, CG: -0.01 %, $p = 0.32$, $df = 47$) with $R^2 = 0.99^\dagger$, $p = 0.03$, $df = 22$. In the CG, joy-of-use (Likert Statement 7) accounted for 73% of the variance in deviation of percentage of neutral NFEs with $R^2 = 0.73^\dagger$, $p = 0.03$, $df = 22$. The CG's response to Likert Statement 14 (see Appendix G) correlated to their deviation of percentage of AU26 NFEs with $\rho = -0.52$, $p = 0.01$ ($R^2 = 0.89^\dagger$, $p < 0.01$), $df = 22$. Twenty-one percent of variance in deviation of percentage of AU20 NFEs (AR: - 0.05%, CG: -0.02%, $p = 0.28$, $df = 47$) was accounted for by shorter task durations in the AR group with $R^2 = 0.21$, $p = 0.03$, $df = 22$.

In summary of the facial expression results, the AR group showed a significantly greater increase of AU12 NFEs (smiling) than the CG. Shorter task duration in the AR group correlated

Table 4.3. Significantly different facial expression features between the AR and paper-based CG, showing mean values and statistical significance.

Facial expression feature	AR mean	CG mean	Result
AU20 MFEs per minute during the task	2.3/min	3.1/min	*0.01*[*]
Deviation of percentage of AU20 MFEs	-0.05%	-0.02%	*0.05*[*]
Deviation of percentage of neutral MFEs	-0.3%	-0.1%	*0.02*[**]
Deviation of percentage of neutral NFEs	-0.3%	-0.1%	*0.02*[**]
Deviation of percentage of AU12 NFEs	0.1%	-0.1%	*0.01*[*]
Deviation of percentage of AU26 MFEs	0.02%	0.10%	*0.03*[*]
Deviation of percentage of AU26 NFEs	-0.05%	0.08%	*0.01*[*]

[*]: U-test, [**]: t-test.

to lower AU20 NFEs. In the CG, effort and joy-of-use correlated to percentage of AU15 MFEs and neutral NFEs respectively.

4.3.2.2.3 Head rotation frequency results

The literature reports that emotion is expressed in the frequency of head rotations [102], [103]. An eight second window of post experience OpenFace head pose data was analysed in the frequency domain using MATLAB™ to evaluate the participant's emotional state at task end. The results showed that the CG exhibited significantly higher amplitudes of the high-range frequencies (said to exclusively express anger emotion [102]) on each axis of head rotation (pitch: 41%, yaw: 35%, roll: 24%) during the 8 second post task sample, with $p < 0.01$, $df = 47$.

4.4 Summary

Study 1 presented a QoE evaluation and comparison of textual procedure assistance instructions presented in AR compared to a paper-based control. This evaluation used an optimal Rubik's Cube® solving procedure as a proof-of-concept for AR and paper-based procedure assistance using a text instruction format.

The AR group performed the Rubik's Cube® task significantly faster than the paper-based control group, and with fewer errors. This finding highlights the utility of AR for informational phase procedure assistance. Longer practice durations in the AR environment suggest AR may require more time for new users to familiarise themselves with this novel assistance medium. The AR group reported significantly more distraction and discomfort and less usefulness with the AR procedure assistance medium. This suggests that AR application design requires careful consideration of user comfort and distraction. Nevertheless, the AR medium was ranked more acceptable than the paper-based medium. The novelty of the AR medium may have had an influence on this result [171], [172].

Longer task durations in the paper-based control environment were seen to correlate to a physiological manifestation of increasing stress (reducing IBIs). The weight given to mental task load on the NASA-TLX questionnaire influenced IBI features in the AR group and BVP features in the CG. Deviation from baseline of systolic BVP was significantly different between the groups. The lower deviation of systolic BVP in the CG correlated to their increase in of percentage of AU26 NFEs (jaw drop, see Table 2.4). Many of these results and the lessons learned from Study 1 inform parts of the methodology of Study 2 that is detailed in the methodology subsection of Chapter 5.

Chapter 5

A QoE Evaluation of a Text-Only and a Combined Text and Interactive Animated 3D Model Instruction Format for AR Procedure Training.

This chapter describes the QoE evaluation of a combined text and animated interactive 3D model instruction format compared to a text-only control for AR-based procedure training. This evaluation is referred to in this chapter as Study 2. The motivation and aims of this study are given. The methodology is described in detail including the QoE metrics and recording instruments used, the experimental protocol, and the GoCube™ training procedure.

5.1 Motivation

AR is showing promise as a training platform [99], [174], with the literature encouraging further research in this area [19]. AR offers improved trainee learning by means of reduced cognitive load during training [44], [72], [175]. This could be achieved by interactive training in AR, which allows for customised training pace [37] and corrective feedback [61]. However, the benefits to learning offered by AR can be impacted by instruction format because of dependency formation [8] and extrinsic cognitive load [60]. Procedural instructions describe how to complete a procedure in a stepwise manner. Examples provide an analogous model showing exactly how a particular task is carried out; they may influence learning by creating trainee dependency. The effort required to carry out procedural instructions may benefit learning. A clear understanding of the influence of procedural and example instruction formats on AR trainee QoE is crucial to realise AR's potential as a procedure training platform. Researchers have called for the evaluation of the influence of training instruction formats on the AR trainee [44]. This provides the motivation for the evaluation of procedural and example training instruction formats within AR. Study 2 evaluates the influence of these instruction formats on the AR trainee's pragmatic and hedonic needs and expectations, motivated by the research sub-questions outlined in Chapter 1. Performing the GoCube™ manipulation procedure involves fine motor bimanual and visual coordination [145], [146] in common with the variety of disciplines in which optical see-through AR HMDs are expected to be adopted for training [31], [32], [39]. This includes object identification, inspection, alignment, adjustment and orientation manipulations, combined with visual comparison and verification [13].

The results of Study 1 have raised some open questions about the influence of the text-based instruction format on reported distraction and lack of hedonic expectation fulfilment. The aim of this study is to evaluate the influence of a combined text and interactive animated 3D model (example) instruction format compared to a text-only (procedural) instruction format, on AR trainee QoE. This includes the influence of Rubik's Cube manipulation on positive transfer to general mental rotation abilities [145]. This aim was supported by the development of a test methodology that incorporated the capture of physiological ratings, facial expressions, eye gaze, mental rotation abilities and self-reported affect, task load, cognitive load and QoE. The AR training application included instruction execution verification. Instruction position was a design consideration in Study 2 towards reducing perceived distraction. In Study 2, the instructions are anchored in the same fixed position in the environment for each participant using environment scanning sensors on the HL2. This afforded the participant the opportunity to focus their attention on the workpiece only as required. This may help to shed light on the influence of these instructions on distraction and expectation fulfilment.

5.2 Methodology

This section describes the methodology that was employed to carry out Study 2. It includes an overview of the task that the participants undertook, the instruction modalities that were involved and the experimental protocol. The protocol is largely the same as in Study 1 as described in Chapter 4, however the test phase consists of four sub-phases, which are training, waiting, recall and transfer [176], [177].

The main test group (referred to hereafter as the TG) experienced an animated 3D Cube model instruction format combined with text as recommended in [8] to aid comprehension. The control group (referred to hereafter as the CG) experienced text-only instructions in AR. Both groups were trained in an AR-based GoCube™ training procedure using the HL2 AR HMD (see Table 2.1). The literature informed the waiting phase duration of thirty seconds used in the protocol [177]. Participant learning was evaluated in a post-training recall phase inspired by [55]. Transfer was evaluated in a comparison of pre- and post-training mental rotation abilities using the standard Vandenberg mental rotation test [153].

Study 2 recorded the participant’s physiological signals, facial expressions, subjective affect, task load and elements of QoE. In Study 2, the E4 was fitted immediately after written consent was provided by the participant. This was done at this stage in Study 2 to allow the maximum amount of time for the E4’s heat flux sensor to acclimatise to the participant’s skin temperature. In addition to the metrics recorded in Study 1, the participant’s eye gaze was recorded using the HL2’s eye tracking sensors. In training, the fulfilment of the trainee’s pragmatic needs were concerned with learning and transfer, which were evaluated in post-training recall and Vandenberg rotations. The instruction formats may also influence the trainee’s hedonic needs and expectations by affecting the usability and interaction quality [5] as seen in Study 1.

In Study 1, correlation analysis across multiple questionnaires corroborated consistent subjective reporting of aspects of QoE. The same set of questionnaires is used in Study 2 for the same reason, and to facilitate a longitudinal study of AR pre-experience and user expectations. Consideration of minimum and maximum physiological ratings in addition to mean values were shown to have utility in the form of significant correlations to subjective experience in Study 1. Study 2 continues to consider minimum, mean and maximum ratings. In Study 1, systolic (peak) and diastolic (trough) BVP amplitudes were considered, inspired by [120]. In Study 2, mean BVP (peak minus trough) is also considered. As part of Study 1, per-minute AU features during baseline and task were created. In addition to this, deviation from baseline of per-minute and percentage AU features were also calculated. Where the makeup of total facial expression is expressed in terms of percentage of various AUs, the increase of one AU occurs in conjunction with a decrease in another. This nonmonotonic nature of AUs normalised on a percentage basis makes interpretation of such facial expression results difficult, as they may be as easily due to a reduction in one facial expression or an increase in another at the same time. Therefore, in Study 2, MFEs and NFEs are normalised on a per minute basis only to create monotonic features, whose results can be interpreted without ambiguity.

The higher distraction reported in the AR environment in Study 1 may have been partially due to the AR instruction position in the user's FOV, and partially due to the HMD. In Study 2, instruction position is considered more carefully to control for this influencing factor. The instructions are not fixed in the participant's view, affording them the opportunity to focus unhindered on the workpiece as needed. The more ergonomic Microsoft™ HoloLens 2™ HMD is used to aid in participant comfort. A gender balanced sample is maintained in Study 2 to continue to evaluate the influence of this static human QoE influencing factor[16], [178]. An electronic version of the Rubik's Cube®, (the GoCube™) was used in Study 2 for robust Cube state tracking. This is required during the psychomotor phase [34] to provide corrective instructions during AR training. The GoCube™ is a network enabled version of the Rubik's Cube®, permitting communication of Cube state to the AR headset. In Study 1 only a subset of lower facial AUs exclusive to certain emotions were used. Classification of facial expressions into emotions can be inaccurate if done out of context [105]. In Study 2 the full set of AUs from Table 2.4 is used without classifying them into representations of certain emotions.

5.2.1 The evaluation task

During the training phase, the participants were instructed in a 14-step [179] GoCube™ manipulation procedure. Training was self-paced [177] and the participant could undergo as many training cycles as they required to learn the GoCube™ manipulation procedure. Each training cycle consisted of two halves, where the participant was required to action a set of 7 instructions. Each training cycle began with the Cube in the solved state. The second set of 7 instructions was the reverse of the first set of 7 instructions, returning the Cube to its solved state by the end of each training cycle. The TG received training instruction using a combined text and interactive animated 3D model of the GoCube™ as shown in Fig. 5.1. The CG only had the benefit of the text instruction from Fig. 5.1.

Task performance was evaluated over three phases, namely, the training, recall and transfer phases. The efficacy of training in this study was measured in terms of the training itself, in the participant's learning and in the transfer to general mental rotation abilities. The training was measured in terms of quantity of training cycles, quantity of errors, instruction response times and overall duration. Learning was measured during a recall where the participant had to

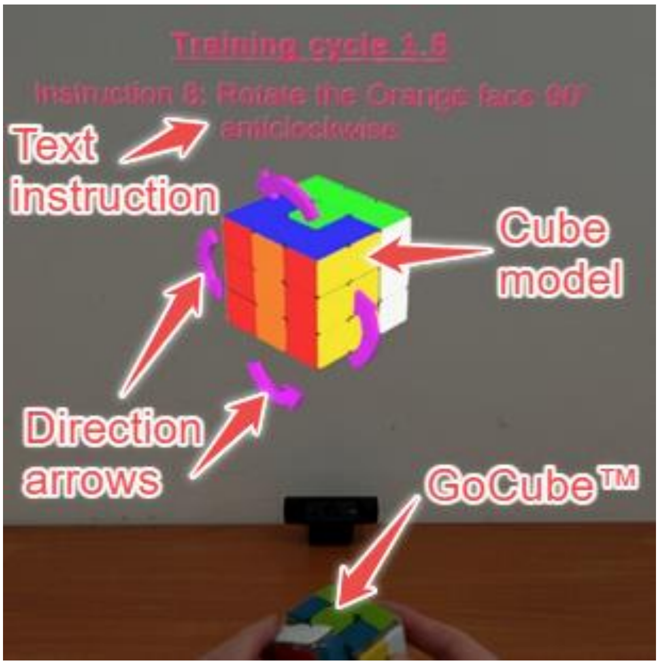


Fig. 5.1. The combined text and interactive animated 3D Cube model, with desk mounted video camera and GoCube™.

perform the procedure as trained from memory. Recall phase performance was recorded in terms of duration, Cube face rotation durations, and accuracy. Transfer was measured in the difference between pre- and post-training Vandenberg test results. These metrics were all measured in real-time as relayed from the GoCube™ over a wireless network to the HL2 as described in the next section.

5.2.2 The metrics captured during Study 2

The following subsections detail the different metrics captured during Study 2. This includes various task performance metrics, implicit metrics in the form of physiological ratings, facial expressions and eye gaze features, and explicit questionnaire responses.

5.2.2.1 Task performance

Task performance metrics consisted of interaction times, error rates and phase durations. Learning was evaluated during a post-training recall phase [56], [176]. During the recall phase, the participant was required to perform the GoCube™ manipulation procedure from memory.

Recall Cube face rotation durations (the time taken to rotate the GoCube™ faces), recall duration (the time it took for the participant to perform the procedure from memory) and recall accuracy (how many mistakes the participant made while performing the procedure from memory) were recorded on the HL2 in real-time.

The recall phase only commenced after a 30-second [177] post-training waiting period. During the waiting period, the participant was required to correctly answer as many arithmetic questions as possible from a list of 10 questions taken directly or adapted from [180] (see Appendix H). This was done to engage their working memory to ensure that the learned material had been schematised to LTM and not temporarily held in WM by means of focus or repetition [15]. If the test subject could not recall the GoCube™ manipulation procedure as trained, then the knowledge was not being recalled from LTM or had been lost from WM, in which case it would never be correctly encoded to LTM.

The influence of the different instruction formats on transfer was evaluated in a post-training Vandenberg rotation test for comparison to mental rotation baselines. The Vandenberg test provides a standard way to evaluate the mental rotations that are involved in rotating the faces of the Rubik's Cube® [145]. The participants mental rotation abilities were recorded for pre- and post-training comparison.

5.2.2.2 Implicit metric capture

The methodology of Study 2 provided for the capture of eye gaze features, physiological ratings and presence of lower facial AUs. Each participant was seated at a table where they were fitted with the HL2. The HL2 eye-tracking sensors were calibrated to each participant's eyes using the eye calibration protocol bundled with the HL2. Eye gaze is intended as an input medium for the HL2, in conjunction with hand tracking and voice commands, in the absence of traditional mouse or keyboard input. The HL2 SDK code was adapted to record eye gaze in HL2 RAM in real-time. This data was then written to memory at the end of the experience so as not to impact the HL2 performance during the evaluation.

The Empatica E4 sensor [129] was used to record the participant's skin temperature, BVP, IBI and EDA in common with Study 1. In Study 2, an update to the E4 firmware had included the calculation of heart rate (HR) in the E4 signal. A desk-mounted video-camera was used in conjunction with OpenFace facial recognition software [118] to record the participant's lower facial AUs [104] (as per Study 1). Recording of these implicit QoE metrics continued from baseline until the end of the recall sub-phase. Deviation from baseline of these implicit QoE metrics was considered to be indicative of the influence of the different instruction formats on user QoE.

5.2.2.3 Explicit metric capture

In post-experience questionnaires, the participant was first asked to use an emotion of their choice to describe their post-experience emotional state (see Appendix I part 1). The participant then completed the SAM questionnaire (see Appendix I part 2). They were then asked to select a label from the 2D emotion space [79] that best described their post-experience emotional state (see Appendix I part 3). Correlates were sought across the three methods of subjective affect reporting to establish consistency in usage of emotion terms by each participant. Usage of conflicting emotion terms across the questionnaires in Appendix I part 1 and part 3 (which were completed seconds apart) would indicate a lack of consistent meaning to the participants that used them. Correlates were also sought between the emotion terms used by the participants, and physiological ratings and facial expressions of emotion, subjective experience and objective task performance metrics. The participant then reported their subjective QoE, cognitive load and task load, using the Likert scale questionnaire (see Appendix J) and NASA-TLX (see Appendix A), respectively.

In the Likert scale questionnaire, aspects of interaction were included in Statements 1, 2, 3, in terms of confidence, comfort and frustration. Usability and interaction were covered by Likert Statements 4, 6, 7 and 8 in terms of joy-of-experience, distraction and stress. Aesthetics were covered in Likert Statements 5 in terms of user interest. Finally, Likert Statement 9 and 10 were designed to capture acceptability. In Study 2, the Likert scale questionnaire also incorporated relevant elements of the Paas [181] and Leppink [101] cognitive load questionnaires. The Paas questionnaire measures total cognitive load, while the Leppink questionnaire measures intrinsic, extraneous and germane cognitive loads independently.

5.2.3 Experimental protocol

The experimental protocol of Study 2 is largely similar to that of Study 1 as outlined in Chapter 4. This section describes the differences that are specific to Study 2, including the test phase, which consists of 4 sub-phases. In addition to the metrics recorded in Study 1, Study 2 recorded eye gaze and subjective cognitive load. In Study 2, the experiment lasted for 40 minutes on average.

5.2.3.1 Phase 1: Sampling and Information Sharing

Convenience sampling resulted in a sample size of 60 test participants [95]. The sample group had an age range from 19 to 62 years old with a mean age of 32. Participants were assigned the TG or CG based on their gender, with an equal distribution of 15 males and 15 females in each group. Each participant was provided with an information sheet explaining the study in full as per Appendix K. Every participant completed and signed a consent form as per Appendix L. This information sharing phase lasted 2 minutes on average, and the signing of the consent form took just over 1 minute and 30 seconds on average across all participants. The end of the consent form included two post signature questions to record interest and expectations (see Appendix L). After giving written consent, participants were fitted with the Empatica E4 sensor [129]. This was done at this stage to allow the maximum time for the temperature heat flux sensor to acclimatise to the participant’s skin temperature. The E4 began recording physiological ratings at this time. Fitting and commencement of recording of physiological ratings took 30 seconds on average, leading to a total of 4 minutes for the sampling and information sharing phase. The participant then proceeded to the screening phase.

5.2.3.2 Phase 2: Screening

The Snellen eyesight test (see Fig. 3.1.a) and the Ishihara colour blind test (see Fig. 3.1.b) lasted 1 minute and 30 seconds each. The interactive digital Vandenberg-based mental rotation test (see Fig. 3.2) was implemented for 1 minute. Twelve participants assigned to the main test group did not have 20/20 vision compared to 19 who were assigned to the control group. Seven of the participants assigned to the main test group indicated varying degrees of red-green colour blindness by not correctly answering all plates, compared to 3 who were assigned to the control group. No participants were excluded during screening in line with ITU-T P.913 recommendations [94]. Participants were not prohibited from wearing their prescription glasses during the test. The screening phase took an average of 4 minutes.

5.2.3.3 Phase 3: Instruction and HL2 eye calibration

Each participant was introduced to the GoCube™ in terms of face colours and face rotation directions. Their understanding of this information was verified using a standard Rubik's Cube®. They were then verbally instructed how to perform the remaining phases as detailed in the following sections. They were fitted with the HL2, which was calibrated to their eyes. This instruction and calibration took a further 4 minutes. The baseline phase then began.

5.2.3.4 Phase 4: Baseline

The start of the 5-minute baseline period was marked by the beginning of recording of eye gaze features using the HL2's eye tracking sensors. The recording of these implicit QoE metrics continued throughout the evaluation and only ceased after the recall phase was complete.

5.2.3.5 Phase 5: Practice

The participant underwent an automated practice phase using their given instruction format. This involved carrying out instructions for rotating each GoCube™ face 90^{o} in both clockwise and anti-clockwise directions (i.e., 12 instructions). It had already been verified that they could do this independently of the HL2 during the instruction and calibration phase (Phase 3). Now the goal was to verify that the participant could see, understand and correctly follow instruction from the HL2. Development of the AR training application included corrective instructions that were issued in the event of trainee mistakes. Upon successful completion of all instructions, the participant automatically progressed to the test phase in which they were trained in a specific GoCube™ manipulation procedure. The average practice phase lasted 69 seconds for the TG and 47 seconds for the CG. Fig. 5.2 shows the test set-up as the participant begins the practice phase.

5.2.3.6 Phase 6 a: Testing: The training sub-phase

The training sub-phase began with the GoCube™ in the solved state. Each training cycle consisted of two halves, where the participant was required to action a set of 14 instructions that began and ended each training cycle with the GoCube™ in the solved state. This was done to standardise the training procedure across all test subjects. The number of training cycles,

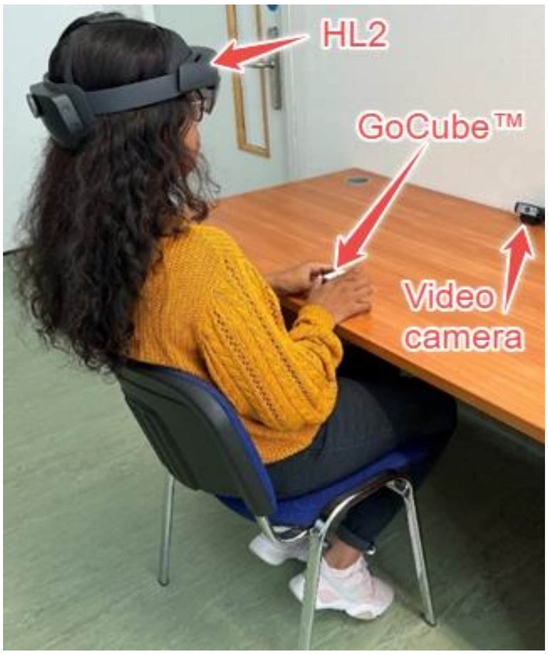


Fig. 5.2. Study 2 test set-up showing a participant wearing the HL2, holding the GoCube™ with the 1080p Logitech desk mounted camera.

total training time and number of errors were all automatically recorded on the HL2 as objective metrics of the influence of the instruction formats on training. The TG required an average training duration of 4 minutes, the CG required an average training duration of 4 minutes and 30 seconds. The participant alerted the researcher once they were confident that they had learned the procedure as trained. The researcher then ended the training phase by means of remote input transmitted over the wireless network to the HL2.

5.2.3.7 Phase 6 b: Testing: The waiting sub-phase

A minimum of 20 seconds of stimulus-free (i.e., the GoCube™) waiting is sufficient to ensure that learned information has either been schematised into LTM or retained in WM by means of repetition [15]. If after 20 seconds, the participant cannot perform the task, the information has either not been learned or has been lost from WM, in which case it will not be learned. The participant waited for a 30 second interval as inspired by [177], during which they performed arithmetic questions taken directly from, or inspired by, [180]. Performing these equations correctly requires WM resources and any training not schematised to LTM will likely be lost during this process. This phase lasted for 30 seconds.

5.2.3.8 Phase 6 c: Testing: The recall sub-phase

In the recall phase, the participant had to reproduce the GoCube™ manipulation procedure as trained. Accuracy, number of errors, Cube face rotation intervals and total recall duration were the objective performance metrics of recall. Recall phase duration was not limited. The TG required an average recall phase duration of 46 seconds to perform the GoCube™ manipulation procedure from memory, compared to 30 seconds for the CG. This duration difference is discussed in detail in the results and discussion section.

5.2.3.9 Phase 6d: Testing: The transfer sub-phase

The influence of the different instruction formats on near transfer was evaluated in a post-training Vandenberg rotation test for comparison to mental rotation baselines. The Vandenberg test provided a convenient standardised means to evaluate near transfer of the mental rotation abilities that were involved in manipulating the Rubik's Cube® [145]. The time allocated to this test was 1 minute.

5.2.3.10 Phase 7: Questionnaires

As part of this explicit measures phase, the participant was first asked to write down an emotion that best described their emotional state (see Appendix I part 1). They were then asked to complete the SAM questionnaire (see Appendix I part 2). They were then asked to select one emotion label from Russel's 2D emotion space taken from [85] (see Appendix I part 3). They then answered the ten-statement five-point Likert scale (see Appendix J). In addition to these statements, there were also three cognitive load questions, one each specific to intrinsic, extrinsic and germane cognitive load during the training phase [101] (see Appendix J, Statements 11 - 13). There were also two cognitive load questions taken from [181] on this questionnaire (see Appendix J, Statements 14 and 15). One was to subjectively evaluate the amount of cognitive effort during the training phase and one for the recall phase. These cognitive load statements were presented on a nine-point scale. The participants then completed the NASA-TLX questionnaire (see Appendix B). In total, the questionnaire phase took just under 6 minutes to complete on average.

5.3 Results and discussion

This section presents and discusses the subjective and objective results. Like Study 1, subjective results consist of affect questionnaires, a Likert scale questionnaire and a task load questionnaire. In Study 2, the Likert scale questionnaire contained statements specific to cognitive load. Objective results consist of task performance and implicit results. Task performance consists of baseline, practice, training, waiting, recall and transfer results. Implicit results consist of physiological ratings, facial expressions and eye gaze results. Subjective results are presented first to facilitate a discussion of correlation results in the later task performance and implicit results sections.

5.3.1 Subjective results from Study 2

This section reports the results of the questionnaires that were completed after giving written consent (see Appendix L) and after the recall phase (see Appendix I, J and B). After the recall phase, the participants firstly answered the combined open-ended emotion label, SAM questionnaire and 2D emotion space questionnaire. This was followed by the Likert scale questionnaire. Finally, the participants answered the NASA-TLX task load questionnaire.

5.3.1.1.1 Pre-experience, interest and expectations

There were no statistically significant differences between the groups for pre-experience in AR (including HMD AR), interest in AR or expectations of joy-of-experience in AR. However, 63% of extraneous cognitive load reported by female participants of the CG (AR male: , CG male: , AR female: , CG female:) was accounted for by variance in pre-experience with AR HMDs. In females of the TG, pre-experience interest correlated moderately to acceptability (see section 5.3.1.3) with ρ = 0.53, *p* = 0.04, *df* = 13, while positive expectations of joy-of-experience correlated to acceptability with ρ = 0.58, *p* = 0.02, *df* = 13.

5.3.1.2 Post experience emotional state

As part of this explicit measures phase, the participant was first asked to write down a term that they felt best reflected their post-experience emotional state (see Appendix I part 1). Fig. 5.3. shows that ‘Happy’ was the most common open-ended emotion term used, being chosen

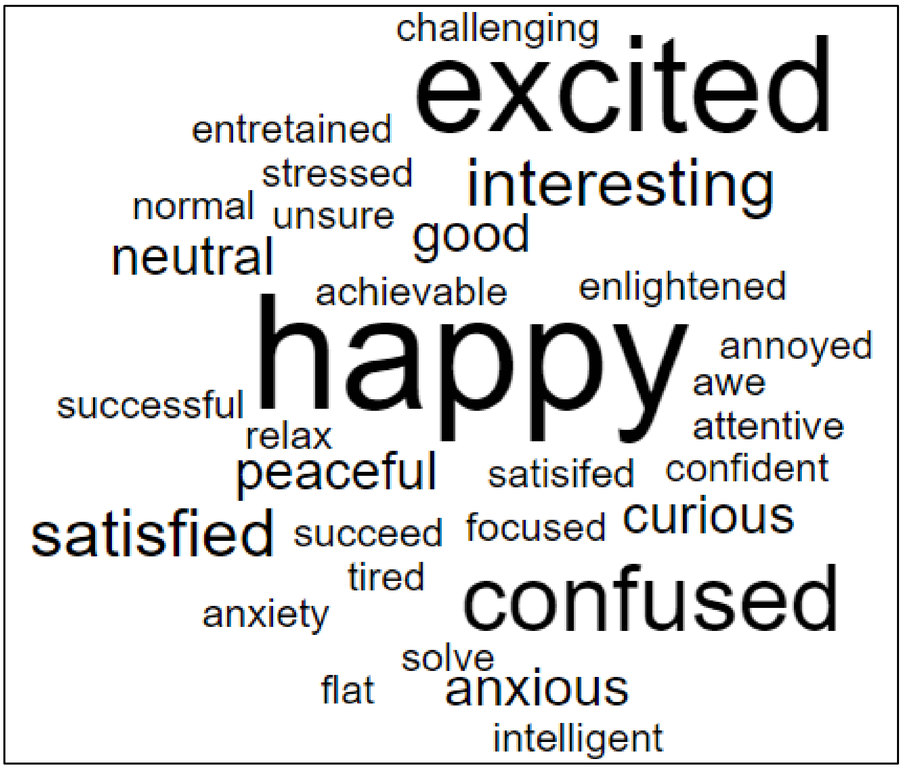


Fig. 5.3. A word cloud of open-ended emotion terms used by the participants of both groups of Study 2. Frequency of term usage is represented by the size of the term.

by 17% of the sample. This was followed by the term 'excited' as the second most used term, being chosen by 13% of the sample. Ten percent of the open-ended terms were not regarded as emotion terms, perhaps being chosen due to language barriers. The full set of open-ended terms used by the participants is shown in Fig. 5.3. A pie-chart showing the percentages of terms used in available in Appendix M for the interested reader.

The participants were then required to complete the SAM affect questionnaire (see Appendix I part 2). The SAM questionnaire responses are shown in Fig. 5.4. There were no statistically significant differences between the groups for valence (p = 0.64), arousal (p = 0.96) or dominance (p = 0.96), df = 58, respectively. When SAM arousal (female: 1.3, male: 1.0, p = 0.36), valence (female: 2.9, male: 2.4, p = 0.11) and dominance (female: 1.3, male: 1.6, p = 0.56) results were combined into ordinal values, they correlated to gender with ρ = 0.51, p < 0.01 (R^2 = 0.35††, p < 0.01), df = 58. SAM valence correlated to the rank given to NASA-TLX frustration across both test groups with ρ = -0.50, p < 0.01 (R^2 = 0.17†, p < 0.04), df = 58, while SAM dominance corelated to the rank given to performance and overall task load with ρ = -

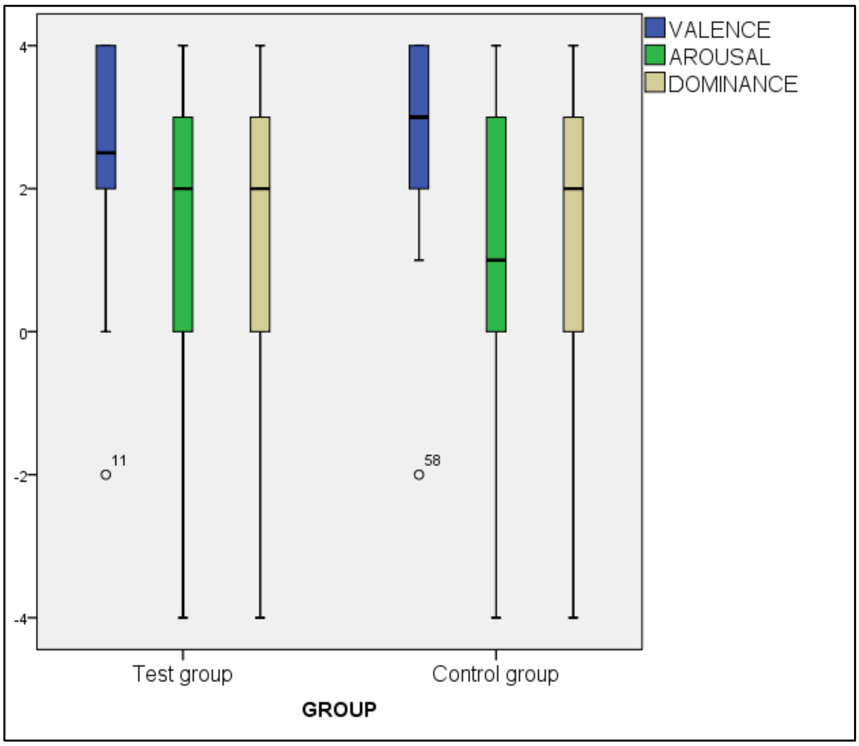


Fig. 5.4. A box plot of the SAM questionnaire responses for the test group and the control group.

0.68, $p < 0.01$ ($R^2 = 0.52^\dagger$, $p < 0.01$) and ρ = -0.59, $p < 0.01$ ($R^2 = 0.99^\dagger$, †, $p < 0.01$), $df = 58$, respectively. This might suggest that greater utility can be derived from consideration of emotion in terms of its valence, arousal and dominance dimensions rather than using labels that may not be understood.

The participants were then asked to choose an emotion term from the 2D emotion space [85] (see Fig. 5.5. or Appendix I part 3). From the 2D emotion space shown in Fig. 5.5, ‘interested’ was the most chosen emotion label being chosen by 13% of the sample, with ‘happy’ in second place chosen by 8%. Once presented with the labels available in the 2D space, only 35% of the sample whose open-ended label did appear in the 2D space stayed with their original choice. Forty percent of the open-ended emotion labels chosen did not appear in the 2D space. There were no statistically significant differences between the groups for open-ended terms ($p = 0.83$) or 2D space terms ($p = 0.99$). The 2D emotion terms chosen by the participants of Study 2 are shown in the word cloud in Fig. 5.6. A pie chart including the percentage of usage of the terms is included in Appendix N. There were no significant correlations between emotion terms to any of the other metrics captured during Study 2, including to SAM questionnaire responses.

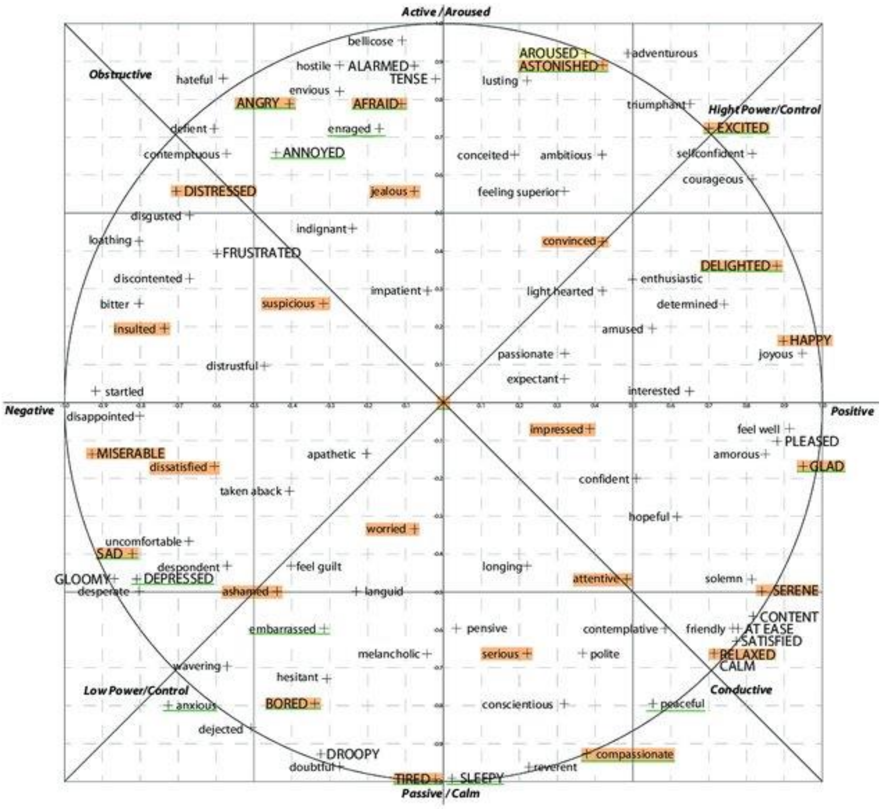


Fig. 5.5. The 2D emotion space as used in Study 2 [85].

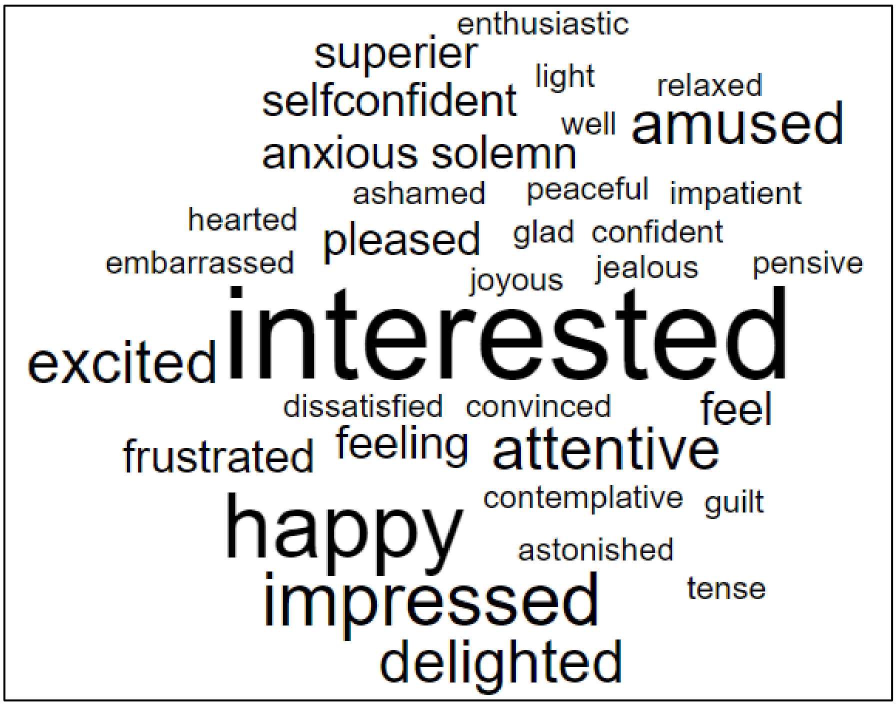


Fig. 5.6. A word cloud of the emotion terms chosen from the 2D emotion space by the participants of both groups of Study 2. Frequency of term usage is represented by the size of the term.

5.3.1.3 Likert scale questionnaire

Figure 5.7 shows the Likert scale results, including the adjectives associated with Statements 1-10, covering confidence, comfort, frustration, joy-of-experience, interest, distraction, stress and acceptability. Likert Statements 11-13 were one question each relating to intrinsic, extrinsic and germane cognitive load respectively, on a nine-point scale. Likert Statements 14 and 15 related to overall cognitive effort invested during training and recall respectively (Appendix J) also reported on a nine-point scale.

There were no statistically significant differences on the Likert scale questionnaire between the groups as shown in Fig. 5.7. There were also no statistically significant differences between the genders. In Study 1, five-point Likert Scale questionnaire MOS results were presented ranging to 1 (strong disagreement) to 5 (strong agreement). In Study 2, Likert Scale MOS results

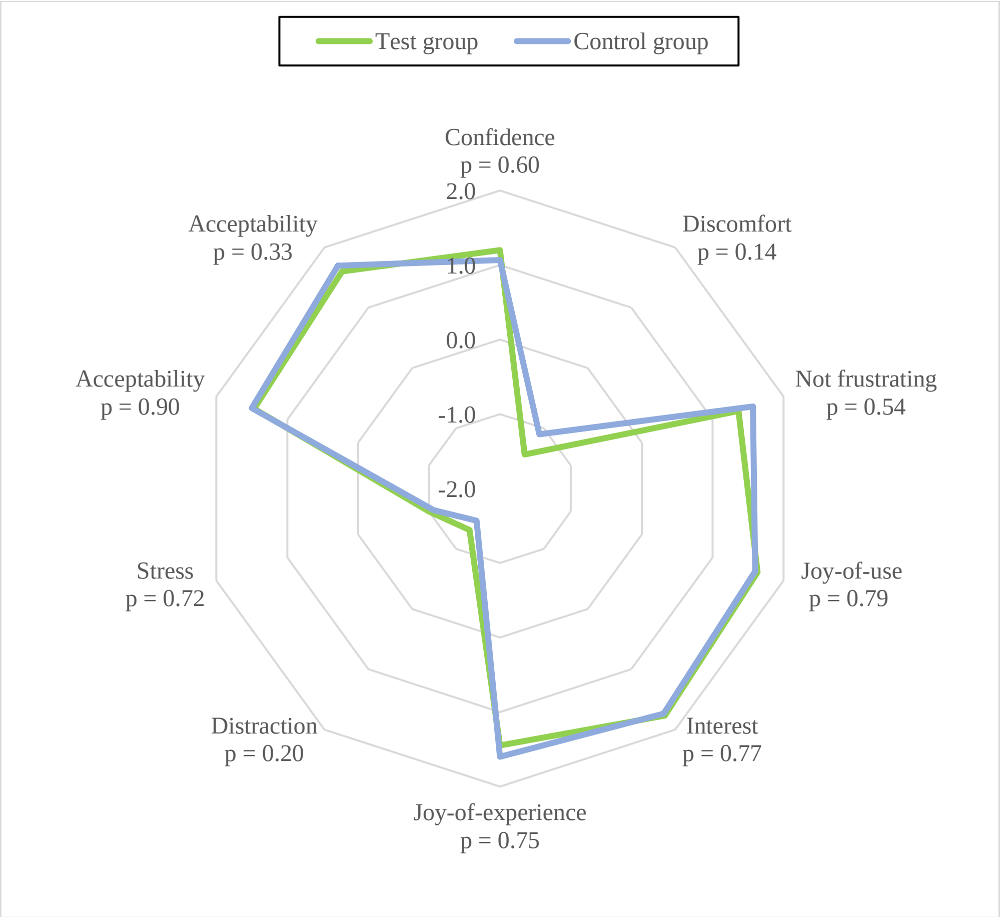


Fig. 5.7. A radar graph showing the Study 2 Likert scale adjectives related to the questionnaire statements 1-10 and the U-test statistical significance between the main test group and the control group.

are presented as positive and negative values centred around the neutral score of zero to better reflect agreement or disagreement with the statements. In this way it is easier to see in Fig. 5.7 (and in Appendix O) that the TG agreed more strongly than the CG in confidence (Statement 1), joy-of-experience (Statement 4) and interest (Statement 5). The CG agreed stronger than the TG for frustration (Statement 3), joy-of-experience (Statement 6, posed to verify conscientious responses in relation to Statement 4 resulting in a correlation of ρ = 0.62, see Fig. 5.7.) and acceptability (Statements 9 and 10). The TG disagreed stronger than the CG on discomfort (Statement 2). The CG disagreed stronger than the TG on distraction (Statement 7), and stress (Statement 8).

Fig. 5.8 and Fig. 5.9 show the results of the Likert scale questionnaire statements relating to cognitive load as analysed on the basis of gender. The full table of Likert scale responses for Study 2 is given in Appendix O. Fig. 5.8 shows that the female participants in the TG reported more cognitive effort invested during training than females in the CG. Fig. 5.9. shows that males in the CG reported more cognitive effort invested during recall than their female counterparts.

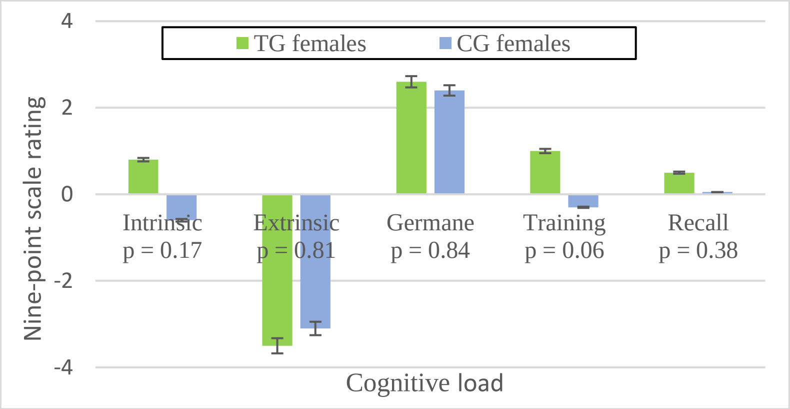


Fig. 5.8. Subjective cognitive load reported by CG and TG females including U-test results.

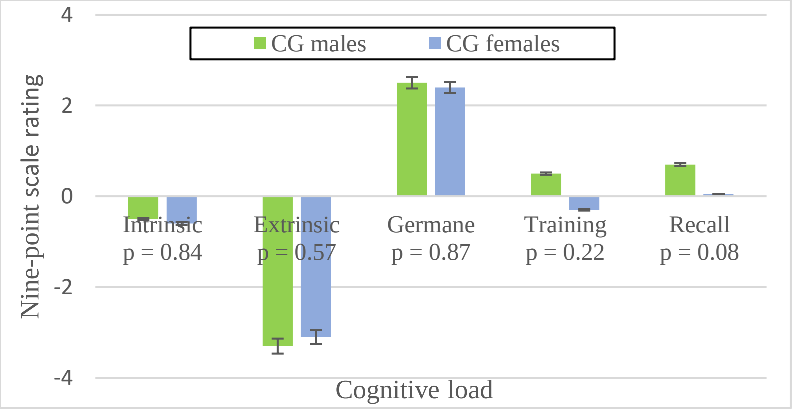


Fig. 5.9. Subjective cognitive load reported by CG males and females in U-test results.

5.3.1.4 NASA-TLX task load

Fig. 5.10. shows that the TG reported a statistically significant higher perception of performance than the CG with p = 0.03, df = 59. This was contributed to most significantly by females as seen in Fig. 5.11. Perhaps this is because males in the TG perceived more mental demand than females in the TG with, p = 0.07, df = 29. This heightened sense of performance in the TG correlated to the deviation from baseline to training of their minimum skin temperature with ρ = -0.59, p = 0.01 (R^2 = 0.97†, p < 0.01), df = 28.

In summary of the subjective results, there were no statistically significant differences in pre-experience interest or expectations which were recorded after the participant's gave written consent. However, in female of the TG, acceptability correlated moderately to pre-experience interest and positive expectations of joy-of-use. There were no statistically significant differences between the open-ended of 2D emotion space terms used by the participants of either test group. There was also a noticeable absence of correlations between these open-ended and 2D emotion space terms. Furthermore, there was also a noticeable absence of correlations between these emotion terms and any of the other metrics recorded during this Study 2.

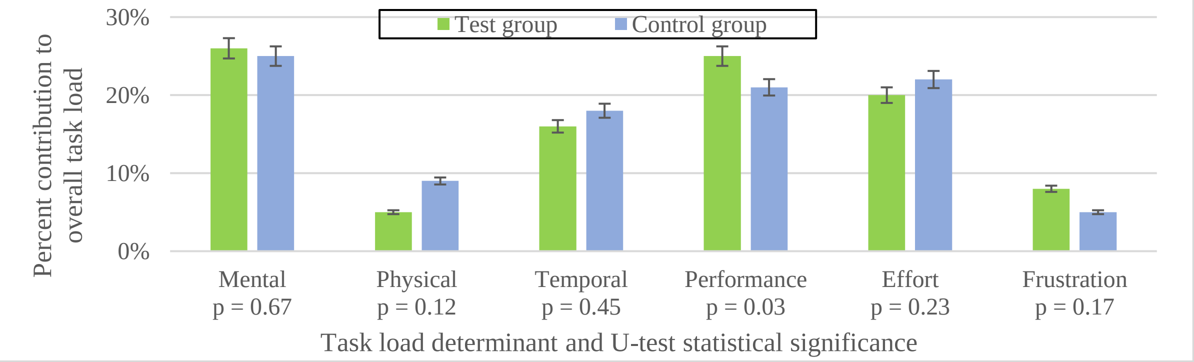


Fig. 5.10. NASA-TLX determinants and U-test significance for the main test group and the control group.

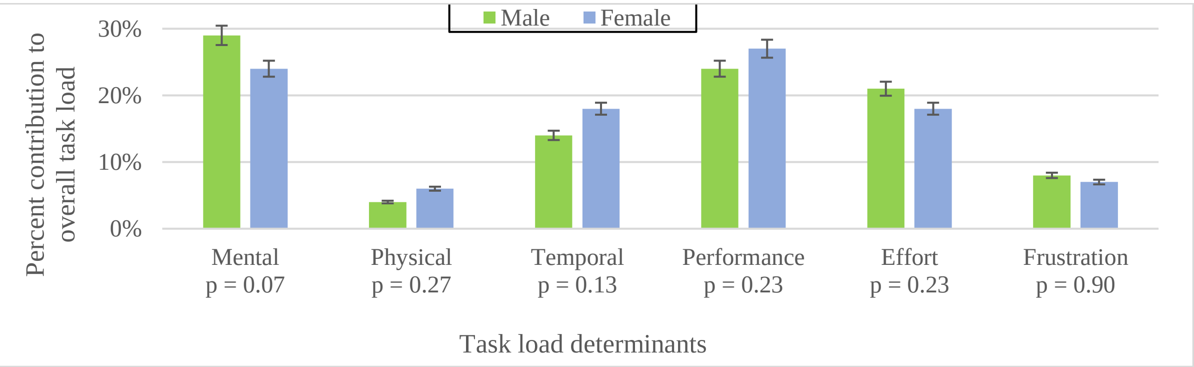


Fig. 5.11. NASA-TLX determinants and U-test significance for male and female participants.

There were no statistically significant differences in the SAM questionnaire responses for valence, arousal or dominance but SAM valence correlated to elements of frustration while SAM dominance corelated to elements of performance and overall task load across both test groups. Ordinal SAM results (arousal, valence and dominance combined) correlated significantly to gender. TG females reported more cognitive effort invested during training than CG females. CG males reported more cognitive effort invested during recall than CG females. Female participants reported less mental load than males on the NASA-TLX questionnaire.

5.3.2 Objective and implicit results from Study 2

The objective results of Study 2 are baseline Vandenberg performance, error rates and durations of the practice phase and test sub-phases. This is followed by a discussion of the physiological and physical implicit metrics including facial expressions and eye gaze results.

5.3.2.1 Task performance results

Task performance results are discussed for the following protocol phases; baseline of mental rotation abilities are reported first, followed by performance during the initial practice phase. This is followed by duration and error rates during the training sub-phase. Arithmetic performance during the waiting phase is included, followed by recall performance and the difference between pre- and post-Vandenberg mental rotations during the transfer phase.

5.3.2.1.1 Baseline phase results

Overall, the two independent groups were closely matched in terms of Vandenberg rotation abilities, with p = 0.94, df = 59. However, when investigated on the basis of gender, Table 5.1 shows that male participants of the CG got a statistically significant number of correct rotations compared to their female counterparts. Statistically significant differences in mental rotation abilities are reported in the literature, where males tend to get more correct rotations and females tend to get less incorrect rotations [182]. This is said to be because females spend more time than males verifying correctly matching stimuli [183]. Results of Study 2 corroborate this belief as shown in Fig. 5.12. Even though the female participants got fewer correct rotations (male: 244, female: 191) they also got fewer wrong rotations (male: 37, female: 32).

Table 5.1. Mean baseline mental rotation scores for correct, incorrect and total rotations by males and females of the main test group (TG) and control group (CG).

	Gender	Correct	Incorrect	Total
TG	**Male**	8	1	9
	Female	7	1	8
	Result	*0.40***	*0.51**	*0.34***
CG	**Male**	9	1	10
	Female	6	1	7
	Result	*0.05**	*0.50**	*0.50**

*: U-test, **: t-test.

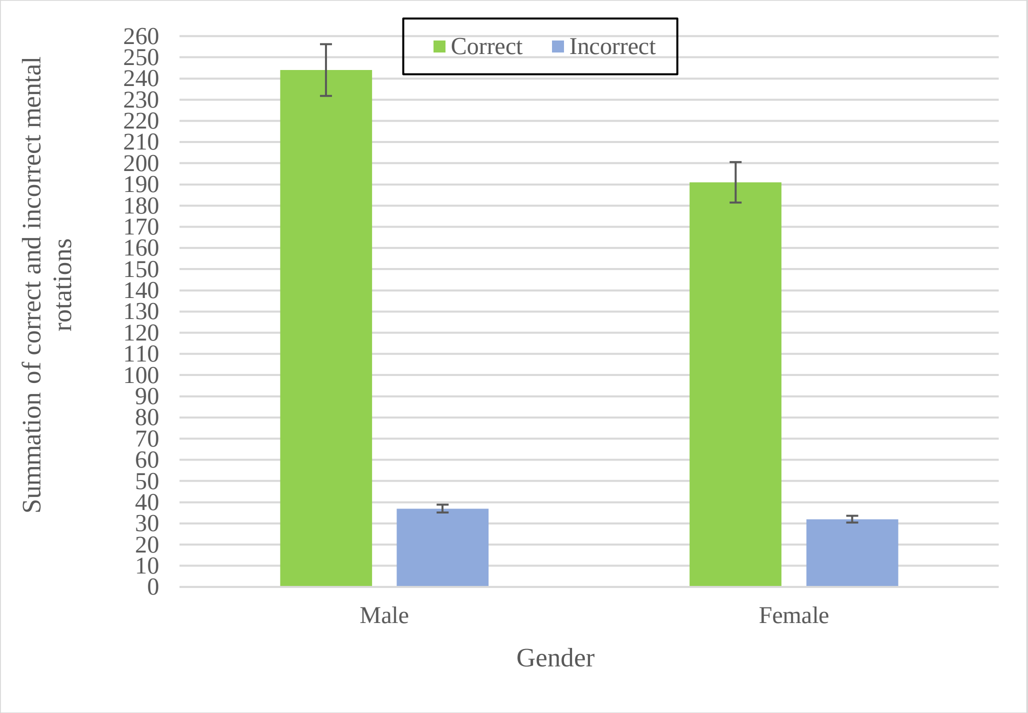


Fig. 5.12. Bar chart of correct and incorrect baseline mental rotation by male and female participant.

5.3.2.1.2 Practice phase

The number of practice phase mistakes per group is shown in Fig. 5.13 with the TG making significantly more mistakes than the CG. During the practice phase, 30% of participants in the TG made at least one mistake (with a maximum of 7 errors) compared to 7% of the CG, with (with a maximum of 5 mistakes) $p < 0.01$, $df = 59$. *The* average practice phase duration was 1 minute and 9 seconds for the TG, and 47 seconds for the CG, with $p < 0.01$, $df = 59$. Practice

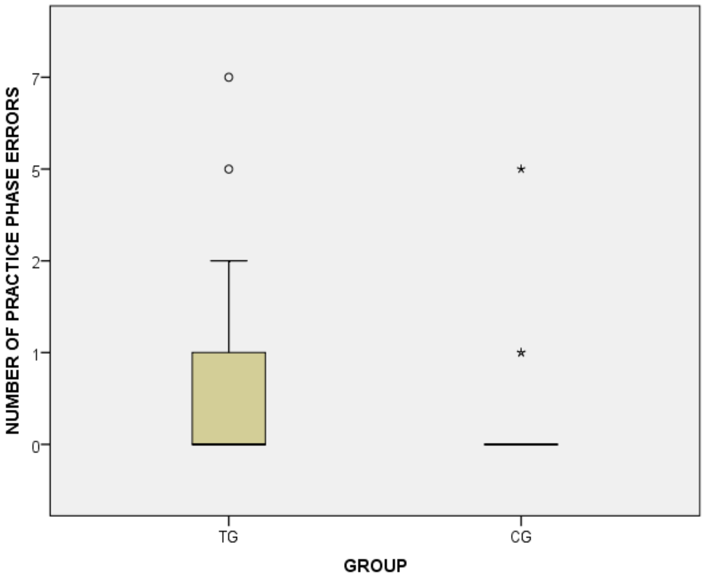


Fig. 5.13. A Boxplot showing the number of mistakes made by participants of the main test group (TG) and text-only control group (CG) during the initial practice phase.

duration correlated significantly to practice instruction response times with $r = 0.88$, $p < 0.01$ ($R^2 = 0.77$, $p < 0.01$), $df = 58$. The mean practice instruction response time was 5.2 seconds in the TG and 4.1 seconds in the CG, with $p = 0.01$, $df = 59$. It seems that less information presented in the text-only instruction caused less confusion resulting in less mistakes and faster practice.

It took the TG participants significantly longer to perform the practice phase instructions. TG participants had more information to look at in the combined text and model instruction format. During the practice phase, the TG participants spent additional time watching the Cube model animate in the TG environment which took longer than it took the CG participants to read the text-only instruction in the control environment (see practice phase eye gaze result in the implicit metrics section which follows). The TG participants' manipulation of the GoCube™ was unhindered by the speed of the animation. They could pre-empt the animation to skip the animation by carrying out the instruction on the GoCube™ before the animation completed. In the CG, mean practice instruction response times of 3.6 seconds for males and 4.7 seconds for females were significantly different, with $p = 0.05$, $df = 29$. CG practice durations of 39 seconds for males and 56 seconds for females were in turn significantly different with $p = 0.03$, $df = 29$.

5.3.2.1.3 Test phase: Training

A statistically significant difference was seen between training instruction response times of 4.6 s in the TG and 3.9 s in the CG, with p = 0.05, df = 59. Thirty three percent of the variance in training duration in males in the TG was accounted for by extrinsic cognitive load with R^2 = 0.76†, p = 0.05, df = 13. It seems that the greater amount of information being presented in the main test group caused more cognitive load in male participants resulting in slower training.

5.3.2.1.4 Test phase: Waiting period

Prior to performing the procedure from memory, the participants underwent a 30-second waiting period during which they were instructed to correctly solve as many arithmetic questions as they could from a set of 10 questions (see Appendix H). During this waiting phase there were no significant differences between the groups for the number of correct questions (TG: 4, CG: 4, p = 0.64, df = 59), incorrect questions (TG: 0, CG: 1, p = 0.08, df = 59) and total questions completed (TG: 4, CG: 5, p = 0.25, df = 59).

5.3.2.1.5 Test phase: Recall

During recall there was a statistically significant difference between the GoCube™ face rotation durations with 3.4 seconds for the TG and 2.5 seconds for the CG, with p = 0.01, df = 59. This result is broken down by gender and group in Table 5.2. This in turn led to the mean recall durations between the groups to be significantly different at 46 seconds for the TG and 30 seconds for the CG, with p < 0.01, df = 59. Broken down by gender within the groups, females in the TG had significantly longer rotation intervals than text-only CG females, causing longer mean recall durations in the TG as seen in Table 5.2. Female TG training instruction response times correlated to their recall Cube face rotations, with r = 0.68, p = 0.01 (R^2 = 0.46, p = 0.01) df = 13. Considered on its own, this might seem to suggest that training instruction format influences recall from memory in females. However, training duration and recall Cube face rotation durations correlated equally to mental rotation baseline in TG females (see Table 5.1) with r = -0.52, p = 0.05. This suggests that mental rotation abilities partially explain female TG Cube face rotation durations during training and recall.

Table 5.2. Mean recall phase GoCube™ face rotation durations and recall phase durations, with U-test significance values (Result) for the test group (TG) and control group (CG).

Feature	Group	Male	Female	Result
GoCube™ face rotation durations	**TG**	3 s	4 s	*0.78*
	CG	2 s	3 s	*0.90*
	Result	*0.44*	*0.01*	
Recall phase duration	**TG**	39 s	54 s	*0.11*
	CG	29 s	32 s	*0.60*
	Result	*0.22*	*< 0.01*	

5.3.2.1.6 Testing phase: Transfer

The mean of the differences between pre- and post-training Vandenberg rotation results is given in Table 5.3. This shows that males in the CG were the only participants not to improve in correct post-training Vandenberg rotations from baseline. In fact, on average, they got marginally fewer correct post-training Vandenberg rotations (-0.13). They were also the only participants to increase in incorrect Vandenberg rotations. These results partially contributed to the closing of what was seen as a significant gender-based discrepancy at baseline. Fig. 5.14 shows the sum of the differences between pre-and post-training mental rotation results. This shows that the statistically significant difference between males and females of the CG seen at baseline was also closed by CG females who also increased in correct mental rotations from baseline. In fig. 5.14 we see that the TG performed better than the CG in terms of both correct and incorrect mental rotations. This hints at a possible benefit of the animated 3D model to transfer of mental rotation abilities. This difference was not statistically significant with p = 0.42 for correct rotations and p = 0.81 for incorrect rotations in this case, and as such, further research in this area is merited. The number of wrong post experience mental rotations in CG males correlated to their subjective dominance (TG: 1.5, CG: 1.5, p = 0.96, df = 29) with ρ = 0.52, p = 0.05, df = 13. This correlation might begin to shed light on the gender-based differences reported in mental rotation abilities but further research in this area is required as males and females of the CG reported the same amount on subjective dominance in SAM questionnaire responses in this work.

Table 5.3. Mean difference between pre- and post-training Vandenberg rotations.

	Gender	Correct	Incorrect	Total
TG	**Male**	2	0	2
	Female	1	0	1
	Result	*0.88*	*0.62*	*0.85*
CG	**Male**	0	1	1
	Female	1	0	1
	Result	*0.35*	*0.40*	*0.41*

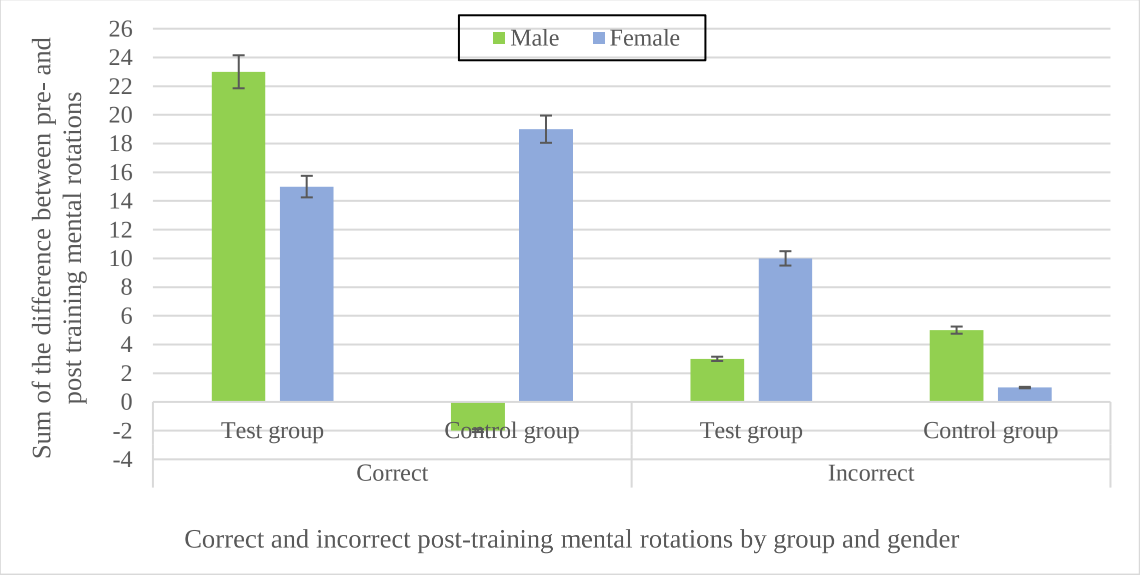


Fig. 5.14. The sum of the differences of correct and incorrect post-training mental rotation results by gender within the groups.

In summary, a statistically significant difference was seen in baseline mental rotation abilities between males and females of the CG in favour of male participants, which is a well-documented phenomenon in the literature [182]–[185]. Instruction response times were significantly quicker in the CG during practice and training. The CG also made fewer mistakes during these phases. CG females were significantly faster than their female TG counterparts during recall, although mental rotation abilities contributed to slower training and recall response times in TG females. What was seen as a significant gender gap in mental rotation baselines was not seen in post-training warranting further investigation in future work. The TG performed better than the CG in post-training Vandenberg mental rotations, although this difference was not statistically significant.

5.3.2.2 Implicit results

Physiological, facial expression and eye gaze data were analysed in the time domain as described in section 3.3. Statistical differences and correlations were sought between this data, the results of which are presented here. The physiological results are discussed in Section 5.3.2.2.1, the eye gaze results are discussed in Section 5.3.2.2.2 and finally, the facial expression results are discussed in Section 5.3.2.2.3.

5.3.2.2.1 Physiological results

Table 5.4 shows mean values of the physiological features that were statistically significant between the TG and CG. See Appendix P for the complete set of all physiological ratings from Study 2. Most of the physiological ratings in Table 5.4 were statistically significant between the female participants. Deviation from baseline to recall of minimum heart rate was statistically significant between male participants only with p = 0.04, df = 29. This suggests that female physiology is more susceptible to changes influenced by AR training than males. These statistically significant differences between the female participants are shown in Table 5.5. In Table 5.5 we see that the female TG participant’s maximum heart rate feature was significantly

Table 5.4. Statistically significant physiological ratings between the main test group (TG) and text-only control group (CG).

Physiological feature	TG	CG	Result
Maximum skin temperature during baseline	34.3 ^{o}C	33.4 ^{o}C	*0.05***
Maximum HR during baseline	92 bpm	85 bpm	*0.03***
Baseline to practice deviation of minimum IBI	0.01 s	0.04 s	*0.05**
Baseline to practice deviation of maximum HR	5 bpm	0 bpm	*0.01***
Baseline to training deviation of maximum HR	-1 bpm	7 bpm	*< 0.01**
Baseline to recall deviation of minimum HR	11 bpm	7 bpm	*0.03**
Baseline to recall deviation of maximum HR	-1 bpm	6 bpm	*0.04**

*: U-test, **: t-test

Table 5.5. Statistically significant physiological features between females of the test (TG) and control group (CG).

Physiological feature	TG	CG	Result
Maximum skin temperature during baseline	34.6 °C	33.0 °C	*0.02**
Maximum HR during baseline	93 bpm	84 bpm	0.03**
Baseline to practice deviation minimum IBI	0.01 s	0.06 s	< 0.01*
Baseline to practice deviation of maximum HR	6 bpm	-1 bpm	0.01*
Baseline to training deviation of maximum HR	-3 bpm	6 bpm	0.01*
Baseline to recall deviation of maximum HR	-2 bpm	6 bpm	0.04**

*: U-test, **: t-test

higher at baseline than their female CG counterparts, and it increased during the practice phase. This higher heart rate in TG females during practice is indicated as higher stress by their significantly shorter deviation from baseline to practice of minimum IBI ratings [162]. However, by the training and recall phases the female TG's maximum heart rate feature had reduced to below baseline levels, while the female CG's continued to remain 6 bpm above their baseline. This significantly reduced maximum heart rate amongst TG females during training correlated negatively to their mean number of training cycles (TG: 3, CG: 4, p = 0.44, df = 29) with r = -0.59, p = 0.02 (R^2 = 0.34, p = 0.02), df = 13, and extraneous cognitive load (TG: -4, CG: -3, p=0.806) with ρ = -0.61, p = 0.02 (R^2 = 1.00†, p = 0.02), df = 13. This suggests that HR is a correlate of task duration and cognitive load.

In summary, the TG females' HR was higher than CG females at baseline. However, during training and recall, HR was higher in CG females. The TG female's lower HR during training correlated to their lower mean quantity of training cycles and extraneous cognitive load. Overall, this suggests that CG females became more physiologically stressed than TG females during training and recall as reflected in significantly increasing maximum HR features. Lower HR was seen to be a negative correlate of training duration and extraneous cognitive load in TG females.

5.3.2.2.2 Eye gaze results

Table 5.6 shows how the TG depended less on the text instruction than the CG during the practice phase. Naturally, use of the 3D model and text in the TG resulted in a higher gaze shift rate. Both groups' eye gaze dwelled on the GoCube™ for an equivalent amount of time. When these results are broken down by gender in Table 5.7, we see an interesting pattern across both test groups in how males and females seem to process information differently. The males spent more time focusing on the instructions and less on the workpiece, while females spent less time on the instruction and more time focusing on the workpiece. In general, it seems that more effort invested during the informational phase reduces time required during the psychomotor phase [30]. Gaze dwell is a correlate of cognitive effort [186], and this result might show that, in the practice phase at least, the males processed the information more during the informational phase, while the females processed it more during the psychomotor phase. On balance, the male approach is marginally quicker than the female approach in both test conditions by circa 1.5 seconds per minute with p = 0.38.

Table 5.6. Practice phase eye gaze features per minute between the test (TG) and control (CG) groups with Spearman's U-test significance.

Eye gaze feature	TG	CG	Result
Text instruction dwell	28 s	36 s	0. 07
Workpiece dwell	23 s	23 s	0.87
Gaze Shifts	109	78	0.26

Table 5.7. Practice phase eye gaze features per minute between males and females of the test (TG) and control (CG) groups.

Group	Gaze feature	Male	Female	Result
TG	**Cube model dwell**	39 s	35 s	0.80*
	Text instruction dwell	27 s	25 s	0.97*
	Workpiece dwell	18 s	26 s	0.80*
	Gaze shifts	106	113	0.59*
CG	**Text instruction dwell**	34 s	31 s	0.79**
	Workpiece dwell	20 s	24 s	0.81*
	Gaze shifts	57	95	0.14*

*: U-test, **: t-test

Naturally, the TG had a higher gaze shift rate because they had the use of both the 3D model and the text instruction formats as well as gaze shift to the physical workpiece. This was true for males with $p = 0.06$ and for females with $p = 0.88$. These p values show that the TG males' gaze shifts contributed to most of this difference because TG and CG females' gaze shift was very similar. The 3D model seems to have reduced dependency on the text instruction in the TG, for males with $p = 0.23$ and for females with $p = 0.43$, with CG participants dwelling on the text instruction for circa 33 s while TG participant's gaze dwelled on the text instruction for circa 26 s.

Table 5.8 shows that the TG used the text instruction almost as much as the CG during the training phase. Table 5.9 shows that the TG used the 3D model far less during training than during the initial practice phase compared to Table 5.7. The position of the instructions in the TG may have influenced this result. As seen in Fig. 5.1, the text instruction appeared before the 3D model in top-down order. The literature suggests that if the 3D model was positioned above the text, the TG participants may have used it more during training [187]. Future research could be conducted to answer the question of how the order of instruction position influences their usage. The initial practice phase may have sufficed for TG participants to use the 3D model to verify their understanding of the text instructions. As they progressed through the training, they will have become more familiar with the procedure. This likely led to reduced need of the 3D model. It seems they did continue to use the text instruction during training to prompt the correct manipulation of the Cube faces until committed to memory by repetition. The benefit of the 3D model during the training phase seems to have been in reduced text instruction usage.

Table 5.8. Training phase eye gaze features between males and females of the test (TG) and control (CG) groups.

Eye gaze feature	TG	CG	Result
Text instruction dwell	26 s	28 s	0.64*
Workpiece dwell	16 s	20 s	0.26*
Gaze Shifts	61	52	0.20**

*: U-test, **: t-test

Table 5.9. Training phase eye gaze features between the test (TG) and control (CG) groups

Group	Gaze feature	Male	Female	Result
TG	**Cube model dwell**	11 s	12 s	0.33*
	Text instruction dwell	22 s	25 s	0.74*
	GoCube™ dwell	15 s	15 s	0.96**
	Gaze shifts	52	66	0.17**
CG	**Text instruction dwell**	22 s	28 s	0.11*
	GoCube™ dwell	16 s	19 s	0.57**
	Gaze shifts	45	52	0.47**

*: U-test, **: t-test

During training, female participants spent as much, or more time, on both the instructions and on the workpiece compared to male participants.

5.3.2.2.3 Facial expression results

Table 5.10 shows the facial expressions that were statistically significant between the groups (see Appendix Q for the full set of facial expression feature results). These are broken down in Table 5.11 by gender within the groups. Deviation from baseline to practice of AU17 MFEs in females of the CG correlated to their deviation from baseline to practice of minimum IBI ratings (TG: 0.001 s, CG: 0.062 s, $p < 0.01$, $df = 29$) with $r = -0.56$, $p = 0.03$ ($R^2 = 0.31$, $p = 0.03$), $df = 13$.

Deviation from baseline to practice of AU17 NFEs in TG females and AU12 NFEs in CG males correlated to deviation from baseline to practice of minimum skin temperature (TG males: 1.0 ^{o}C, TG females: 0.5 ^{o}C, CG males: 0.3 ^{o}C, CG females: 0.2 ^{o}C) with $r = 0.71$, $p = 0.03$ ($R^2 = 0.33$, $p = 0.03$), $df = 13$, and $r = -0.57$, $p = 0.03$ ($R^2 = 0.48$, $p < 0.01$), $df = 13$, respectively. Deviation from baseline to practice of AU12 NFEs in males of the CG correlated to their minimum EDA ratings during practice (TG males: 4.7 μS, CG males: 5.1 μS, $p = 0.15$, $df = 29$) with $r = -0.64$, $p = 0.01$ ($R^2 = 0.41$, $p = 0.01$), $df = 13$.

Table 5.10. Statistically significant differences between the test group (TG) and control group's (CG) facial expressions showing mean values, standard deviations and statistical test result.

AU feature	TG	CG	SD	Result
Baseline to practice deviation of AU10 NFEs	-0.6/min	0.7/min	1.8	*< 0.01**
Baseline to practice deviation of AU12 NFEs	1/min	3/min	3.8	*0.01**
Baseline to practice deviation of AU14 NFEs	-1/min	1/min	1.8	*< 0.01**
Baseline to practice deviation of AU15 NFEs	-1/min	2/min	2.9	*< 0.01**
Baseline to practice deviation of AU17 NFEs	-1/min	5/min	6.4	*< 0.01***
Baseline to practice deviation of AU20 NFEs	1/min	4/min	5.1	*0.01**
Baseline to practice deviation of AU23 NFEs	1 /min	8/min	8.2	*< 0.01***
Baseline to practice deviation of AU10 MFEs	-0.5/min	1.0/min	2.1	*< 0.01**
Baseline to practice deviation of AU12 MFEs	2/min	3/min	3.7	*0.03**
Baseline to practice deviation of AU14 MFEs	-1/min	1/min	2.4	*< 0.01**
Baseline to practice deviation of AU15 MFEs	-1/min	4/min	4.5	*< 0.01**
Baseline to practice deviation of AU17 MFEs	-1/min	15/min	17.0	*< 0.01**
Baseline to practice deviation of AU20 MFEs	2/min	9/min	1.0	*< 0.01**
Baseline to practice deviation of AU23 MFEs	3/min	12/min	14.7	*0.02**
Baseline to practice deviation of AU26 MFEs	2/min	4/min	4.4	*0.03**
Baseline to practice deviation of AU28 MFEs	-0.1/min	0.1/min	0.5	*0.01**
Baseline to training deviation of AU25 NFEs	3/min	6/min	5.2	*0.01***
Baseline to training deviation of AU23 MFEs	-4min	3/min	13.5	*0.05***
Recall AU20 NFEs	4.6/min	7.1/min	5.5	*0.03**
Recall AU25 MFEs	8/min	5/min	6.7	*0.02**

*: U-test, **: t-test

Table 5.11. Statistically significant different facial expression features by gender

Facial expression feature	Group	Male	Female	Result
Baseline to practice deviation of AU10 NFEs	**TG**	-0.8/min	-0.5/min	0.20^{**}
	CG	0.3/min	0.6/min	0.35^{*}
	Result	0.10^{*}	0.01^{*}	
Baseline to practice deviation of AU14 NFEs	**TG**	-1/min	0/min	0.34^{**}
	CG	1.0/min	0.6/min	0.79^{*}
	Result	0.05^{*}	$< 0.01^{*}$	
Baseline to practice deviation of AU15 NFEs	**TG**	-1/min	1/min	0.63^{**}
	CG	3/min	2/min	0.23^{*}
	Result	*0.01*	*0.02*	
Baseline to practice deviation of AU17 NFEs	**TG**	2/min	-3/min	0.04^{**}
	CG	4/min	6/min	0.35^{*}
	Result	0.23^{**}	$< 0.01^{*}$	
Baseline to practice deviation of AU20 NFEs	**TG**	1.2/min	0.7/min	0.80^{**}
	CG	5/min	3/min	0.84^{*}
	Result	*0.27**	*0.01**	
Baseline to practice deviation of AU23 NFEs	**TG**	1/min	1/min	0.90^{**}
	CG	8.34/min	7.96/min	0.61^{*}
	Result	*< 0.01*	*0.02*	
Baseline to practice deviation of AU10 MFEs	**TG**	-2/min	1/min	0.33^{**}
	CG	0.8/min	1.0/min	0.34^{*}
	Result	0.16^{**}	0.03^{*}	
Baseline to practice deviation of AU14 MFEs	**TG**	0/min	-2/min	1.00^{*}
	CG	4/min	1/min	0.40^{*}
	Result	0.05^{*}	$<0.01^{*}$	
Baseline to practice deviation of AU15 MFEs	**TG**	6/min	-3/min	0.23^{*}
	CG	7/min	5/min	0.45^{*}
	Result	*0.03*	*<0.01*	
Baseline to practice deviation of AU17 MFEs	**TG**	5/min	-7/min	0.10^{*}
	CG	17/min	13/min	0.23^{*}
	Result	*0.01*	*<0.01*	
Baseline to practice deviation of AU20 MFEs	**TG**	2/min	3/min	0.80^{**}
	CG	9/min	8/min	0.65^{*}
	Result	0.04^{*}	0.10^{*}	

Baseline to practice deviation of AU26 MFEs	**TG**	1.52/min	1.54/min	0.98^{**}
	CG	3/min	5/min	0.19^{*}
	Result	0.29^{*}	0.03^{*}	
Baseline to training deviation of AU25 NFEs	**TG**	2/min	4/min	0.33^{*}
	CG	8/min	5/min	0.22^{**}
	Result	0.01^{**}	0.50^{*}	
Recall AU20 NFEs	**TG**	3/min	6/min	0.22^{*}
	CG	8/min	7/min	0.70^{**}
	Result	0.03^{*}	0.49^{*}	
Recall AU25 MFEs	**TG**	9/min	8/min	0.65^{*}
	CG	5/min	6/min	0.67^{**}
	Result	0.23^{*}	0.37^{*}	

*: U-test, **: t-test

For all female participants (both groups), 14% of the variance in deviation from baseline to practice of AU14 NFEs was accounted for by variance in deviation from baseline to practice of diastolic BVP (TG: 94 nW, CG:20 nW) with $R^2 = 0.14$, $p = 0.04$, $df = 28$. Twenty five percent of the variance seen in their deviation from baseline to practice of AU15 MFEs accounted for the variance seen in their total eye gaze shifts during practice (TG males: 81, TG females: 124, CG males: 65 CG: females: 65) with $R^2 = 0.16$, $p = 0.03$, $df = 28$ and 26% of the variance seen in their deviation from baseline to practice of minimum IBI ratings (see Appendix P) with $R^2 = 0.26$, $p < 0.01$, $df = 28$. Deviation from baseline to practice of AU15 MFEs was significantly different between male and female participants of the TG with $p = 0.01$, $df = 29$. In males of the CG, 22% of the variance in deviation from baseline to practice of AU17 MFEs was accounted for by variance in gaze shift rate with $R^2 = 0.22$, $p = 0.08$, $df = 13$.

Deviation from baseline to practice of AU17 MFEs in males of the CG, AU20 NFEs in females of the TG and AU20 MFEs in males of both groups correlated to practice duration (TG male: 63 s, TG female: 75 s, CG male: 39 s, CG female: 56 s) with $r = -0.61$, $p = 0.02$ ($R^2 = 0.37$, $p = 0.02$), $df = 13$, $r = 0.74$, $p < 0.01$ ($R^2 = 0.55$, $p < 0.01$), $df = 13$ and $r = 0.50$, $p = 0.01$ ($R^2 = 0.31$, $p = 0.03$), $df = 28$ respectively. Deviation from baseline to practice of AU17 MFEs in CG males and AU20 NFEs in TG females also correlated to their practice instruction response times (TG male: 5 s, TG female: 5 s, CG male: 4 s, CG female: 5 s) with $r = 0.61$, $p = 0.02$ ($R^2 = 0.37$, $p = 0.02$), $df = 28$ and $r = 0.89$, $p < 0.01$ ($R^2 = 0.80$, $p < 0.01$), $df = 28$.

In summary, deviation from baseline to practice of AU15 MFEs was significantly different between male and female participants. In females, this correlated to minimum IBI ratings during practice and to their practice phase eye gaze fixations. In males, deviation of from baseline to practice of AU17 MFEs correlated to their practice eye gaze fixations, instruction response times and practice duration. Deviation from baseline to practice of AU20 (NFE and MFE) was seen to correlate to practice duration. These condensed results suggests that AU20, and to a lesser extent AU17, are the best facial expression candidates for reproducibility as implicit indicators of AR users experience of task duration and that AU15 is an implicit facial expression indicator of stress in female AR trainees.

5.4 Summary

This study evaluated the influence of a combined text and animated interactive 3D model instruction format compared to a text only control on AR trainee QoE. A between-groups study design compared text-based instructions against text combined with an interactive animated 3D model. This evaluation used a fully featured AR GoCube™ manipulation training application in which both independent test groups benefitted from psychomotor phase [34] corrective instructions in the event of trainee errors. This was enabled by wireless Cube state tracking. Eye gaze, facial expression and physiological features were used to compliment subjective reports of affect, cognitive load, task load and QoE.

The combined text and interactive animated 3D model instruction format yielded slower instruction response times and more mistakes during practice and training. Results suggest the lesser amount of information presented in the text-only instruction format cause less extraneous cognitive load, which led to fewer mistakes and shorter training times. This trend continued into recall where TG participants were slower in performing Cube face rotations from memory, predominantly the female members of the group. TG female mental rotation baseline correlated to their training instruction response times and Cube face rotation intervals during recall. This might suggest that female trainees benefit from text-only instruction formats to improve training speed in line with the literature [179]. Faster recall speed from memory in female trainees due to text-only instruction during training cannot be ruled out due to the correlation seen here.

In addition to this, AR training in GoCube™ manipulation may have played some role in closing the significant difference between the genders in general mental rotation abilities. In general, the TG performed better than the CG in post-training mental rotation abilities. This lays the foundations for further research into the influence of AR training instructions on transfer to general mental rotation abilities.

HR and IBI features showed that TG female participants were significantly more stressed during an initial practice phase. However, CG females had a significantly higher heart rate during training and recall. There were multiple significantly different facial expression features between the test groups. The majority of these occurred as deviations from baseline during the initial practice phase. These facial expressions correlated to HR, IBI, EDA skin temperature, eye gaze, cognitive load, distraction and frustration. AU20 facial expressions were a common correlate of task duration and IBI ratings were a correlate of elements of mental task load.

The critique of the literature given in Chapter 2 showed that the terms ‘delight’ and ‘annoyance’ are explicitly used in the definition of QoE. They represent diametrically opposing ends of a spectrum of emotions that reflect the degree of fulfilment of a user’s pragmatic and hedonic needs and expectations. However, delight and annoyance are depicted throughout the literature in 2D emotion space graphs as having different amounts of arousal and valence [79], [80], [83]–[85]. This gave rise to the question, what is the significance of this asymmetry to the definition of QoE? To help answer this question, the methodology of Study 2 in particular, was designed to evaluate the significance of emotion semantics to the participants as part of research sub question 2. As part of the explicit measures phase of Study 2, the participants were asked to report their post experience emotion state using open-ended terms, the SAM affect questionnaire and a label from the 2D emotion space taken from [85]. Elements of the participants’ emotional state were also recorded in their facial expressions and physiological ratings. The initial reasoning was that the presence of statistically significant correlations within and between the emotion terms would signify that the terms had a strong meaning to the participants. This may then necessitate a change in the definition of QoE using more symmetrically opposed emotions. A lack of statistically significant correlations within and between the emotion terms would signify no strong meaning to the participants, in which case, perhaps no change would necessarily be required to the definition of QoE as far as the general population is concerned.

As it transpired, there were no statistically significant correlations seen within or between the emotion terms used by the participants, suggesting no strong meaning of these terms to the participants. However, there is also an argument for re-evaluating the definition of QoE in the absence of statistically significant correlations to the emotion terms to facilitate meaningful academic discourse and interdisciplinary collaboration amongst scientists. QoE needs to be measurable, and therefore emotion terms that correlate in a statistically significant way to other manifestations of emotions such as physiological ratings and facial expressions, would be of greater utility to allow participants to report the emotion component of their QoE. Statistically significant correlations between SAM valence, arousal, and dominance responses that were seen to elements of frustration, performance, task load and gender, might suggest that more utility can be derived by communicating emotion in terms of these constituent components. The use of valence, arousal and dominance to communicate the central role of emotion in QoE may bring QoE research more in line with affective computing, human-computer interaction and machine learning. These fields of research commonly use valence, arousal and dominance dimensions for classification of emotion. However, continued research is needed in the form of correlation analysis to discover a consistent and measurable means of utility for communicating emotion.

There were no statistically significant differences in pre-experience interest or expectations recorded after the participant’s gave written consent. However, in female of the TG, acceptability correlated moderately to pre-experience interest and positive expectations of joy-of-use.

In Chapter 6 which follows, the thesis is concluded by revisiting the research questions and how the results of Study 1 and Study 2 have answered them. Future research opportunities that arise following from the research reported in this thesis are presented. This comes with recommendations for future methodologies including a cost/value analysis of the instruments used in this research. Future AR methodology recommendations also come in the form of AR augmentation design recommendations. Finally, the limitations of this research are acknowledged including their influence on the interpretability of the results of this thesis.

CHAPTER 6

Conclusion and Future Work

6.1 Thesis conclusions

This thesis addresses the topic of text and 3D AR instruction formats applied to procedure assistance and training over two studies. AR promises great utility for these roles in its potential to adapt to frequently changing procedures and to ensure correct learning during training. To fully realise this potential of AR for these roles, the QoE implications of instruction design decisions need to be well understood and the impact of relevant human, system and context influencing factors on QoE needs to be studied. This research focused on the influence of text and 3D instruction formats on AR user QoE for the procedure assistance and training roles across two studies. In Study 1, AR's utility for presenting text-based procedure assistance instruction was compared to a paper-based medium. Text only instruction was used to control for clarity, precision, and user comprehension. In Study 2, an interactive 3D model of the workpiece, combined with text, was compared against a text only instruction format for the training role. The 3D format was used as it is one of the main advantages offered by AR for interactive training.

Both Study 1 and Study 2 used a Rubik's Cube® style task. In Study 1, the AR application's ability to optimally solve the Rubik's Cube® from any of 10^{19} possible states was intended to provide a proof of concept for the adaptability of AR procedure assistance such as in mass customisation. In Study 2, the use of the Rubik's Cube® style workpiece was used to evaluate its influence on transfer to general mental rotation abilities as well as learning of a specific Cube manipulation procedure. Study 1 and Study 2 were conducted with the aim of answering two distinct questions arising from the literature as highlighted in Chapter 1 and Chapter 2. Study 1 was conducted with the aim of answering research question 1:

How does text instruction in AR influence user QoE for procedure assistance compared to a paper-based control?

Study 2 was conducted with the aim of answering research question 2:

How does a combined text and interactive animated 3D model instruction format influence user QoE for procedure training compared to a text-only instruction format?

Study 1 answered research question 1 by showing that AR yielded procedure assistance gains over paper-based instruction in terms of procedure completion duration and error reduction. This confirmed that AR better fulfilled the user's pragmatic procedure assistance needs. However, the AR group reported significantly more distraction and discomfort and less usefulness with the AR procedure assistance medium. Consequently, it seems that the Meta 2™ AR HMD application did not fulfil the user's hedonic needs and expectations to the same level as the paper-based instruction format. Positive arousal and valence correlated significantly to positive joy-of-use across the sample of Study 1. In the control group, positive valence also correlated moderately to interest and ease ease-of-use. Correlations between EDA and IBI to mental task load suggests that the subjective experience of mental task load can be measured in these physiological ratings. Minimum physiological features typically deviated more from baseline than mean and maximum ratings. IBI and BVP deviations showed increasing stress in both groups with this being marginally higher in the AR group. Longer task durations partially accounted for increasing stress across the entire Study 1 sample. A combination of correlation and regression analysis showed how higher skin temperature was partially accounted for by higher subjective joy-of-use in the control group.

Study 2 answered research question 2 by showing that the text-only instruction format resulted in quicker mean instruction response times and fewer mistakes than the combined text and model instruction format. This suggests that the text-only instruction format better fulfilled the AR trainee's pragmatic needs in terms of the training itself. The combined text and model instruction format may have contributed to slower Cube face rotation intervals during training and recall in females, but this was also linked to their mental rotation baseline.

A statistically significant gender difference seen in baseline mental rotation abilities in favour of males during Study 2 was not present in post-training mental rotation performance due to an improvement in post-training female performance. It is believed the use of the Rubik's Cube® as a workpiece during training transfer positively to general mental rotation abilities. This warrants further investigation in future work. The TG performed better than the CG in post-training Vandenberg mental rotations suggesting that the 3D visualisation of the Cube may have benefitted more to transfer to mental rotation abilities. More stress was indicated in higher heart rate amongst CG female participants than their TG female counterparts during training and recall. Heart rate was a correlate of task duration and extraneous cognitive load. AU20 facial expressions were a common correlate of task durations across both Study1 and Study 2.

There was a noticeable absence of correlations between open-ended and 2D emotion space terms used by the participants. Furthermore, there was also a noticeable absence of correlations between these emotion terms and any of the other metrics recorded during this Study 2. However, valence correlated to elements of frustration while dominance corelated to elements of performance and overall task load across both test groups. Arousal, valence and dominance combined correlated significantly to gender. TG females reported more cognitive effort invested during training than CG females while CG males reported more cognitive effort invested during recall than CG females.

The overarching research questions were broken down into five research sub-questions that were common to both Study 1 and Study 2.

The first sub-question was*, how do the different instruction formats influence the user's pragmatic needs and expectations?* In Study 1, the AR-based instruction medium yielded faster procedure completion durations and reduced errors compared to the paper-based instruction medium for procedure assistance. In Study 2, the text-only instruction format resulted in quicker mean instruction response times and fewer mistakes than the combined text and model instruction format. The users of the text-only instruction format also performed quicker in recalling the procedure from memory. This suggests that the text-only instruction format better fulfilled the AR trainee's pragmatic needs in terms of the training itself.

The second sub-question was, *what do users self-report in terms of the degree of fulfilment of their hedonic needs and expectations when experiencing the instruction formats?* In Study 1, the users of the AR-based instruction medium reported significantly more distraction and discomfort and less usefulness of the format. The AR hardware seems to have partially contributed to discomfort responses for wearers of spectacles. In Study 1, the CG reported higher valence and higher joy-of-experience. CG valence correlated to higher interest and ease-of-use. Their higher valence correlated more strongly to joy-of-experience than the AR groups. In Study 2, the combined instruction format resulted in a significantly higher perception of task performance. Users of the combined instruction format reported more confidence, joy-of-experience and interest. Users of the text-only instruction format report marginally more frustration, less distraction, less stress and greater acceptability. Valence and dominance correlated to elements of frustration, performance and overall task load respectively. Valence, arousal and dominance combined correlated to gender. There were no statistically significant correlations seen between the emotion terms used by the participants to communicate their post-experience emotion state and any of the other metrics recorded during Study 2. This calls into question the utility of such emotion terms. There were no statistically significant differences in pre-experience interest or expectations which were recorded after the participant's gave written consent. However, in female of the TG, acceptability correlated moderately to pre-experience interest and positive expectations of joy-of-use.

The third sub-question was, *can physiological measurements and facial expressions support a better understanding of user responses in the context of a QoE evaluation of the different instruction formats?* In Study 1, peripheral skin temperature was found to have the most discriminatory utility of joy-of-experience and affect between AR-based and paper-based procedure assistance instruction usage, while correlating moderately to MFEs of AU15 for both test groups. EDA and IBI features were seen to correlate to mental task load components. In Study 2, the higher perceived performance using the combined model and text instruction format correlated to deviation from baseline of a minimum skin temperature physiological feature.

The fourth sub-question was, *what is the influence of gender on the degree of fulfilment of pragmatic needs of the user of the different instruction formats?* In Study 1, there was significant correlation seen between gender and task performance. This includes mental rotation baseline. In Study 2, CG females were significantly faster than TG females during recall. Lower mental rotation abilities at baseline correlated to slower training and recall response times in TG females. What was seen as a significant gender gap in mental rotation baselines was not seen in post-training.

The fifth and final sub-question was, *how do different cognitive loads inherent in the different instruction formats influence user QoE?* In Study 1, there was no significant correlation between subjective cognitive load and any of the other recorded metrics. In Study 2, males using the combined instruction format reported higher mental demand than their female counterparts during both training and recall although this wasn't statistically significant.

The results of this research give rise to many future research opportunities. This not only includes a requirement for validation of the results and further investigation of the correlations seen in Study 1 and Study 2, but also to address questions raised in protocol design, instrument usage and the influence of different types and positions of AR instruction formats of user QoE. These future research opportunities are detailed in the following section. This is followed by methodology recommendations arising from the lessons learned during this research. This includes a cost-benefit style evaluation of the instruments used in this research and also some AR instruction design recommendations.

6.1.1 Future Work & Research opportunities

Results emanating from Study 2 have raised a question about the influence of the order of instruction position on instruction usage. The literature suggests that examples are used when present, in preference to procedural instructions, as the path of least cognitive effort [60]. In Study 2, this was only seen during the initial practice phase. Future work will involve further analysis on the eye gaze data to investigate 3D model usage over time as trainees progress from novice to expert. Scan path analysis will also shed light on the order of how the participants used the instructions (text first or model first). This analysis alone will not answer the question of the influence of the order of instruction positions on example usage (e.g., top down or left to right [187]) in preference of procedural instructions for configurations not used in this work.

In Study 2, there were several significant correlations between physiological and physical manifestations of emotion to objective and subject metrics. For example, heart rate correlated to task duration and extraneous cognitive load, and IBI correlated to eye gaze shift rate and AU15 MFEs to name but a few. The absence of correlations in Study 2 between open-ended and 2D emotion space terms to any of the other metrics captured might call into question the consensual understanding of such emotion terms. More research is needed in this area. In future work, careful participant guidance could be offered to ensure emotion open-ended and 2D emotion space terms are correctly selected without introducing bias into the data. For example, only if a participant chooses an open-ended term that cannot be classified as an emotion, could they be offered a further opportunity to select a different term of their choice. Similarly, if the term they chose is known to exist in the 2D emotion space, this could be first pointed out to them prior to letting them decide if they wish to choose the same term from the 2D emotion space or select a different one. Following this, correlation analysis could be carried out within and between these terms and other manifestations of emotion.

The 2D emotion space is described as a convenient tool for self-reporting the cognitive conceptualisation of emotion. The initial development of the 2D emotion space by J. Russell was essentially an exercise in semantic consensus. Ready access to modern electronic sensors provides an opportunity for future work to evaluate or improve the accuracy of the 2D emotion space label positions based on correlation analysis to physiological ratings. The current 2D space's arousal axis (typically the y-axis) is rather arbitrarily scaled from 0 to 1. There is potential for the development of a 2D emotion space where the arousal axis is scaled in units of a (or representing a combination of) physiological measure(s). This would not be without its challenges as regression analysis may be required to extrapolate the position of some emotion labels so as not to require experiences that elicit the full range of negative emotions to achieve a comprehensively labelled 2D space. Also, the subjective valence component of such emotion labels will require subjective reporting. Improvement of the 2D emotion space may produce an instrument that better allows users to communicate their emotion state in terms of commonly used emotion labels. However, the results from Study 2 seem to suggest that more utility could be derived from use of the valence, arousal and dominance components of emotion to communicate emotion state. Further research is required in the form of correlation analysis to identify the terms that correlate strongest to various manifestations of emotion (e.g., physiological rating, facial expressions). Identifying such terms poses one challenge. Encouraging people to correctly use any novel terms of utility in favour of commonly used terms may pose another challenge. The use of valence, arousal, and dominance terms may bring the field of QoE research more in line with those of affective computing, human-computer interaction, user-experience and machine learning, that currently use these terms.

The literature claims that training with a Rubik's Cube® transfers to an improvement in general mental rotation abilities [145]. Results emanating from Study 2 seem to suggest that such an effect is also possible in AR training, which warrants further investigation. At baseline in Study 2, there was a statistically significant difference between male and female participants of the control group for correct mental rotations in favour of the males. Statistically significant differences in mental rotation scores are commonly reported in the literature. What was seen in the results of Study 1 and Study 2 was that although male participants tend to get more correct rotations, females tend to get less incorrect rotations. The literature states that this is because females take more time to ensure correct rotations. This is exactly what was witnessed during this research. Even after females had chosen the correct shape, they proceeded with a process of elimination of all remaining alternative choices to make sure it was correct, while being aware that only one correct shape is present in the options. This phenomenon may have even been reflected in a statistically significant correlation between subjective dominance and more incorrect rotations in male participants. In Study 2, there were no statistically significant differences seen between the genders after having trained using the GoCube™ in AR. This was partially due to male participants of the control group getting more incorrect post-training mental rotations than at baseline but also due to female participants of the control group getting more correct post-training mental rotations. Further research in this area could definitively conclude if training in AR using a Rubik's Cube® type of workpiece benefits transfer to general mental rotation abilities. A gender balanced sample could highlight to what extent, and in what way, this is true for males and females.

6.1.2 Lessons learned and methodological recommendations

The literature review conducted as part of this research summarises the state-of-the-art in QoE evaluation of AR-based procedure assistance and training applications. Coupled with the development of the methodology and protocols for capturing a comprehensive set of implicit and explicit metrics (see Table 2.5), this work can be adapted to evaluate the influence of various AR augmentation formats and positions in different contexts. This includes using various tasks other than the Rubik's Cube® task considered in this work. Open questions arising from some of the inconclusive results of this research indicate areas where the methodology and protocol could be adapted to provide more conclusive results and to reduce redundancy and improve efficiency.

Eye gaze results gave rise to a question about protocol design in human trials. Eye gaze results emanating from Study 2 showed that the 3D model was only heavily used during the initial practice phase. In Study 2, dwell on the 3D model decreased during the test compared to the initial practice phase. Where novel implicit metrics such as eye gaze are intended to be analysed as part of an evaluation, consideration should be given to analysis of these metrics during any such practice phase also. This could provide valuable insight into the learnability or initial stress levels where participants are first introduced to novel immersive technologies such as AR.

The literature reports a complex relationship between the influence of novel technologies on acceptability. For example, in Study 1, acceptability was the only QoE aspect ranked higher by the AR group while all other QoE aspects were ranked higher for the paper-based control medium. In both Study 1 and Study 2, Likert statements 10 (see Appendices G and M) were posed as proxies to determine acceptability. Such statements are better described as attitude towards use. However, usefulness is considered to be a the most important metric of acceptability [158]. As such, statements regarding usefulness should be designed to capture acceptability of the technology being evaluated.

In terms of implicit metrics, facial expressions and physiological metrics only correlated moderately to subjective reports of affect, task load, cognitive load and QoE. In combination with subjective reports, these implicit metrics help to provide a more in depth understanding of user experience. However, this suggest that such implicit metrics would have only moderate utility in determining user QoE if used on their own. However, more research is needed including correlation analysis of these implicit metrics and subjectively reported experience. The Empatica E4 is a medical grade device that provides a convenient non-intrusive wristwatch form factor device for recording a large set of physiological ratings. It is easy to use and does not add much overhead to the testing methodology in terms of set-up time.

The NASA-TLX questionnaire takes a considerable amount of time to complete, typically adding circa 5 minutes onto the duration of a test. Questions adapted from the Leppink questionnaire could be used as an alternative where only cognitive load is being sought. The raw weight of the mental determinant of the NASA-TLX questionnaire can be used to record cognitive load, but the remaining determinants add a lot of overhead and should only be used where information about physical effort, performance perception and time pressure is warranted. Questions adapted from the Leppink cognitive load questionnaire may be used to record cognitive load in more detail (intrinsic, extraneous and germane) in more detail than those adapted from the Paas questionnaire.

The use of optical see-through AR HMDs and the Rubik's Cube® as a proof-of-concept workpiece in this research allowed for an evaluation of the influence of specific human, system and context influencing factors on AR user QoE. Regarding the challenge of visual workpiece state-tracking of highly configurable workpieces, such as the Rubik’s Cube®, custom development of a software template is an efficient means of tracking of a large number of workpiece states. This approach does not depend on the inordinate number of graphical templates that would be required of other template matching AR approaches. Target object state-tracking control of procedural AR application execution can be influenced by context factors largely outside of the control of the AR developer, even within a controlled laboratory setting. Coupled with a requirement for controlled and repeatable experimentation, workpiece tracking workarounds are so widespread as to have resulted in an over dependence of the Wizard-of-Oz approach in academic AR research. This is where the PI, or participant, controls procedural progression via alternative means such as user input instead of automatic progression defined by workpiece state. State-of-the-art mixed reality HMDs now feature a comprehensive suite of sensors to improve environmental awareness above and beyond that of visual perception alone. In Study 1, keyboard input was employed in the AR condition to provide the same level of instruction progression control in the AR condition as in the paper-based condition. HL2 Wi-Fi sensors were employed in Study 2 as a robust solution for workpiece state tracking for AR application execution control. These approaches to workpiece state tracking can be adapted in future work depending on the context of the research. If optical see-through AR HMDs are being used, the instruction design recommendations provided in the following section should be adhered to.

6.1.3 AR augmentation design recommendations

An attempt to directly overlay tile colour augmentations on the Rubik’s Cube® in this research made the vergence accommodation conflict problem evident; this is the human inability to focus on two different depth planes at the same time. This becomes evident if focus on both the workpiece and the augmentations is required. However, this research results in some augmentation design recommendations to overcome this challenge:

1. Instead of overlaying a target object that requires user focus, position augmentations in proximity to the target object instead (e.g. by means of object tracking). This will allow the AR user to subconsciously shift focus from augmentation to target object instead of trying to focus on both at once. Billboarding (augmentations within borders with solid background colour) is

recommended in the wild where environmental background colours are unknown to the developer in advance.

2. Direct overlay of target objects that do not require focus or acute attention should not cause vergence accommodation conflict. The AR user may not even realise they are experiencing diplopia when large featureless target objects are being directly overlaid. It is the presence of features such as the grid pattern and colours on the Rubik’s Cube® that allow the user to notice that their vision is crossed when these features cross over and become blurred. That is not to suggest that direct overlay of large featureless objects may not cause user discomfort after prolonged use.

Commercial video pass-through headsets (e.g., the Meta™ Quest™ and the Apple™ Vision Pro™) seem to be currently a popular solution to this challenge. This is where the user sees a video of their environment as opposed to seeing it directly. In this way, target objects and their augmentations are presented to the user on the same depth plane avoiding diplopia. Video pass-through headsets have the added benefit of joint AR/VR functionality. This is a hardware solution to the vergence accommodation conflict problem. Video-pass through HMDs could be used in future AR research to avoid the vergance-accommodation problem.

6.2 Limitations and their implications for the results

Convenience sampling resulted in mean sample ages of 32 years in both Study 1 and Study 2. The age distributions are skewed towards younger participants s where the mean age of an adult sample group should be 48 years old. However, there were no significant correlations seen between age and any of the other metrics recorded during this research in the samples of either Study 1 or Study 2. This suggests that the youthfully skewed age range of the samples had a limited influence, if any, on the results presented in this thesis.

This research only considered a single type of workpiece to evaluate the utility of AR for procedure assistance and training. This was a Rubik’s Cube®. To validate the reported results, the experimental evaluations can be repeated using other workpieces. This could be done to assess repeatability of the results reported herein.

As part of the explicit measures phase of Study 2, the participants used open ended terms to describe their post-experience emotional state. Considering the twelve nationalities represented in the sample, potential language barriers meant that 10% of the terms used could not be categorised as emotions. This included terms such as ‘succeed’, ‘challenging’ and ‘achievable’ as per Fig. 5.3 and Fig. 5.6. Only 35% of the sample whose open-ended term did appear in the 2D space persisted with their original choice. There is the chance that this is because, without guidance, they couldn’t find their original term even if it was present in the 2D space. There were no statistically significant correlations seen between either of these categories of emotion terms or to any of the other metrics recorded during Study 2. One limitation of this research is that guidance interventions were not offered in order to avoid the potential for biasing the terms chosen by the participants. As such, language barriers and a potential inability to find one’s open-ended term in the 2D space may have partially contributed to the lack of statistically significant correlations between the open-ended and 2D emotion labels used by the participants. Another limitation is that the order of the questionnaire statements was not randomised to control for the influence of questionnaire fatigue on participant responses. It is possible that these limitations have influenced the results of this work and further research is required to address this open question.

REFERENCES

[1] D. Gorecky, S. Worgan, and G. Meixner, *COGNITO: a cognitive assistance and training system for manual tasks in industry*. 2011, p. 56. doi: 10.1145/2074712.2074723.

[2] C. Keighrey, R. Flynn, S. Murray, and N. Murray, “A QoE evaluation of immersive augmented and virtual reality speech language assessment applications,” in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–6. doi: 10.1109/QoMEX.2017.7965656.

[3] J. Blattgerste, P. Renner, B. Strenge, and T. Pfeiffer, “In-Situ Instructions Exceed Side-by-Side Instructions in Augmented Reality Assisted Assembly,” in *Proceedings of the 11th PErvasive Technologies Related to Assistive Environments Conference*, Corfu Greece: ACM, Jun. 2018, pp. 133–140. doi: 10.1145/3197768.3197778.

[4] S. Fox, “The importance of information and communication design for manual skills instruction with augmented reality,” *J. Manuf. Technol. Manag.*, vol. 21, no. 2, pp. 188–205, Jan. 2010, doi: 10.1108/17410381011014369.

[5] I. Wechsung, K.-P. Engelbrecht, C. Kühnel, S. Möller, and B. Weiss, “Measuring the Quality of Service and Quality of Experience of multimodal human–machine interaction,” *J. Multimodal User Interfaces*, vol. 6, no. 1, pp. 73–85, Jul. 2012, doi: 10.1007/s12193-011-0088-y.

[6] V. Vieira, D. Rafael, and R. Agnihotri, “Augmented reality generalizations: A meta-analytical review on consumer-related outcomes and the mediating role of hedonic and utilitarian values,” *J. Bus. Res.*, vol. 151, pp. 170–184, Nov. 2022, doi: 10.1016/j.jbusres.2022.06.030.

[7] X. Yang, “Augmented Reality in Experiential Marketing: The Effects on Consumer Utilitarian and Hedonic Perceptions and Behavioural Responses,” in *Information Technology in Organisations and Societies: Multidisciplinary Perspectives from AI to Technostress*, Z. W. Y. Lee, T. K. H. Chan, and C. M. K. Cheung, Eds., Emerald Publishing Limited, 2021, pp. 147–174. doi: 10.1108/978-1-83909-812-320211006.

[8] E. Eiriksdottir and R. Catrambone, “Procedural Instructions, Principles, and Examples: How to Structure Instructions for Procedural Tasks to Enhance Performance, Learning,

and Transfer," *Hum. Factors J. Hum. Factors Ergon. Soc.*, vol. 53, no. 6, pp. 749–770, Dec. 2011, doi: 10.1177/0018720811419154.

[9] I. Radu and B. Schneider, "What Can We Learn from Augmented Reality (AR)?," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, in CHI '19. New York, NY, USA: ACM, 2019, p. 544:1-544:12. doi: 10.1145/3290605.3300774.

[10] R. E. Mayer and R. Moreno, "Aids to computer-based multimedia learning," *Learn. Instr.*, vol. 12, no. 1, pp. 107–119, Feb. 2002, doi: 10.1016/S0959-4752(01)00018-4.

[11] R. E. Mayer, "Multimedia aids to problem-solving transfer," *Int. J. Educ. Res.*, vol. 31, no. 7, pp. 611–623, Jan. 1999, doi: 10.1016/S0883-0355(99)00027-0.

[12] S. Antifakos, F. Michahelles, and B. Schiele, "Proactive Instructions for Furniture Assembly," in *UbiComp 2002: Ubiquitous Computing*, vol. 2498, G. Borriello and L. E. Holmquist, Eds., in Lecture Notes in Computer Science, vol. 2498. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2002, pp. 351–360. doi: 10.1007/3-540-45809-3_27.

[13] U. Neumann and A. Majoros, "Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance," in *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium (Cat. No.98CB36180)*, Mar. 1998, pp. 4–11. doi: 10.1109/VRAIS.1998.658416.

[14] D. Perkins and G. Salomon, "Transfer Of Learning,", International encyclopaedia of education, vol. 2, Jul. 1992.

[15] M. Terrell, "Anatomy of learning: Instructional design principles for the anatomical sciences," *Anat. Rec. B. New Anat.*, vol. 289B, no. 6, pp. 252–260, Nov. 2006, doi: 10.1002/ar.b.20116.

[16] S. Möller and A. Raake, *Quality of Experience, Advanced Concepts, Applications and Methods*. Springer, 2013.

[17] P. Gerjets, K. Scheiter, and R. Catrambone, "Can learning from molar and modular worked examples be enhanced by providing instructional explanations and prompting self-explanations?," *Learn. Instr.*, vol. 16, no. 2, pp. 104–121, Apr. 2006, doi: 10.1016/j.learninstruc.2006.02.007.

[18] P. Gerjets, K. Scheiter, and R. Catrambone, "Designing Instructional Examples to

Reduce Intrinsic Cognitive Load: Molar versus Modular Presentation of Solution Procedures," *Instr. Sci.*, vol. 32, Jan. 2004, doi: 10.1023/B:TRUC.0000021809.10236.71.

[19] N. Gavish *et al.*, "Evaluating virtual reality and augmented reality training for industrial maintenance and assembly tasks," *Interact. Learn. Environ.*, vol. 23, no. 6, pp. 778–798, Nov. 2015, doi: 10.1080/10494820.2013.815221.

[20] D. W. F. Van Krevelen and R. Poelman, "A Survey of Augmented Reality Technologies, Applications and Limitations," *Int. J. Virtual Real.*, vol. 9, no. 2, pp. 1–20, Jan. 2010, doi: 10.20870/IJVR.2010.9.2.2767.

[21] P. Milgram, H. Takemura, A. Utsumi, and F. Kishino, "Augmented reality: a class of displays on the reality-virtuality continuum," presented at the Photonics for Industrial Applications, H. Das, Ed., Boston, MA, Dec. 1995, pp. 282–292. doi: 10.1117/12.197321.

[22] G. Evans, J. Miller, M. Iglesias Pena, A. MacAllister, and E. Winer, "Evaluating the Microsoft HoloLens through an augmented reality assembly application," presented at the SPIE Defense + Security, J. (Jack) N. Sanders-Reed and J. (Trey) J. Arthur, Eds., Anaheim, California, United States: Degraded environments: sensing, processing, and display. Vol. 10197. International Society for Optics and Photonics, May 2017, p. 101970V. doi: 10.1117/12.2262626.

[23] D. Ungureanu *et al.*, "HoloLens 2 Research Mode as a Tool for Computer Vision Research," Aug. 2020, Accessed: Nov. 25, 2020. [Online]. Available: https://arxiv.org/abs/2008.11239v1

[24] A. State, G. Hirota, D. T. Chen, W. F. Garrett, and M. A. Livingston, "Superior Augmented Reality Registration by Integrating Landmark Tracking and Magnetic Tracking," in *Proceedings of the 23rd Annual Conference on Computer Graphics and Interactive Techniques*, in SIGGRAPH '96. New York, NY, USA: ACM, 1996, pp. 429–438. doi: 10.1145/237170.237282.

[25] X. Wang, S. K. Ong, and A. Y. C. Nee, "A comprehensive survey of augmented reality assembly research," *Adv. Manuf.*, vol. 4, no. 1, pp. 1–22, Mar. 2016, doi: 10.1007/s40436-015-0131-4.

[26] J. M. Coughlan and J. Miele, "AR4VI: AR as an Accessibility Tool for People with

Visual Impairments," in *2017 IEEE International Symposium on Mixed and Augmented Reality (ISMAR-Adjunct)*, Oct. 2017, pp. 288–292. doi: 10.1109/ISMAR-Adjunct.2017.89.

[27] N. Hrishikesh and J. J. Nair, "Interactive learning system for the hearing impaired and the vocally challenged," in *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, Sep. 2016, pp. 1079–1083. doi: 10.1109/ICACCI.2016.7732188.

[28] U. Neumann and S. You, "Natural feature tracking for augmented reality," *IEEE Trans. Multimed.*, vol. 1, no. 1, pp. 53–64, Mar. 1999, doi: 10.1109/6046.748171.

[29] J. Rambach, A. Pagani, M. Schneider, O. Artemenko, and D. Stricker, "6DoF Object Tracking based on 3D Scans for Augmented Reality Remote Live Support," *Computers*, vol. 7, no. 1, p. 6, Mar. 2018, doi: 10.3390/computers7010006.

[30] U. Neumann and A. Majoros, "Cognitive, performance, and systems issues for augmented reality applications in manufacturing and maintenance," in *Proceedings. IEEE 1998 Virtual Reality Annual International Symposium (Cat. No.98CB36180)*, Mar. 1998, pp. 4–11. doi: 10.1109/VRAIS.1998.658416.

[31] S. J. Henderson and S. Feiner, "Evaluating the benefits of augmented reality for task localization in maintenance of an armored personnel carrier turret," in *2009 8th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2009, pp. 135–144. doi: 10.1109/ISMAR.2009.5336486.

[32] S. Aldekhyl, R. B. Cavalcanti, and L. M. Naismith, "Cognitive load predicts point-of-care ultrasound simulator performance," *Perspect. Med. Educ.*, vol. 7, no. 1, pp. 23–32, Feb. 2018, doi: 10.1007/s40037-017-0392-7.

[33] V. Paelke, "Augmented reality in the smart factory: Supporting workers in an industry 4.0. environment," in *Proceedings of the 2014 IEEE Emerging Technology and Factory Automation (ETFA)*, Sep. 2014, pp. 1–4. doi: 10.1109/ETFA.2014.7005252.

[34] S. J. Henderson and S. K. Feiner, "Augmented reality in the psychomotor phase of a procedural task," in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp. 191–200. doi: 10.1109/ISMAR.2011.6092386.

[35] L. Hou, X. Wang, L. Bernold, and P. E. D. Love, "Using Animated Augmented Reality to Cognitively Guide Assembly," *J. Comput. Civ. Eng.*, vol. 27, no. 5, pp. 439–451,

Aug. 2013, doi: 10.1061/(ASCE)CP.1943-5487.0000184.

[36] F. Loch, F. Quint, and I. Brishtel, “Comparing Video and Augmented Reality Assistance in Manual Assembly,” in *2016 12th International Conference on Intelligent Environments (IE),* Sep. 2016, pp. 147–150. doi: 10.1109/IE.2016.31.

[37] S. Egger-Lampl, C. Gerdenitsch, L. Deinhard, R. Schatz, and P. Hold, “Assembly Instructions with AR: Towards measuring Interactive Assistance Experience in an Industry 4.0 Context,” in *2019 Eleventh International Conference on Quality of Multimedia Experience (QoMEX),* Jun. 2019, pp. 1–3. doi: 10.1109/QoMEX.2019.8743266.

[38] A. Cachada *et al.*, “Maintenance 4.0: Intelligent and Predictive Maintenance System Architecture,” in *2018 IEEE 23rd International Conference on Emerging Technologies and Factory Automation (ETFA),* Sep. 2018, pp. 139–146. doi: 10.1109/ETFA.2018.8502489.

[39] A. Uva, M. Gattullo, V. Manghisi, D. Spagnulo, G. Cascella, and M. Fiorentino, “Evaluating the effectiveness of spatial augmented reality in smart manufacturing: a solution for manual working stations,” *Int. J. Adv. Manuf. Technol.*, vol. 94, no. 1–4, pp. 509–521, Jan. 2018, doi: 10.1007/s00170-017-0846-4.

[40] M. Broecker, R. T. Smith, and B. H. Thomas, “Adaptive substrate for enhanced spatial augmented reality contrast and resolution,” in *2011 10th IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2011, pp. 251–252. doi: 10.1109/ISMAR.2011.6092401.

[41] “A Review on Industrial Augmented Reality Systems for the Industry 4.0 Shipyard - IEEE Journals & Magazine.” Accessed: Nov. 09, 2018. [Online]. Available: https://ieeexplore.ieee.org/document/8298525

[42] Y. Cho, J. Lee, and U. Neumann, “A Multi-ring Color Fiducial System and an Intensity-invariant Detection Method for Scalable Fiducial-Tracking Augmented Reality,” in *In IWAR*, 1998, pp. 147–165.

[43] S. R. R. Sanches, D. M. Tokunaga, V. F. Silva, A. C. Sementille, and R. Tori, “Mutual occlusion between real and virtual elements in Augmented Reality based on fiducial markers,” in *2012 IEEE Workshop on the Applications of Computer Vision (WACV)*, Jan. 2012, pp. 49–54. doi: 10.1109/WACV.2012.6163037.

[44] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche, “An augmented reality training platform for assembly and maintenance skills,” *Robot. Auton. Syst.*, vol. 61, no. 4, pp. 398–403, Apr. 2013, doi: 10.1016/j.robot.2012.09.013.

[45] T.-Y. Lin *et al.*, “Microsoft COCO: Common Objects in Context.” arXiv, Feb. 20, 2015. Accessed: Oct. 03, 2023. [Online]. Available: http://arxiv.org/abs/1405.0312

[46] V. Krauß, “Current Practices, Challenges, and Design Implications for Collaborative AR/VR Application Development,” p. 15, 2021.

[47] F. Quint, F. Loch, and P. Bertram, “The Challenge of Introducing AR in Industry - Results of a Participative Process Involving Maintenance Engineers,” *Procedia Manuf.*, vol. 11, pp. 1319–1323, Dec. 2017, doi: 10.1016/j.promfg.2017.07.260.

[48] D. K. Bengler and R. Passaro, “Augmented Reality in Cars Requirements and Constraints,” *ISMAR06 Ind. Track*, Jul. 2006.

[49] A. Tang, C. Owen, F. Biocca, and W. Mou, “Comparative Effectiveness of Augmented Reality in Object Assembly,” *NEW Horiz.*, no. 5, p. 8, 2003.

[50] A. Leykin and M. Tuceryan, “Automatic determination of text readability over textured backgrounds for augmented reality systems,” in *Third IEEE and ACM International Symposium on Mixed and Augmented Reality*, Nov. 2004, pp. 224–230. doi: 10.1109/ISMAR.2004.22.

[51] M. Fiorentino, S. Debernardis, A. E. Uva, and G. Monno, “Augmented Reality Text Style Readability with See-Through Head-Mounted Displays in Industrial Context,” *Presence Teleoperators Virtual Environ.*, vol. 22, no. 2, pp. 171–190, Aug. 2013, doi: 10.1162/PRES_a_00146.

[52] J. L. Gabbard, J. E. Swan, D. Hix, R. S. Schulman, J. Lucas, and D. Gupta, “An empirical user-based study of text drawing styles and outdoor background textures for augmented reality,” in *IEEE Proceedings. VR 2005. Virtual Reality, 2005.*, Mar. 2005, pp. 11–18. doi: 10.1109/VR.2005.1492748.

[53] C. Liu, S. Huot, J. Diehl, W. Mackay, and M. Beaudouin-Lafon, “Evaluating the benefits of real-time feedback in mobile augmented reality with hand-held devices,” in *Proceedings of the 2012 ACM annual conference on Human Factors in Computing Systems - CHI ’12*, Austin, Texas, USA: ACM Press, 2012, p. 2973. doi: 10.1145/2207676.2208706.

[54] M. Ljubojevic, V. Vaskovic, S. Stankovic, and J. Vaskovic, "Using supplementary video in multimedia instruction as a teaching tool to increase efficiency of learning and quality of experience," *Int. Rev. Res. Open Distrib. Learn.*, vol. 15, no. 3, Jun. 2014, doi: 10.19173/irrodl.v15i3.1825.

[55] E. S. Wilschut, R. Könemann, M. S. Murphy, G. J. W. van Rhijn, and T. Bosch, "Evaluating learning approaches for product assembly: using chunking of instructions, spatial augmented reality and display based work instructions," in *Proceedings of the 12th ACM International Conference on PErvasive Technologies Related to Assistive Environments*, in PETRA '19. New York, NY, USA: Association for Computing Machinery, Jun. 2019, pp. 376–381. doi: 10.1145/3316782.3322750.

[56] R. E. Mayer, "Techniques that increase generative processing in multimedia learning: Open questions for cognitive load research," in *Cognitive load theory*, R. Moreno, Ed., New York, NY, US: Cambridge University Press, 2010, pp. 153–177. doi: 10.1017/CBO9780511844744.010.

[57] C. Gerdenitsch, L. Deinhard, B. Kern, P. Hold, and S. Egger-Lampl, "Cognitive Assistance to Support Maintenance and Assembly Tasks: Results on Technology Acceptance of a Head-Mounted Device," 2021, pp. 276–284. doi: 10.1007/978-3-030-72632-4_20.

[58] E. Hynes, R. Flynn, B. Lee, and N. Murray, "A Quality of Experience Evaluation Comparing Augmented Reality and Paper Based Instruction for Complex Task Assistance," in *2019 IEEE 21st International Workshop on Multimedia Signal Processing (MMSP)*, Sep. 2019, pp. 1–6. doi: 10.1109/MMSP.2019.8901705.

[59] A. Tang, C. Owen, F. Biocca, and W. Mou, "Comparative Effectiveness of Augmented Reality in Object Assembly," *Proc. SIGCHI Conf. Hum. Factors Comput. Syst.*, no. 5, pp. 73–80, 2003, doi: 1-58113-630-7/03/0004.

[60] R. E. Mayer, "Techniques that reduce extraneous cognitive load and manage intrinsic cognitive load during multimedia learning," in *Cognitive load theory*, R. Moreno, Ed., New York, NY, US: Cambridge University Press, 2010, pp. 131–152. doi: 10.1017/CBO9780511844744.009.

[61] S. Webel, U. Bockholt, T. Engelke, M. Peveri, M. Olbrich, and C. Preusche, "Augmented Reality Training for Assembly and Maintenance Skills," *BIO Web Conf.*,

vol. 1, p. 00097, 2011, doi: 10.1051/bioconf/20110100097.

[62] S. Webel, U. Bockholt, and J. Keil, “Design Criteria for AR-Based Training of Maintenance and Assembly Tasks,” in *Virtual and Mixed Reality - New Trends*, vol. 6773, R. Shumaker, Ed., in Lecture Notes in Computer Science, vol. 6773. , Berlin, Heidelberg: Springer Berlin Heidelberg, 2011, pp. 123–132. doi: 10.1007/978-3-642-22021-0_15.

[63] R. T. Azuma, “A Survey of Augmented Reality,” *Presence Teleoperators Virtual Environ. 6 No 4*, vol. No. 4, no. 6, pp. 355–385, Aug. 1997, doi: 10.1162/pres.1997.6.4.355.

[64] G. Klinker, D. Stricker, and D. Reiners, “Augmented Reality: A Balance Act between High Quality and Real-Time Constraints,” *Mix. Real.-Merging Real Virtual Worlds*, pp. 325–346, 1999.

[65] E. Kruijff, J. E. Swan, and S. Feiner, “Perceptual issues in augmented reality revisited,” in *2010 IEEE International Symposium on Mixed and Augmented Reality*, Oct. 2010, pp. 3–12. doi: 10.1109/ISMAR.2010.5643530.

[66] K. Čopič Pucihar, P. Coulton, and J. Alexander, “Evaluating dual-view perceptual issues in handheld augmented reality: device vs. user perspective rendering,” in *Proceedings of the 15th ACM on International conference on multimodal interaction - ICMI ’13*, Sydney, Australia: ACM Press, 2013, pp. 381–388. doi: 10.1145/2522848.2522885.

[67] R. L. S. Silva, P. S. Rodrigues, D. Mazala, and G. Giraldi, “Applying Object Recognition and Tracking to Augmented Reality for Information Visualization,” *Tech. Rep. Tech. Rep. LNCC Braz.*, p. 7, 2004.

[68] Y. Pang, M. L. Yuan, A. Y. C. Nee, S. K. Ong, and K. Youcef-toumi, *A Markerless Registration Method for Augmented Reality based on Affine Properties*, vol. Volume 50. Proceedings of the 7th Australasian User interface conference., 2006.

[69] D. E. Qeshmy, J. Makdisi, E. H. D. Ribeiro da Silva, and J. Angelis, “Managing Human Errors: Augmented Reality systems as a tool in the quality journey,” *Procedia Manuf.*, vol. 28, pp. 24–30, Jan. 2019, doi: 10.1016/j.promfg.2018.12.005.

[70] S. Büttner, M. Prilla, and C. Röcker, “Augmented Reality Training for Industrial Assembly Work - Are Projection-based AR Assistive Systems an Appropriate Tool for

Assembly Training?," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu HI USA: ACM, Apr. 2020, pp. 1–12. doi: 10.1145/3313831.3376720.

[71] "ITU-T P.851 : Subjective quality evaluation of telephone services based on spoken dialogue systems." Accessed: Jun. 27, 2022. [Online]. Available: https://tinyurl.com/54v2nvw

[72] S. Werrlich, A. Daniel, A. Ginger, P.-A. Nguyen, and G. Notni, "Comparing HMD-Based and Paper-Based Training," in *2018 IEEE International Symposium on Mixed and Augmented Reality (ISMAR)*, Oct. 2018, pp. 134–142. doi: 10.1109/ISMAR.2018.00046.

[73] R. De Amicis, A. Ceruti, D. Francia, L. Frizziero, and B. Simões, "Augmented Reality for virtual user manual," *Int. J. Interact. Des. Manuf. IJIDeM*, vol. 12, no. 2, pp. 689–697, May 2018, doi: 10.1007/s12008-017-0451-7.

[74] P. Reichl *et al.*, "Towards a comprehensive framework for QOE and user behavior modelling," in *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX)*, Pylos-Nestoras: IEEE, May 2015, pp. 1–6. doi: 10.1109/QoMEX.2015.7148138.

[75] Y. Chen, K. Wu, and Q. Zhang, "From QoS to QoE: A Tutorial on Video Quality Assessment," *IEEE Commun. Surv. Tutor.*, vol. 17, no. 2, pp. 1126–1165, Secondquarter 2015, doi: 10.1109/COMST.2014.2363139.

[76] 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," p. 6, 2009.

[77] "New Recommendation for Subjective Video Quality Assessment Methods for Recognition Tasks[v1] | Preprints." Accessed: Nov. 17, 2022. [Online]. Available: https://www.preprints.org/manuscript/202101.0435/v1

[78] K. Brunnström *et al.*, "Qualinet White Paper on Definitions of Quality of Experience," Mar. 2013, Accessed: Nov. 10, 2017. [Online]. Available: https://hal.archives-ouvertes.fr/hal-00977812/document

[79] J. A. Russell, "A circumplex model of affect," *J. Pers. Soc. Psychol.*, vol. 39, no. 6, pp. 1161–1178, Dec. 1980.

[80] G. K. Verma and U. S. Tiwary, “Affect representation and recognition in 3D continuous valence–arousal–dominance space,” *Multimed. Tools Appl.*, vol. 76, no. 2, pp. 2159–2183, Jan. 2017, doi: 10.1007/s11042-015-3119-y.

[81] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994, doi: 10.1016/0005-7916(94)90063-9.

[82] L. F. Barrett and J. A. Russell, “The Structure of Current Affect: Controversies and Emerging Consensus,” *Curr. Dir. Psychol. Sci.*, vol. 8, no. 1, p. 5, 1999.

[83] K. R. Scherer, “What are emotions? And how can they be measured?,” *Soc. Sci. Inf.*, vol. 44, no. 4, pp. 695–729, Dec. 2005, doi: 10.1177/0539018405058216.

[84] J. Posner, J. A. Russell, and B. S. Peterson, “The circumplex model of affect: An integrative approach to affective neuroscience, cognitive development, and psychopathology,” *Dev. Psychopathol. Camb.*, vol. 17, no. 3, pp. 715–34, Jul. 2005.

[85] G. Paltoglou and M. Thelwall, “Seeing Stars of Valence and Arousal in Blog Posts,” *IEEE Trans. Affect. Comput.*, vol. 4, no. 1, pp. 116–123, Jan. 2013, doi: 10.1109/T-AFFC.2012.36.

[86] J. A. Russell, “Pancultural Aspects of the Human Conceptual Organization of Emotions,” *J. Pers. Soc. Psychol.*, vol. 45, no. 6, 1983, doi: 1281.

[87] A. Perkis *et al.*, “QUALINET White Paper on Definitions of Immersive Media Experience (IMEx),” *ArXiv200707032 Cs*, Jun. 2020, Accessed: Jul. 15, 2020. [Online]. Available: https://tinyurl.com/ye24vcs2

[88] K. De Moor, F. Mazza, I. Hupont, M. Ríos Quintero, T. Mäki, and M. Varela, “Chamber QoE: a multi-instrumental approach to explore affective aspects in relation to quality of experience,” presented at the IS&T/SPIE Electronic Imaging, B. E. Rogowitz, T. N. Pappas, and H. de Ridder, Eds., San Francisco, California, USA, Feb. 2014, p. 90140U. doi: 10.1117/12.2042243.

[89] B. M. Veljkovic, “RECOGNITION OF FACIAL MICRO EXPRESSIONS OF EMOTIONS DEPENDING ON PROFESSIONAL ORIENTATION, SATISFACTION WITH LIFE AND SOCIAL BACKGROUND OF STUDENTS,” *TEME*, p. 1061, Jan. 2019, doi: 10.22190/TEME1804061V.

[90] C. Keighrey, R. Flynn, S. Murray, S. Brennan, and N. Murray, “Comparing user QoE

via physiological and interaction measurements of immersive AR and VR speech language therapy applications.," *25th ACM Int. Conf. Multimed. ACM MM 2017*, vol. Thematic Workshop.

[91] U. Engelke, H. Nguyen, and S. Ketchell, "Quality of augmented reality experience: A correlation analysis," in *2017 Ninth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2017, pp. 1–3. doi: 10.1109/QoMEX.2017.7965638.

[92] R. Likert, "A technique for the measurement of attitudes," *Arch. Psychol.*, vol. 22 140, pp. 55–55, 1932.

[93] S. S. Sabet, C. Griwodz, and S. Möller, "Influence of primacy, recency and peak effects on the game experience questionnaire," in *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, in MMVE '19. Amherst, Massachusetts: Association for Computing Machinery, Jun. 2019, pp. 22–27. doi: 10.1145/3304113.3326113.

[94] "ITU-T P. 913, Series P: Terminals and subjective and objective assessment methods." Accessed: Oct. 15, 2018. [Online]. Available: https://www.itu.int/rec/T-REC-P.913-201603-I/en

[95] "ITU-T P.919 : Subjective test methodologies for 360° video on head-mounted displays." Accessed: Nov. 07, 2022. [Online]. Available: bit.ly/3KxfU2B

[96] T. Hoβfeld, R. Schatz, and S. Egger, "SOS: The MOS is not enough!," in *2011 Third International Workshop on Quality of Multimedia Experience*, Sep. 2011, pp. 131–136. doi: 10.1109/QoMEX.2011.6065690.

[97] S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research," in *Advances in Psychology*, vol. 52, Elsevier, 1988, pp. 139–183. doi: 10.1016/S0166-4115(08)62386-9.

[98] S. G. Hart and M. Field, "NASA-TASK LOAD INDEX (NASA-TLX); 20 YEARS LATER," *Proc. Hum. Factors Ergon. Soc. Annu. Meet. Los Angel. CA Sage Publ.*, vol. Vol. 50., no. No. 9., p. 5, 2006.

[99] L. M. Naismith, J. J. H. Cheung, C. Ringsted, and R. B. Cavalcanti, "Limitations of subjective cognitive load measures in simulation-based procedural training," *Med. Educ.*, vol. 49, no. 8, pp. 805–814, 2015, doi: 10.1111/medu.12732.

[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," *Educ. Psychol.*, vol. 38, no. 1, pp. 63–71, Mar. 2003, doi: 10.1207/S15326985EP3801_8.

[101] J. Leppink, F. Paas, C. P. M. Van der Vleuten, T. Van Gog, and J. J. G. Van Merriënboer, "Development of an instrument for measuring different types of cognitive load," *Behav. Res. Methods*, vol. 45, no. 4, pp. 1058–1072, Dec. 2013, doi: 10.3758/s13428-013-0334-1.

[102] Y. Ding, L. Shi, and Z. Deng, "Low-level Characterization of Expressive Head Motion through Frequency Domain Analysis," *IEEE Trans. Affect. Comput.*, pp. 1–1, 2018, doi: 10.1109/TAFFC.2018.2805892.

[103] E. Hynes, R. Flynn, B. Lee, and N. Murray, "A QoE Evaluation of an Augmented Reality Procedure Assistance Application," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2020, pp. 1–4. doi: 10.1109/QoMEX48832.2020.9123097.

[104] G. Donato, M. S. Bartlett, J. C. Hager, P. Ekman, and T. J. Sejnowski, "Classifying facial actions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 21, no. 10, pp. 974–989, Oct. 1999, doi: 10.1109/34.799905.

[105] L. F. Barrett, B. Mesquita, and M. Gendron, "Context in Emotion Perception," *Curr. Dir. Psychol. Sci.*, vol. 20, no. 5, pp. 286–290, Oct. 2011, doi: 10.1177/0963721411422522.

[106] Y. Huang, F. Chen, S. Lv, and X. Wang, "Facial Expression Recognition: A Survey," *Symmetry*, vol. 11, no. 10, p. 1189, Oct. 2019, doi: 10.3390/sym11101189.

[107] A. K. Davison, C. Lansley, N. Costen, K. Tan, and M. H. Yap, "SAMM: A Spontaneous Micro-Facial Movement Dataset," *IEEE Trans. Affect. Comput.*, vol. 9, no. 1, pp. 116–129, Jan. 2018, doi: 10.1109/TAFFC.2016.2573832.

[108] M. Takalkar, M. Xu, Q. Wu, and Z. Chaczko, "A survey: facial micro-expression recognition," *Multimed. Tools Appl.*, vol. 77, no. 15, pp. 19301–19325, Aug. 2018, doi: 10.1007/s11042-017-5317-2.

[109] S. Polikovsky, Y. Kameda, and Y. Ohta, "Facial micro-expressions recognition using high speed camera and 3D-gradient descriptor," pp. 16–16, Jan. 2009, doi: 10.1049/ic.2009.0244.

[110] "EBSCOhost | 131051581 | A survey: facial micro-expression recognition."

Accessed: Dec. 04, 2019. [Online]. Available: https://link.springer.com/article/10.1007/s11042-017-5317-2

[111] Z. Zeng, M. Pantic, G. I. Roisman, and T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 31, no. 1, pp. 39–58, Jan. 2009, doi: 10.1109/TPAMI.2008.52.

[112] W.-J. Yan, Q. Wu, J. Liang, Y.-H. Chen, and X. Fu, "How Fast are the Leaked Facial Expressions: The Duration of Micro-Expressions," *J. Nonverbal Behav.*, vol. 37, no. 4, pp. 217–230, Dec. 2013, doi: 10.1007/s10919-013-0159-8.

[113] T. Pfister, Xiaobai Li, G. Zhao, and M. Pietikäinen, "Recognising spontaneous facial micro-expressions," in *2011 International Conference on Computer Vision*, Nov. 2011, pp. 1449–1456. doi: 10.1109/ICCV.2011.6126401.

[114] S. Du, Y. Tao, and A. M. Martinez, "Compound facial expressions of emotion," *Proc. Natl. Acad. Sci.*, vol. 111, no. 15, pp. E1454–E1462, Apr. 2014, doi: 10.1073/pnas.1322355111.

[115] P. Ekman and R. L. Erika, "What the Face Reveals: Basic and Applied Studies of Spontaneous Expression Using the Facial Action Coding System (FACS), Second Edition."

[116] P. Lucey, J. F. Cohn, T. Kanade, J. Saragih, Z. Ambadar, and I. Matthews, "The Extended Cohn-Kanade Dataset (CK+): A complete dataset for action unit and emotion-specified expression," in *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops*, Jun. 2010, pp. 94–101. doi: 10.1109/CVPRW.2010.5543262.

[117] T. Kanade, J. F. Cohn, and Y. Tian, "Comprehensive database for facial expression analysis," in *Proceedings Fourth IEEE International Conference on Automatic Face and Gesture Recognition (Cat. No. PR00580)*, Mar. 2000, pp. 46–53. doi: 10.1109/AFGR.2000.840611.

[118] T. Baltrušaitis, P. Robinson, and L. P. Morency, "OpenFace: An open source facial behavior analysis toolkit," in *2016 IEEE Winter Conference on Applications of Computer Vision (WACV)*, Mar. 2016, pp. 1–10. doi: 10.1109/WACV.2016.7477553.

[119] M. F. Valstar *et al.*, "FERA 2015 - second Facial Expression Recognition and Analysis challenge," in *2015 11th IEEE International Conference and Workshops on*

Automatic Face and Gesture Recognition (FG), May 2015, pp. 1–8. doi: 10.1109/FG.2015.7284874.

[120] J. Aigrain, M. Spodenkiewicz, S. Dubuisson, M. Detyniecki, D. Cohen, and M. Chetouani, "Multimodal Stress Detection from Multiple Assessments," *IEEE Trans. Affect. Comput.*, vol. 9, no. 4, pp. 491–506, Oct. 2018, doi: 10.1109/TAFFC.2016.2631594.

[121] S. Ollander, C. Godin, A. Campagne, and S. Charbonnier, "A comparison of wearable and stationary sensors for stress detection," in *2016 IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, Oct. 2016, pp. 004362–004366. doi: 10.1109/SMC.2016.7844917.

[122] 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 *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6. doi: 10.1109/QoMEX.2016.7498964.

[123] C. Keighrey, R. Flynn, S. Murray, and N. Murray, "A Physiology-based QoE Comparison of Interactive Augmented Reality, Virtual Reality and Tablet-based Applications," *IEEE Trans. Multimed.*, pp. 1–1, 2020, doi: 10.1109/TMM.2020.2982046.

[124] A. Hirway, Y. Qiao, and N. Murray, "A QoE and Visual Attention Evaluation on the Influence of Audio in 360° Videos," in *2020 IEEE 21st International Symposium on "A World of Wireless, Mobile and Multimedia Networks" (WoWMoM)*, Aug. 2020, pp. 191–193. doi: 10.1109/WoWMoM49955.2020.00045.

[125] A. N. Moraes, R. Flynn, A. Hines, and N. Murray, "Evaluating the User in a Sound Localisation Task in a Virtual Reality Application," in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2020, pp. 1–6. doi: 10.1109/QoMEX48832.2020.9123136.

[126] D. P. Salgado *et al.*, "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 on - MMSys '18*, Amsterdam, Netherlands: ACM Press, 2018, pp. 517–520. doi: 10.1145/3204949.3208118.

[127] U. Engelke *et al.*, “Psychophysiology-Based QoE Assessment: A Survey,” *IEEE J. Sel. Top. Signal Process.*, vol. 11, no. 1, pp. 6–21, Feb. 2017, doi: 10.1109/JSTSP.2016.2609843.

[128] D. Concannon, R. Flynn, and N. Murray, “A quality of experience evaluation system and research challenges for networked virtual reality-based teleoperation applications,” in *Proceedings of the 11th ACM Workshop on Immersive Mixed and Virtual Environment Systems*, in MMVE ’19. Amherst, Massachusetts: Association for Computing Machinery, Jun. 2019, pp. 10–12. doi: 10.1145/3304113.3326119.

[129] C. McCarthy, N. Pradhan, C. Redpath, and A. Adler, “Validation of the Empatica E4 wristband,” in *2016 IEEE EMBS International Student Conference (ISC)*, May 2016, pp. 1–4. doi: 10.1109/EMBSISC.2016.7508621.

[130] T. B. Rodrigues, C. Ó. Catháin, N. E. O. Connor, and N. Murray, “A QoE Evaluation of Haptic and Augmented Reality Gait Applications via Time and Frequency-Domain Electrodermal Activity (EDA) Analysis,” in *2022 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct)*, Oct. 2022, pp. 297–302. doi: 10.1109/ISMAR-Adjunct57072.2022.00067.

[131] C. Setz, B. Arnrich, J. Schumm, R. La Marca, G. Tröster, and U. Ehlert, “Discriminating Stress From Cognitive Load Using a Wearable EDA Device,” *IEEE Trans. Inf. Technol. Biomed.*, vol. 14, no. 2, pp. 410–417, Mar. 2010, doi: 10.1109/TITB.2009.2036164.

[132] A. Yüce, H. Gao, G. L. Cuendet, and J.-P. Thiran, “Action Units and Their Cross-Correlations for Prediction of Cognitive Load during Driving,” *IEEE Trans. Affect. Comput.*, vol. 8, no. 2, pp. 161–175, Apr. 2017, doi: 10.1109/TAFFC.2016.2584042.

[133] N. Li and C. Busso, “Predicting Perceived Visual and Cognitive Distractions of Drivers With Multimodal Features,” *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 1, pp. 51–65, Feb. 2015, doi: 10.1109/TITS.2014.2324414.

[134] V. Zurloni, B. Diana, M. Elia, and L. Anolli, “Imposing Cognitive Load to Detect Prepared Lies: A T-Pattern Approach,” in *Discovering Hidden Temporal Patterns in Behavior and Interaction: T-Pattern Detection and Analysis with THEME^{TM}*, M. S. Magnusson, J. K. Burgoon, and M. Casarrubea, Eds., in Neuromethods. , New York, NY: Springer, 2016, pp. 63–82. doi: 10.1007/978-1-4939-3249-8_3.

[135] J. S. Lerner, R. E. Dahl, A. R. Hariri, and S. E. Taylor, “Facial Expressions of Emotion Reveal Neuroendocrine and Cardiovascular Stress Responses,” *Biol. Psychiatry*, vol. 61, no. 2, pp. 253–260, Jan. 2007, doi: 10.1016/j.biopsych.2006.08.016.

[136] H. Gao, A. Yüce, and J.-P. Thiran, “Detecting emotional stress from facial expressions for driving safety,” in *2014 IEEE International Conference on Image Processing (ICIP)*, Oct. 2014, pp. 5961–5965. doi: 10.1109/ICIP.2014.7026203.

[137] C. Viegas, S.-H. Lau, R. Maxion, and A. Hauptmann, “Towards Independent Stress Detection: A Dependent Model Using Facial Action Units,” in *2018 International Conference on Content-Based Multimedia Indexing (CBMI)*, Sep. 2018, pp. 1–6. doi: 10.1109/CBMI.2018.8516497.

[138] K. Kotovsky, J. R. Hayes, and H. A. Simon, “Why are some problems hard? Evidence from Tower of Hanoi,” *Cognit. Psychol.*, vol. 17, no. 2, pp. 248–294, Apr. 1985, doi: 10.1016/0010-0285(85)90009-X.

[139] “MacAllister et al. - 2017 - Comparing Visual Assembly Aids for Augmented Reali.pdf.” Accessed: Oct. 20, 2020. [Online]. Available: https://lib.dr.iastate.edu/cgi/viewcontent.cgi?article=1196&context=me_conf

[140] P. Plapper, M. Minoufekr, S. Kolla, and A. Sanchez, *Augmented Reality in Manual Assembly Processes*. 2020.

[141] N. Pathomaree and S. Charoenseang, “Augmented reality for skill transfer in assembly task,” in *ROMAN 2005. IEEE International Workshop on Robot and Human Interactive Communication, 2005.*, Aug. 2005, pp. 500–504. doi: 10.1109/ROMAN.2005.1513829.

[142] I. E. Nugraha, T. W. Sen, R. B. Wahyu, B. Sulistyo, and Rosalina, “Assembly instruction with augmented reality on Android application ‘Assembly with AR,’” in *2017 4th International Conference on New Media Studies (CONMEDIA)*, Nov. 2017, pp. 32–37. doi: 10.1109/CONMEDIA.2017.8266027.

[143] S. Webel, U. Bockholt, T. Engelke, N. Gavish, M. Olbrich, and C. Preusche, “An augmented reality training platform for assembly and maintenance skills,” *Robot. Auton. Syst.*, vol. 61, no. 4, pp. 398–403, Apr. 2013, doi: 10.1016/j.robot.2012.09.013.

[144] M. Funk, T. Kosch, and A. Schmidt, “Interactive Worker Assistance: Comparing the

Effects of In-Situ Projection, Head-Mounted Displays, Tablet, and Paper Instructions," p. 6.

[145] J. Valerie, G. Aylward, and K. Varma, "I Solved it! Using the Rubik's Cube to Support Mental Rotation in a Middle School Science Classroom," *Int. Soc. Learn. Sci.*, Jun. 2020, doi: 10.22318/icls2020.653.

[146] J. Valerie, "Supporting Middle School Students' Spatial Skills Through Rubik'S Cube Play," May 2020, Accessed: Jun. 24, 2021. [Online]. Available: https://tinyurl.com/k4v4wb4t

[147] S. Vogt, A. Khamene, and F. Sauer, "Reality Augmentation for Medical Procedures: System Architecture, Single Camera Marker Tracking, and System Evaluation," *Int. J. Comput. Vis.*, vol. 70, no. 2, p. 179, Nov. 2006, doi: 10.1007/s11263-006-7938-1.

[148] T. Rokicki, "Why It's Almost Impossible to Solve a Rubik's Cube in Under 3 Seconds." Accessed: Jul. 09, 2019. [Online]. Available: https://tinyurl.com/34vauyp8

[149] P. K. Kaiser, "Prospective Evaluation of Visual Acuity Assessment: A Comparison of Snellen Versus ETDRS Charts in Clinical Practice (An AOS Thesis)," *Trans. Am. Ophthalmol. Soc.*, vol. 107, pp. 311–324, Dec. 2009.

[150] Colblindor, "Digital color blindness test," Ishihara plates. Accessed: Dec. 02, 2020. [Online]. Available: bit.ly/3M6wkjB

[151] Committee on Vision, Assembly of Behavioural and Social Sciences National Research Council, "Procedures for Tesing Color Vision: Report of," *Natl. Acad. PRESS*, 1981.

[152] "3D Mental Rotation Training." Accessed: Jul. 13, 2023. [Online]. Available: https://vample.com/tools/mental-rotation/

[153] S. G. Vandenberg and A. R. Kuse, "Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization.," *Percept. Mot. Skills*, vol. 47, no. 2, pp. 599–604, Dec. 1978, doi: 10.2466/pms.1978.47.2.599.

[154] T. Kösa and F. Karakuş, "The effects of computer-aided design software on engineering students' spatial visualisation skills," *Eur. J. Eng. Educ.*, vol. 43, pp. 1–13, Sep. 2017, doi: 10.1080/03043797.2017.1370578.

[155] J. R. Lewis, "IBM computer usability satisfaction questionnaires: Psychometric evaluation and instructions for use," *Int. J. Human–Computer Interact.*, vol. 7, no. 1,

pp. 57–78, Jan. 1995, doi: 10.1080/10447319509526110.

[156] J. P. Chin, V. A. Diehl, and K. L. Norman, “Development of an instrument measuring user satisfaction of the human-computer interface,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, in CHI ’88. New York, NY, USA: Association for Computing Machinery, May 1988, pp. 213–218. doi: 10.1145/57167.57203.

[157] F. D. Davis, “User acceptance of information technology: system characteristics, user perceptions and behavioral impacts,” *Int. J. Man-Mach. Stud.*, vol. 38, no. 3, pp. 475–487, Mar. 1993, doi: 10.1006/imms.1993.1022.

[158] P. Legris, J. Ingham, and P. Collerette, “Why do people use information technology? A critical review of the technology acceptance model,” *Inf. Manage.*, vol. 40, no. 3, pp. 191–204, Jan. 2003, doi: 10.1016/S0378-7206(01)00143-4.

[159] “C930s Pro HD Webcam - Logitech.” Accessed: Feb. 15, 2023. [Online]. Available: https://tinyurl.com/4fwzcb2e

[160] D. Muhamad, “CNN and LSTM-Based Emotion Charting Using Physiological Signals,” *Dar MN Akram MU Khawaja SG Pujari CNN LSTM-Based Emot. Charting Using Physiol. Signals Sens. 2020 20 4551 Httpsdoiorg103390s20164551*.

[161] D. P. Salgado, R. Flynn, E. L. M. Naves, and N. Murray, “The Impact of Jerk on Quality of Experience and Cybersickness in an Immersive Wheelchair Application,” in *2020 Twelfth International Conference on Quality of Multimedia Experience (QoMEX)*, May 2020, pp. 1–6. doi: 10.1109/QoMEX48832.2020.9123086.

[162] J. Zhai and A. Barreto, “Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables,” in *2006 International Conference of the IEEE Engineering in Medicine and Biology Society*, Aug. 2006, pp. 1355–1358. doi: 10.1109/IEMBS.2006.259421.

[163] M. Valstar *et al.*, “AVEC 2016 - Depression, Mood, and Emotion Recognition Workshop and Challenge,” *ArXiv160501600 Cs*, May 2016, Accessed: Aug. 26, 2019. [Online]. Available: http://arxiv.org/abs/1605.01600

[164] E. Por, M. van Kooten, and V. Sarkovic, “Nyquist–Shannon sampling theorem”.

[165] T. Blu, P. Thevenaz, and M. Unser, “Linear interpolation revitalized,” *IEEE Trans. Image Process.*, vol. 13, no. 5, pp. 710–719, May 2004, doi: 10.1109/TIP.2004.826093.

[166] D. D. Salvucci and J. H. Goldberg, “Identifying fixations and saccades in eye-tracking protocols,” in *Proceedings of the symposium on Eye tracking research & applications - ETRA ’00*, Palm Beach Gardens, Florida, United States: ACM Press, 2000, pp. 71–78. doi: 10.1145/355017.355028.

[167] “Concise Guide to APA Style, Seventh Edition,” https://apastyle.apa.org. Accessed: May 02, 2023. [Online]. Available: https://apastyle.apa.org/products/concise-guide

[168] T. Rokicki, H. Kociemba, M. Davidson, and J. Dethridge, “The Diameter of the Rubik’s Cube Group Is Twenty,” *SIAM Rev.*, vol. 56, no. 4, pp. 645–670, Jan. 2014, doi: 10.1137/140973499.

[169] J. Pride, “From Superflip to Solved,” Academic Excellence Showcase Schedule. [Online]. Available: https://digitalcommons.wou.edu/aes_event/2017/all/226

[170] AndroidSteve, “Rubik-Cube-Wizard,” GitHub. Accessed: Dec. 29, 2020. [Online]. Available: https://github.com/AndroidSteve/Rubik-Cube-Wizard

[171] F. Loch, F. Quint, and I. Brishtel, “Comparing Video and Augmented Reality Assistance in Manual Assembly,” in *2016 12th International Conference on Intelligent Environments (IE)*, Sep. 2016, pp. 147–150. doi: 10.1109/IE.2016.31.

[172] T. H. C. Chiang, S. J. H. Yang, and G.-J. Hwang, “An Augmented Reality-based Mobile Learning System to Improve Students’ Learning Achievements and Motivations in Natural Science Inquiry Activities,” *J. Educ. Technol. Soc.*, vol. 17, no. 4, pp. 352–365, 2014.

[173] M. M. Bradley and P. J. Lang, “Measuring emotion: The self-assessment manikin and the semantic differential,” *J. Behav. Ther. Exp. Psychiatry*, vol. 25, no. 1, pp. 49–59, Mar. 1994, doi: 10.1016/0005-7916(94)90063-9.

[174] G. Makransky, S. Borre-Gude, and R. E. Mayer, “Motivational and cognitive benefits of training in immersive virtual reality based on multiple assessments,” Journal of Computer Assisted Learning. Accessed: Aug. 19, 2019. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1111/jcal.12375

[175] S. Werrlich, K. Nitsche, and G. Notni, “Demand Analysis for an Augmented Reality based Assembly Training,” in *Proceedings of the 10th International Conference on PErvasive Technologies Related to Assistive Environments*, Island of Rhodes Greece: ACM, Jun. 2017, pp. 416–422. doi: 10.1145/3056540.3076190.

[176] H. P. Bahrick, “Two-phase model for prompted recall,” *Psychol. Rev.*, vol. 77, no. 3, pp. 215–222, May 1970, doi: http://dx.doi.org/10.1037/h0029099.

[177] T. O. Nelson and R. J. Leonesio, “Allocation of self-paced study time and the “labor-in-vain effect,” *J. Exp. Psychol. Learn. Mem. Cogn.*, pp. 676–686, 1988.

[178] N. Murray, B. Lee, Y. Qiao, and G. Miro-Muntean, “The influence of human factors on olfaction based mulsemedia quality of experience,” in *2016 Eighth International Conference on Quality of Multimedia Experience (QoMEX)*, Jun. 2016, pp. 1–6. doi: 10.1109/QoMEX.2016.7498975.

[179] A. Duenser, H. Kaufmann, K. Steinbügl, and J. Glück, *Virtual and Augmented Reality as Spatial Ability Training Tools*, vol. 158. 2006. doi: 10.1145/1152760.1152776.

[180] P. Lemaire, H. Abdi, and M. Fayol, “The Role of Working Memory Resources in Simple Cognitive Arithmetic.” 1996.

[181] F. Paas, P. Ayres, and M. Pachman, “Assessment of Cognitive LoAd in muLtimediA LeArning theory, methods and Applications,” 2008, pp. 11–35.

[182] A. C. Neubauer, S. Bergner, and M. Schatz, “Two- vs. three-dimensional presentation of mental rotation tasks: Sex differences and effects of training on performance and brain activation,” *Intelligence*, vol. 38, no. 5, pp. 529–539, Sep. 2010, doi: 10.1016/j.intell.2010.06.001.

[183] M. Peters, “Sex differences and the factor of time in solving Vandenberg and Kuse mental rotation problems,” *Brain Cogn.*, vol. 57, no. 2, pp. 176–184, Mar. 2005, doi: 10.1016/j.bandc.2004.08.052.

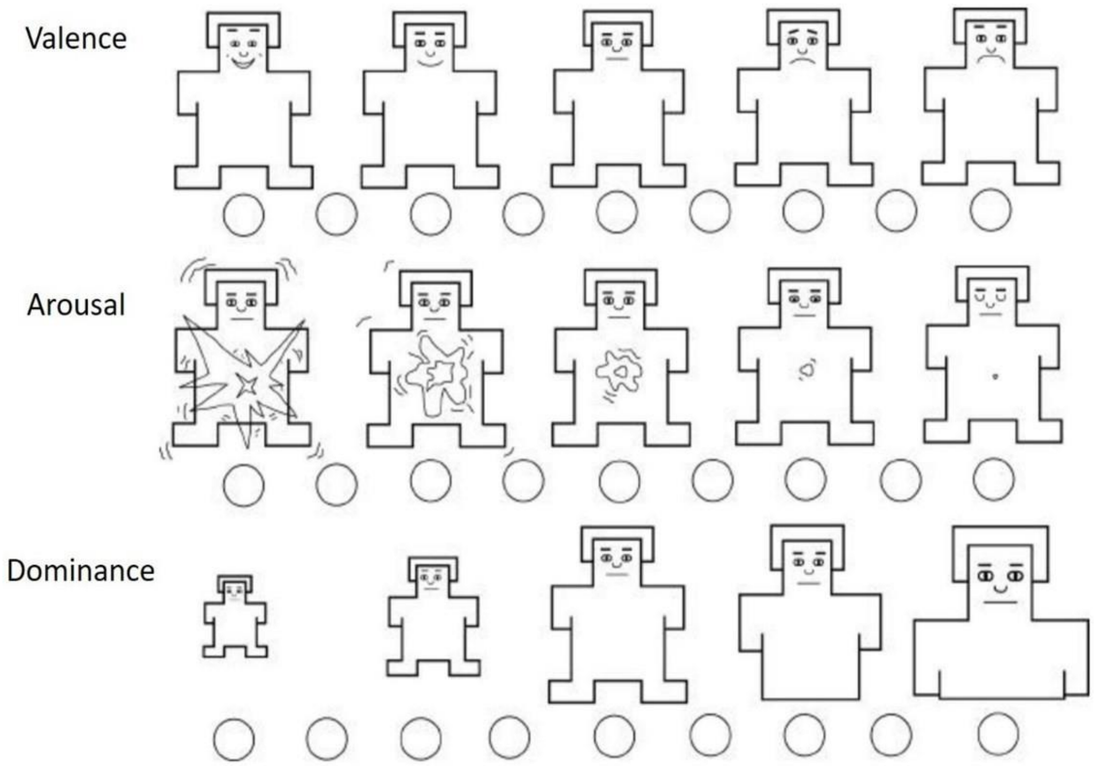
[184] T. D. Parsons *et al.*, “Sex differences in mental rotation and spatial rotation in a virtual environment,” *Neuropsychologia*, vol. 42, no. 4, pp. 555–562, Jan. 2004, doi: 10.1016/j.neuropsychologia.2003.08.014.

[185] M. Peters, W. Lehmann, S. Takahira, Y. Takeuchi, and K. Jordan, “Mental Rotation Test Performance in Four Cross-Cultural Samples (N = 3367): Overall Sex Differences and the Role of Academic Program in Performance,” *Cortex*, vol. 42, pp. 1005–1014, Jan. 2006, doi: 10.1016/s0010-9452(08)70206-5.

[186] S. Aldekhyl, R. B. Cavalcanti, and L. M. Naismith, “Cognitive load predicts point-of-care ultrasound simulator performance,” *Perspect. Med. Educ.*, vol. 7, no. 1, pp. 23–32, Feb. 2018, doi: 10.1007/s40037-017-0392-7.

[187] C. D. Gilbert and W. Li, “Top-down influences on visual processing,” *Nat. Rev. Neurosci.*, vol. 14, no. 5, pp. 350–363, May 2013, doi: 10.1038/nrn3476.

Appendix A – The Self-Assessment Manikin (SAM) Questionnaire



Appendix B - The NASA task load index (TLX) Questionnaire

Task load factor	Definition
Mental Demand	**How much mental add perceptual activity was required** (for example, thinking, deciding, calculating, remembering, looking, searching, etc)? Was the task easy or demanding, simple or complex, forgiving or exacting?
Physical Demand	**How much physical activity was required** (for example, pushing, pulling, turning, controlling, activating, etc.)? Was the task easy or demanding, slow or brisk, slack or strenuous, restful or laborious?
Temporal Demand Level	**How much time pressure did you feel** due to the rate or pace at which the tasks or task elements occurred? Was the pace slow and leisurely or rapid and frantic?
Performance Level	**How successful do you think you were in accomplish the goals of the task set by the experimenter** (or yourself)? How satisfied were you with your performance in accomplish these goals?
Effort Level	**How hard did you have to work** (mentally and physically) to accomplish your level of performance?
Frustration Level	**How insecure, discouraged, irritated, stressed**, and annoyed versus secure, gratified, content, relaxed, and complacent did you feel during the task?

For each of the six scales, evaluate the task you recently performed by cross on the scale's location that matches your experience.
Consider your responses carefully in distinguishing among the different task conditions and consider each individually.

1. Mental Demand (How mentally demanding was the task?/ How much mental and perceptual activity did you spend for this task?)

Very Low Very High

2. Physical Demand (How physically demanding was the task?/ How much physical activity did you spend for this task?)

Very Low Very High

3. Temporal Demand (How hurried or rushed was the pace of the task?/ How much time pressure did you feel in order to complete this task?)

Very Low Very High

4. Performance (How successful were you in accomplishing what you were asked to do?/ How successful do you think you were in accomplishing the goals of the task?)

Good Poor

5. Effort (How hard did you have to work to accomplish your level of performance?)

Very Low Very High

6. Frustration (How insecure, discouraged, irritated, stressed, and annoyed were you during this task?)

Very Low Very High

For each pair, choose the factor that contributed more to your experience of task load

□ Temporal Demand	□ Mental Demand
□ Performance	□ Mental Demand
□ Mental Demand	□ Effort
□ Temporal Demand	□ Effort
□ Physical Demand	□ Performance
□ Performance	□ Temporal Demand
□ Effort	□ Physical Demand
□ Mental Demand	□ Physical Demand
□ Performance	□ Frustration
□ Effort	□ Performance
□ Frustration	□ Effort
□ Frustration	□ Mental Demand
□ Physical Demand	□ Temporal Demand
□ Physical Demand	□ Frustration
□ Temporal Demand	□ Frustration

Appendix C - Study 1 Control group instruction manual

Rubik’s Cube

Solving Instructions

Step 1: Rotate the face with the Red tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 2: Rotate the face with the Orange tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 3: Rotate the face with the Yellow tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 4: Rotate the face with the Green tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 5: Rotate the face with the Red tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 6: Rotate the face with the Green tile at its centre 90 degrees ANTI-clockwise.

Next instruction overleaf ->

Step 7: Rotate the face with the Blue tile at its centre 90 degrees ANTI-clockwise.

Next instruction overleaf ->

Step 8: Rotate the face with the Yellow tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 9: Rotate the face with the Red tile at its centre 90 degrees ANTI-clockwise.

Next instruction overleaf ->

Step 10: Rotate the face with the Yellow tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 11: Rotate the face with the White tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 12: Rotate the face with the Green tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 13: Rotate the face with the Orange tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 14: Rotate the face with the White tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 15: Rotate the face with the Blue tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 16: Rotate the face with the Red tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 17: Rotate the face with the Blue tile at its centre 90 degrees ANTI-clockwise.

Next instruction overleaf ->

Step 18: Rotate the face with the Orange tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 19: Rotate the face with the Blue tile at its centre 90 degrees clockwise.

Next instruction overleaf ->

Step 20: Rotate the face with the White tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 21: Rotate the face with the Blue tile at its centre 180 degrees clockwise.

Next instruction overleaf ->

Step 22: The Cube should now be solved.

Appendix D - Study 1 Information Sheet

Principle Investigator*: Eoghan Hynes* ***Contact****: e.hynes@research.ait.ie*

A Quality of Experience evaluation of paper-based and augmented reality-based procedure assistance instruction formats.

A brief explanation of title:
This experiment consists of a quality of experience (QoE) evaluation of paper-based and augmented reality (AR) -based procedure assistance instruction formats. For this, you will follow either paper-based or AR-based instructions to assist you in a Rubik's Cube® solving procedure. AR combines virtual augmentation with the user's view of their environment. This allows users to view information and interact with their environment at the same time.

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 to ask questions. Thank you for reading this information document.

What is the purpose of this evaluation?

In this experiment, I aim to evaluate the influence of paper-based and AR-based procedure assistant instruction formats. The procedure used to this evaluation is an optimal Rubik's Cube® solving procedure.

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 you will be required to sign a consent form.

What does the experiment involve?

This experiment should last for approximately thirty minutes including questionnaire completion. Participants will be seated in a laboratory in the AIT Engineering Building. The lab will consist of a chair, table, AR headset and keyboard OR a paper instruction manual, a desk-mounted video camera, and an E4 wearable wrist band. Participants will attempt to solve a Rubik's Cube® by following instructions on either a head mounted AR device, or with a paper-based instruction manual. The participant will be fitted with an Empatica E4 wearable wrist band which will record your blood volume pulse, galvanic skin response, skin temperature, heart inter-beat interval and accelerometer data during the test. Head pose data will be gathered by the video camera. This is saved as spreadsheet data. The camera will not record a video of you. The participant will be asked to fill out questionnaires after the test to report their emotional state, task load, and QoE.

What do I have to do?

On the day of the test, participants will undergo visual and mental ability screening. The visual screening process involves testing the participant's visual perception using the standard Snellen eye chart. Testing for colour perception will use standard Ishihara colour-blind plates. Testing for mental rotation abilities will used the standard Vandenberg mental rotation 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?

Some people may find a testing environment stressful. Should a participant at any point feel any discomfort, it is important to communicate this to the Principle Investigator.

Will my participation be confidential?

Any information collected during this test will be strictly confidential. All data will be anonymised and securely stored. It will not be possible to identify you from the data collected.

What will happen to the results of the research project?

The results of this experiment will be published in in top tier research journals. It will be presented at international conferences as part of my research programme.

Thank You

Thank you for taking the time to read this information sheet and I hope you will decide to participate in this evaluation. Solving a Rubiks cube® is a very fulfilling experience!

Appendix E - Study 1 Consent Form

Title of Project:

A QoE evaluation of paper-based and AR-base procedure assistance instruction formats.

Name of Researcher:

Eoghan Hynes

Please Tick the Box

1. I confirm that I have read the information sheet dated ___/___/2018 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 am 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. □

6. I agree to take part in the above study. □

7. I do not know how to solve a Rubik's cube®. YES □ NO □

8. Sex M □ F □

9. Age ____

________________________ _______________ ____________________

Name of Participant Date Signature

_________________________ ________________ ____________________

Assessor Date Signature

10. Please read the statement below and tick the appropriate box:

I expect that attempting to solve a Rubik's Cube® will be an enjoyable experience:

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Appendix F - Study 1 Post Test Questionnaire

Q1: The Instructions were useful.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q2: Following the instructions was not interesting.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q3: I became physically uncomfortable during the experience.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q4: My experience was not frustrating.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q5: I felt confident in my ability to follow the instructions.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q6: Learning to use the instructions correctly was not easy.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q7: I really enjoyed my experience.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q8: The instructions were distracting.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

Q9: My experience was stressful.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Q10: I would like to experience this form of instruction again.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Q11: Attempting to solve a Rubik's Cube was an enjoyable experience.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Q12: Moving onto the next instruction was easy.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Q13: Using the instructions felt intuitive.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Q14: The mode of instruction was not natural.

Strongly disagree	**Disagree**	**Neutral**	**Agree**	**Strongly agree**

Appendix G – Likert scale questionnaire responses for Study 1 including mean opinion scores, standard deviations and U-test results.

No. Statement	AR MOS	AR SD	CG MOS	CG SD	Result
1. The instructions were useful.	4.75	0.44	4.96	0.20	0.04
2. Following the instructions was not interesting.	2.21	1.10	1.83	0.70	0.29
3. I became physically uncomfortable during the experience.	2.00	1.02	1.29	0.69	0.01
4. My experience was not frustrating.	4.21	1.14	4.63	0.77	0.13
5. I felt confident in my ability to follow the instructions.	4.63	0.50	4.75	0.44	0.36
6. Learning to use the instructions correctly was not easy.	1.46	0.51	1.38	0.58	0.46
7. I really enjoyed my experience.	4.38	0.82	4.63	0.50	0.41
8. The instructions were distracting.	1.75	0.53	1.42	0.58	0.03
9. My experience was stressful.	1.71	0.86	1.46	0.51	0.43
10. I would like to experience this form of instruction again.	4.25	0.74	4.17	0.76	0.71
11. Attempting to solve a Rubik's Cube® was an enjoyable experience.	4.58	0.72	4.75	0.44	0.58
12. Moving on to the next instruction was easy.	4.33	0.87	4.63	0.58	0.23
13. Using the instructions felt intuitive.	4.04	0.96	4.17	0.82	0.75
14. The mode of instruction was not natural.	2.25	0.94	1.88	0.90	0.13

Appendix H - Study 2 Waiting Phase Arithmetic Questionnaire.

1. 8 + 4 = _____

2. 34 + 77 = _____

3. 112 – 21 = _____

4. 9 * 8 = 72? True□ False □

5. 7 * 13 = 91? True□ False □

6. 101 – 9 = _____

7. 434 + 87 = _____

8. 16 + 9 + 66 = _____

9. 39 / 3 = _____

10. 9 * 13 = 117? True□ False □

Appendix I - Study 2 Affect Questionnaire

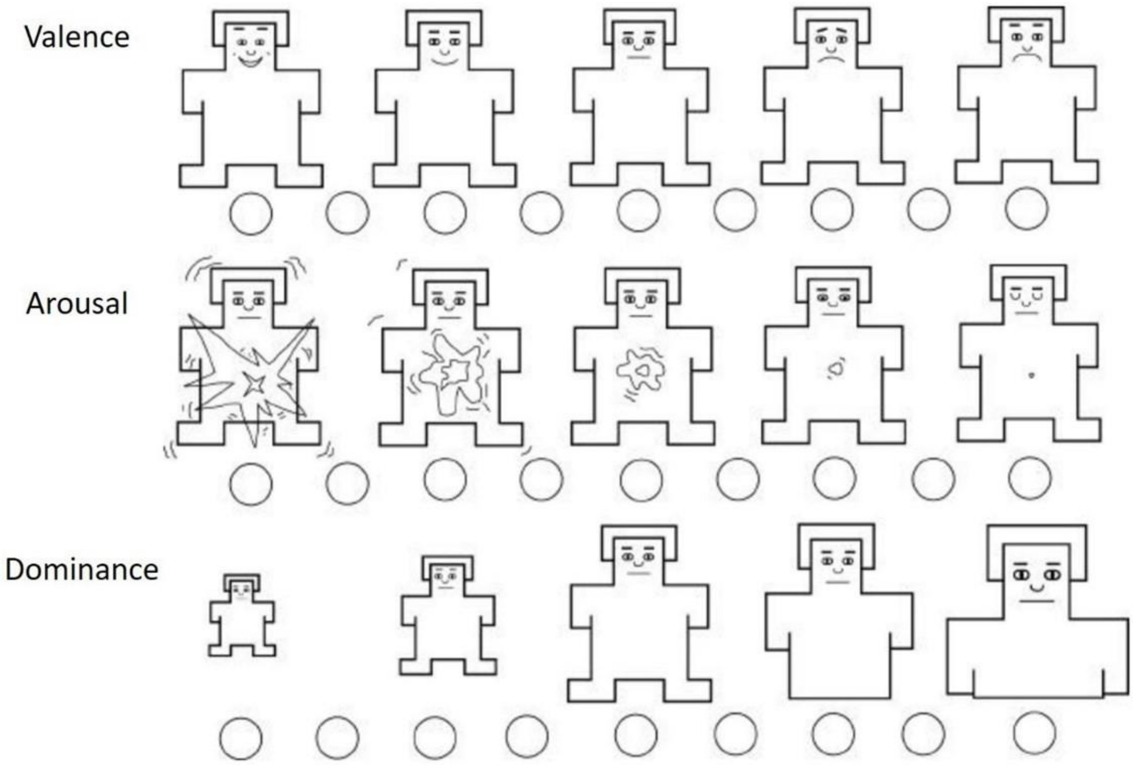
Part 1

On the line below, please write one word that best describes your current emotional state:

____

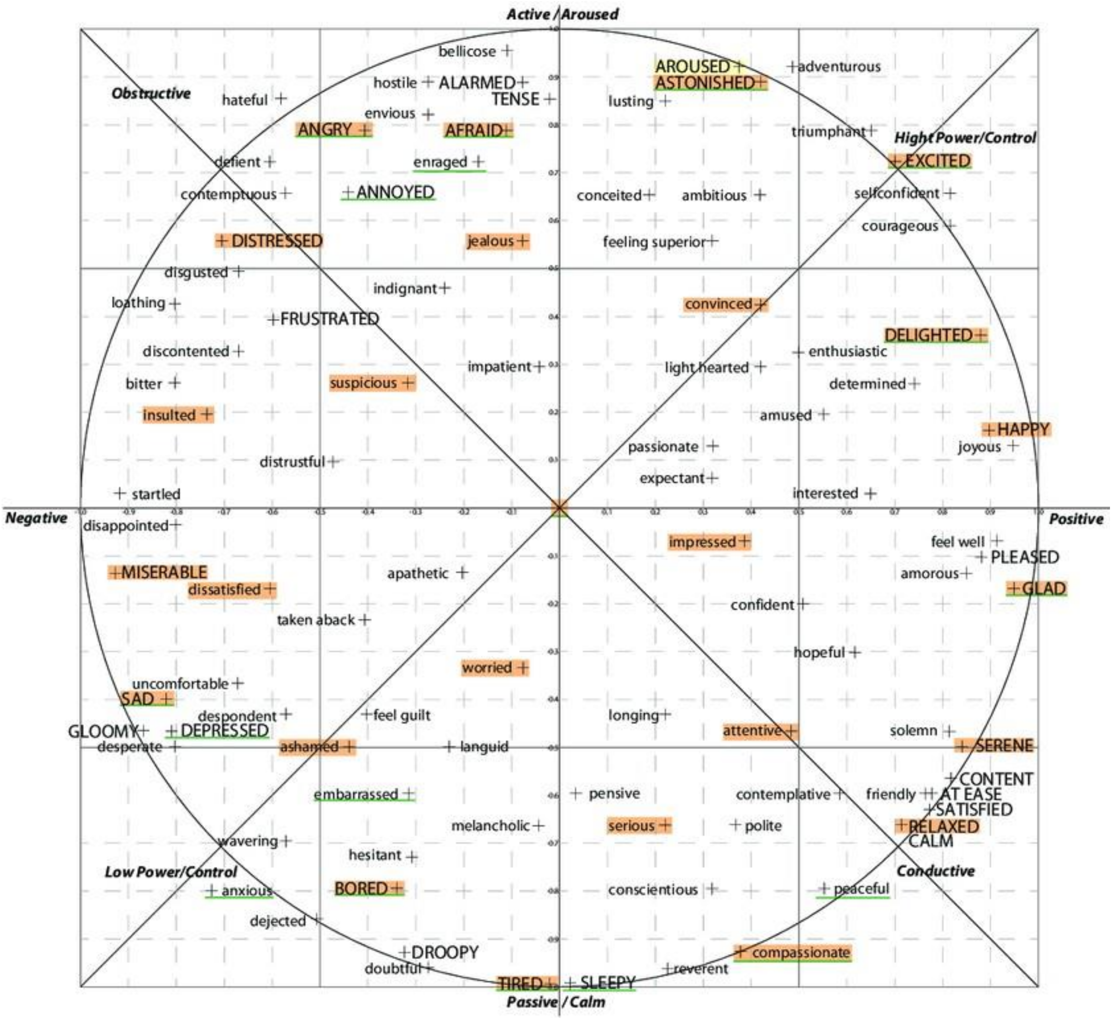
Part 2

Please tick one circle from each line below to describe the given dimension of your current emotional state:



Part 3

Please circle the word below that most closely describes your current emotional state:



Appendix J - Study 2 Post Test Questionnaire

1. I was confident in my ability to carry out the instructions.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

2. I became physically uncomfortable during the experience.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

3. The training was not frustrating.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

4. Training in Augmented Reality was an enjoyable experience.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

5. Augmented Reality is an interesting technology.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

6. I enjoyed my experience of training in Augmented Reality.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

7. The instructions were distracting.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

8**. The training was stressful.**

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

9. Augmented Reality is a good training platform.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

10. I would like to experience Augmented Reality again.

Strongly disagree	Disagree	Neutral	Agree	Strongly agree

11. Training covered concepts that I perceived as complex.

Strongly disagree	Disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Agree	Strongly agree

12. The training instructions were very unclear.

Strongly disagree	Disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Agree	Strongly agree

13. The training enhanced my understanding of the Go Cube.

Strongly disagree	Disagree	Moderately disagree	Slightly disagree	Neutral	Slightly agree	Moderately agree	Agree	Strongly agree

14. During training I invested:

Very very low mental effort	Very low mental effort	Low mental effort	Rather low mental effort	Neither low nor high mental effort	Rather high mental effort	High mental effort	Very high mental effort	Very very high mental effort

15. During the recall phase I invested:

Very very low mental effort	Very low mental effort	Low mental effort	Rather low mental effort	Neither low nor high mental effort	Rather high mental effort	High mental effort	Very high mental effort	Very very high mental effort

Appendix K - Study 2 Information Sheet

***Researcher**: Eoghan Hynes* ***Contact**: A00107408@ait.ie*

Title:

A Quality of Experience Evaluation of Augmented Reality Procedure Training Instruction Formats.

Introduction

I am inviting you to take part in a research evaluation to be carried out in the Engineering Department on the Athlone campus of the Technological University of the Shannon. The aim of this document is to explain why the research is being carried out and what it will involve.

A brief explanation of the experiment

Augmented reality (AR) is a technology that integrates computer generated augmentations into the user's view of their environment. In training apps these augmentations take the form of instructions. The goal of training is learning of a skill. In this evaluation you will learn a Go Cube™ (an electronic Rubik's Cube) manipulation procedure. We will evaluate your experience of the instructions.
If you are not clear on any points, please do not hesitate to ask questions. Thank you for reading this.

What is the purpose of this experiment?

AR is a promising training platform. AR training instruction design must focus on efficiency in terms of both trainee learning and headset resource usage due to relatively limited availability of processing, memory, and power. In this study, I aim to evaluate how text and animated instruction formats influence trainee QoE. This evaluation will evaluate combined text and animated 3D instructions against a text only control for a Go Cube training procedure. This evaluation will use the Microsoft™ HoloLens2™ mixed reality headset.

Do I have to take part?

It is entirely up to you to decide whether you wish to take part in this experiment. You can choose not to take part at your discretion. If you do decide to take part, you can keep this information sheet and you will be required to sign a consent form.

What does the experiment involve?

The evaluation will last about 45 minutes. Participants will be seated in a controlled laboratory in the TUS Athlone Engineering Building. The lab will consist of a chair, table, Microsoft™ HoloLens 2 ™ Mixed Reality headset, desk mounted video cameras and the Empatica E4 wrist band and a computer monitor used during screening. The E4 wrist band will record blood volume pulse, galvanic skin response, skin temperature, heart inter-beat interval and accelerometer data. Head pose data including facial expressions will be gathered by the video cameras. Eye gaze data will be recorded by sensors in the AR headset. The participant will be asked to fill out questionnaires before and after the evaluation to report their emotional state, task load and aspects of their QoE. One questionnaire involves arithmetic equations, designed to induce cognitive load, to control for long term learning.

What do I have to do?

During the evaluation, participants will undergo visual and cognitive screening. Participants will be trained in a Go Cube manipulation procedure by following a set of 14 AR instructions. A training cycle consists of a set of 14 instructions. You can repeat training cycles until you have learned the procedure. Your goal is to learn the Go Cube manipulation procedure in as few training cycles as possible. Learning will be evaluated in a post training recall phase in which **you will have to perform the Go Cube manipulation procedure as taught during training, but without any assistance**. Prior to the recall phase, you will be asked to perform some arithmetic to control for long term learning.

What are the possible disadvantages and risks of taking part?

Some people may find the AR headset mildly uncomfortable, and the evaluation mildly stressful due to cognitive load. Should a participant at any point feel any discomfort, it is important to communicate this to the researcher conducting the evaluation. You may opt out of the evaluation at any point. If you are pregnant or suspect that you may be pregnant, please let the administrator of the evaluation know.

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 manner, and it will not be possible to recognise you from this experiment as your details will be recorded by reference number only and not by name.

What will happen to the results of the research project?

The results of this evaluation will be published in my thesis and corresponding peer reviewed publications as part of my research programme.

Thank You

Thank you for taking the time to read this information sheet and I hope you will decide to participate in this evaluation.

Appendix L - Study 2 Consent Form

Title:
A Quality of Experience Evaluation of Augmented Reality Procedure Training Instruction Formats.

Researcher:
Eoghan Hynes: A00107408@student.ait.ie

Please Tick the Box

1. I confirm that I have read the information sheet on ___/___/2022 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 am not pregnant and/or I am not experiencing any signs of pregnancy. □
5. I understand that my participation is voluntary. □
6. I agree to take part in the above study. □
7. I have used augmented reality before. Y/N
8. Age ____

______________________ ____________ ________________
Name of Participant Date Signature

_______________________ ____________ ________________
Researcher Date Signature

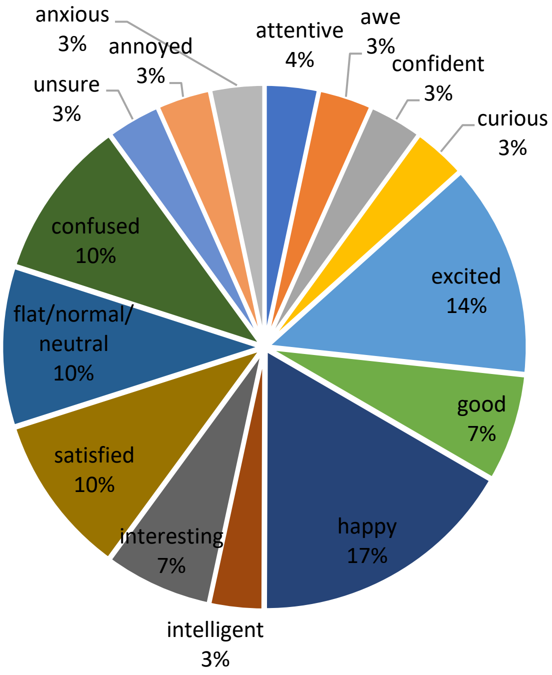
Augmented reality is an interesting technology:

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

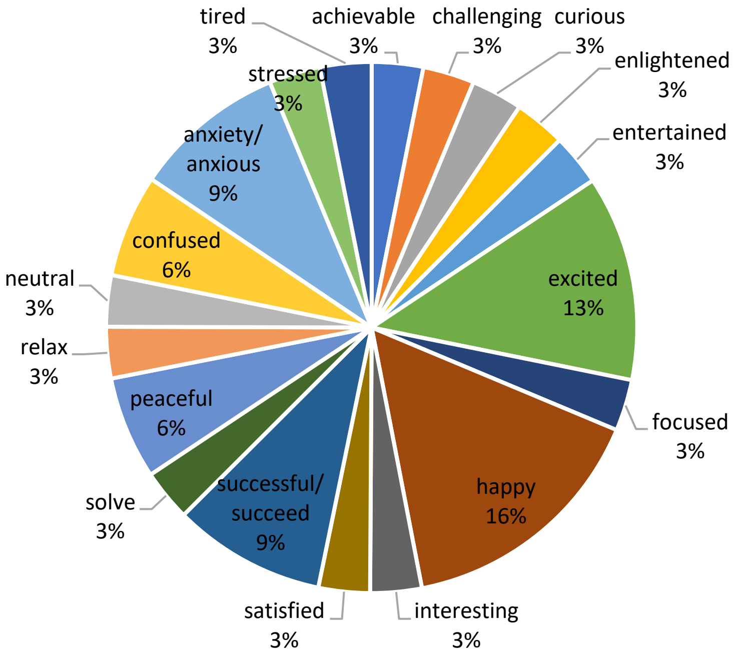
I expect that training in Augmented Reality will be an enjoyable experience:

Strongly Disagree	Disagree	Neutral	Agree	Strongly Agree

Appendix M - Study 2 Pie charts of the open-ended emotion terms used by the test group and the control group

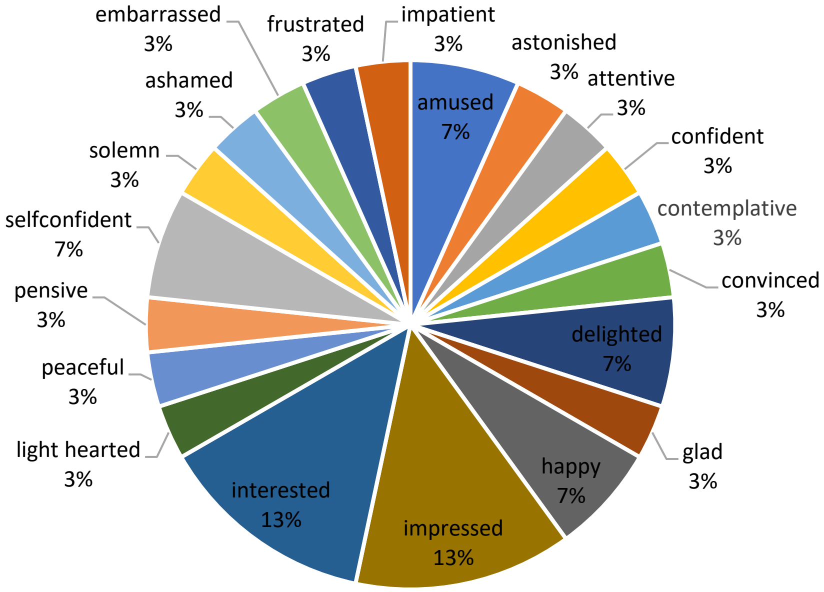


Test Group

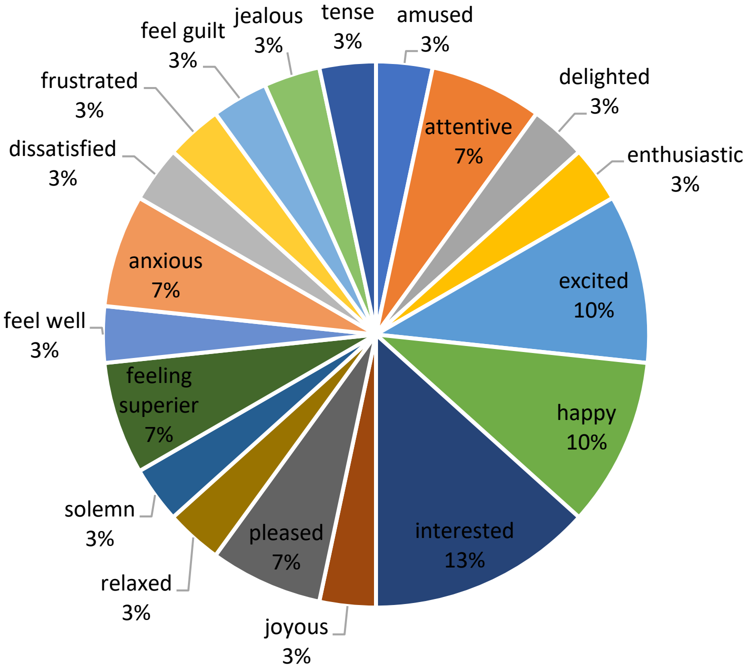


Control Group

Appendix N - Study 2 Pie charts of the 2D motion space terms used by the test group and the control group



Test group



Control group

Appendix O – Study 2 Likert scale questionnaire statements, mean opinion scores, standard deviations (SD) and U-test results

No.	**Statement**	**TG MOS**	**TG SD**	**CG MOS**	**CG SD**	**Result**
1	I was confident in my ability to carry out the instructions.	1.21	0.78	1.07	0.91	*0.60*
2	I became physically uncomfortable during the experience.	-1.81	0.79	-1.10	1.03	*0.14*
3	The training was not frustrating.	1.39	1.14	1.57	0.77	*0.54*
4	Training in Augmented Reality was an enjoyable experience.	1.64	0.48	1.60	0.50	*0.79*
5	Augmented Reality is an interesting technology.	1.79	0.42	1.73	0.45	*0.77*
6	I enjoyed my experience of training in Augmented Reality.	1.57	0.57	1.60	0.50	*0.75*
7	The instructions were distracting.	-1.36	0.68	-1.47	0.86	*0.20*
8	The training was stressful.	-1.04	1.11	-1.47	1.08	*0.72*
9	Augmented Reality is a good training platform.	1.46	0.58	1.50	0.51	*0.90*
10	I would like to experience Augmented Reality again.	1.61	0.50	1.70	0.54	*0.33*
11	Training covered concepts that I perceived as complex.	0.18	2.72	-0.57	2.62	*0.30*
12	The training instructions were very unclear.	-3.46	0.64	-3.23	1.89	*0.32*
13	Training enhanced my understanding of the GoCube™.	2.64	1.25	2.47	1.46	*0.89*
14	During training I invested this amount of cognitive effort.	0.16	1.53	0.07	1.72	*0.13*
15	During recall I invested this amount of cognitive effort.	0.02	1.88	0.37	1.45	*0.40*

Appendix P – Study 2 physiological rating means and significance

Physiological feature	TG	CG	Result
Minimum skin temperature during baseline	32.6 °C	33.0 °C	$\mathit{0.71}^{*}$
Minimum skin temperature during practice	33.3 °C	33.3 °C	$\mathit{0.62}^{*}$
Minimum skin temperature during training	33.5 °C	33.3 °C	$\mathit{0.42}^{*}$
Minimum skin temperature during recall	33.7 °C	33.4 °C	$\mathit{0.62}^{*}$
Mean skin temperature during baseline	33.4 °C	33.2 °C	$\mathit{0.68}^{**}$
Mean skin temperature during practice	33.6 °C	33.4 °C	$\mathit{0.22}^{*}$
Mean skin temperature during training	33.7 °C	33.4 °C	$\mathit{0.50}^{*}$
Mean skin temperature during recall	33.8 °C	33.4 °C	$\mathit{0.92}^{*}$
Maximum skin temperature during baseline	34.3 °C	33.4 °C	$\mathit{0.05}^{**}$
Maximum skin temperature during practice	33.7 °C	33.5 °C	$\mathit{0.23}^{*}$
Maximum skin temperature during training	33.8 °C	33.5 °C	$\mathit{0.33}^{*}$
Maximum skin temperature during recall	33.8 °C	33.5 °C	$\mathit{0.14}^{*}$
Baseline to practice deviation of min. skin temp.	0.7 °C	0.3 °C	$\mathit{0.42}^{*}$
Baseline to training deviation of min. skin temp.	0.9 °C	0.2 °C	$\mathit{0.71}^{*}$
Baseline to recall deviation of min. skin temp.	1.1 °C	0.3 °C	$\mathit{0.34}^{*}$
Baseline to practice deviation of mean skin temp.	0.2 °C	0.2 °C	$\mathit{0.91}^{*}$
Baseline to training deviation of mean skin temp.	0.2 °C	0.2 °C	$\mathit{0.64}^{**}$
Baseline to recall deviation of mean skin temp.	0.3 °C	0.2 °C	$\mathit{0.32}^{**}$
Baseline to practice deviation of max. skin temp.	-0.6 °C	0.0 °C	$\mathit{0.35}^{*}$
Baseline to training deviation of max. skin temp.	-0.5 °C	0.1 °C	$\mathit{0.38}^{*}$
Baseline to recall deviation of max. skin temp	-0.5 °C	0.0 °C	$\mathit{0.33}^{*}$
Minimum IBI during baseline	0.63 s	0.65 s	$\mathit{0.18}^{*}$
Minimum IBI during practice	0.65 s	0.70 s	$\mathit{0.24}^{*}$
Minimum IBI during training	0.63 s	0.66 s	$\mathit{0.34}^{**}$
Minimum IBI during recall	0.65 s	0.69 s	$\mathit{0.20}^{**}$
Mean IBI during baseline	0.67 s	0.70 s	$\mathit{0.63}^{**}$
Mean IBI during practice	0.66 s	0.70 s	$\mathit{0.38}^{**}$
Mean IBI during training	0.66 s	0.69 s	$\mathit{0.54}^{**}$

Mean IBI during recall	0.65 s	0.70 s	*0.24***
Maximum IBI during baseline	0.69 s	0.74 s	*0.14***
Maximum IBI during practice	0.67 s	0.70 s	*0.67***
Maximum IBI during training	0.69 s	0.73 s	*0.23***
Maximum IBI during recall	0.65 s	0.70 s	*0.17***
Baseline to practice deviation of minimum IBI	0.01 s	0.04 s	*0.05**
Baseline to training deviation of minimum IBI	0.00 s	0.01 s	*0.07**
Baseline to recall deviation of minimum IBI	0.02 s	0.04 s	*0.79***
Baseline to practice deviation of mean IBI	-0.01 s	0.00 s	*0.59***
Baseline to training deviation of mean IBI	-0.01 s	0.00 s	*0.33***
Baseline to recall deviation of mean IBI	-0.01 s	0.00 s	*0.35**
Baseline to practice deviation of maximum IBI	-0.03 s	-0.04 s	*0.18***
Baseline to training deviation of maximum IBI	-0.01 s	-0.02 s	*0.78***
Baseline to recall deviation of maximum IBI	-0.04 s	-0.04 s	*0.98**
Minimum EDA during baseline	3.7 µS	3.2 µS	*0.62**
Minimum EDA during practice	4.0 µS	3.3 µS	*0.74**
Minimum EDA during training	3.6 µS	3.1 µS	*0.83**
Minimum EDA during recall	3.9 µS	4.3 µS	*0.39**
Mean EDA during baseline	4.5 µS	3.9 µS	*0.59**
Mean EDA during practice	4.3 µS	3.8 µS	*0.72**
Mean EDA during training	4.1 µS	3.9 µS	*0.62**
Mean EDA during recall	4.2 µS	4.6 µS	*0.43**
Maximum EDA during baseline	5.7 µS	5.0 µS	*0.67**
Maximum EDA during practice	4.9 µS	4.3 µS	*0.77**
Maximum EDA during training	5.0 µS	5.0 µS	*0.44**
Maximum EDA during recall	4.8 µS	4.9 µS	*0.47**
Baseline to practice deviation of minimum EDA	0.3 µS	0.2 µS	*0.66**
Baseline to training deviation of minimum EDA	-0.1 µS	-0.1 µS	*0.18**
Baseline to recall deviation of minimum EDA	0.1 µS	1.1 µS	*0.95**
Baseline to practice deviation of mean EDA	-0.2 µS	-0.2 µS	*0.68**
Baseline to training deviation of mean EDA	-0.4 µS	-0.1 µS	*0.67**

Baseline to recall deviation of mean EDA	-0.3 µS	0.7 µS	0.71^{*}
Baseline to practice deviation of maximum EDA	-0.8 µS	-0.7 µS	0.97^{*}
Baseline to training deviation of maximum EDA	-0.7 µS	0.0 µS	0.56^{*}
Baseline to recall deviation of maximum EDA	-0.9 µS	-0.1 µS	0.92^{*}
Diastolic BVP during baseline	-435 nW	-331 nW	0.07^{*}
Diastolic BVP during practice	-320 nW	-313 nW	0.63^{*}
Diastolic BVP during training	-377 nW	-306 nW	0.15^{*}
Diastolic BVP during recall	-320 nW	-251 nW	0.07^{*}
Mean BVP during baseline	-0.03 nW	-0.02 nW	0.21^{*}
Mean BVP during practice	0.01 nW	0.02 nW	0.23^{*}
Mean BVP during training	0.03 nW	0.01 nW	0.44^{*}
Mean BVP during recall	0.01 nW	-0.17 nW	0.64^{*}
Systolic BVP during baseline	346 nW	296 nW	0.12^{*}
Systolic BVP during practice	273 nW	259 nW	0.28^{*}
Systolic BVP during training	315 nW	283 nW	0.44^{*}
Systolic BVP during recall	273 nW	249 nW	0.36^{*}
Baseline to practice deviation of diastolic BVP	115 nW	18 nW	0.09^{*}
Baseline to training deviation of diastolic BVP	58 nW	25 nW	0.44^{*}
Baseline to recall deviation of diastolic BVP	115 nW	80 nW	0.34^{*}
Baseline to practice deviation of mean BVP	0.04 nW	0.05 nW	0.19^{*}
Baseline to training deviation of mean BVP	0.06 nW	0.03 nW	0.99^{*}
Baseline to recall deviation of mean BVP	0.04 nW	-0.15 nW	0.55^{*}
Baseline to practice deviation of systolic BVP	-73 nW	-37 nW	0.11^{*}
Baseline to training deviation of systolic BVP	-31 nW	-13 nW	0.40^{*}
Baseline to recall deviation of systolic BVP	-73 nW	-48 nW	0.10^{*}
Minimum HR during baseline	73.6 bpm	76.2 bpm	0.37^{**}
Minimum HR during practice	80 bpm	78 bpm	0.56^{**}
Minimum HR during training	76 bpm	76 bpm	0.15^{*}
Minimum HR during recall	84 bpm	83 bpm	0.07^{*}
Mean HR during baseline	82 bpm	80 bpm	0.52^{**}
Mean HR during practice	83 bpm	81 bpm	0.53^{**}

Mean HR during training	84.16 bpm	84.19 bpm	0.51*
Mean HR during recall	87.3 bpm	87.2 bpm	0.87*
Maximum HR during baseline	92 bpm	85 bpm	0.03**
Maximum HR during practice	87 bpm	85 bpm	0.50**
Maximum HR during training	91 bpm	92 bpm	0.96*
Maximum HR during recall	90.4 bpm	90.5 bpm	0.95*
Baseline to practice deviation of minimum HR	6 bpm	2 bpm	0.09*
Baseline to training deviation of minimum HR	2 bpm	0 bpm	0.45**
Baseline to recall deviation of minimum HR	11 bpm	7 bpm	0.03*
Baseline to practice deviation of mean HR	0.9 bpm	1.0 bpm	0.99*
Baseline to training deviation of mean HR	1.7 bpm	3.8 bpm	0.47*
Baseline to recall deviation of mean HR	5 bpm	7 bpm	0.82*
Baseline to practice deviation of maximum HR	5 bpm	0 bpm	0.01**
Baseline to training deviation of maximum HR	-1 bpm	7 bpm	< 0.01*
Baseline to recall deviation of maximum HR	-1 bpm	6 bpm	0.04*

Appendix Q – Study 2 Facial Expressions and Significance

AU feature	TG	CG	SD	Result
Baseline AU10 NFEs	1.09/min	1.07/min	4.5σ	*0.41**
Baseline AU12 NFEs	1.6/min	2.2/min	4.2σ	*0.86**
Baseline AU14 NFEs	1.9/min	1.8/min	11.8σ	*0.22**
Baseline AU15 NFEs	3/min	2/min	11.5σ	*0.12**
Baseline AU17 NFEs	5/min	4/min	9.8σ	*0.06***
Baseline AU20 NFEs	2.2/min	1.7/min	11.4σ	*0.35**
Baseline AU23 NFEs	6/min	5/min	12.7σ	*0.74**
Baseline AU25 NFEs	3/min	2/min	4.1σ	*0.28**
Baseline AU26 NFEs	2/min	1/min	2.5σ	*0.92**
Baseline AU28 NFEs	0.034/min	0.033/min	0.3σ	*0.33**
Baseline AU10 MFEs	1/min	2/min	1.5σ	*0.77**
Baseline AU12 MFEs	2/min	3/min	3.5σ	*0.58**
Baseline AU14 MFEs	2.4/min	1.6/min	4.1σ	*0.68**
Baseline AU15 MFEs	3.1/min	2.8/min	3.7σ	*0.46**
Baseline AU17 MFEs	12.6/min	12.3/min	9.8σ	*0.06**
Baseline AU20 MFEs	5/min	3/min	3.8σ	*0.33**
Baseline AU23 MFEs	15/min	10/min	12.7σ	*0.07**
Baseline AU25 MFEs	3/min	2/min	3.5σ	*0.02**
Baseline AU26 MFEs	3/min	2/min	2.5σ	*0.21**
Baseline AU28 MFEs	0.05/min	0.10/min	4.5σ	*0.91**
Practice AU10 NFEs	0.8/min	0.7/min	1.8σ	*0.77**
Practice AU12 NFEs	3.1/min	2.9/min	3.8σ	*0.78**
Practice AU14 NFEs	0.9/min	0.8/min	1.5σ	*0.64**
Practice AU15 NFEs	1.7/min	2.4/min	2.8σ	*0.88**
Practice AU17 NFEs	4/min	5/min	5.2σ	*0.41**
Practice AU20 NFEs	3/min	4/min	4.5σ	*0.69**
Practice AU23 NFEs	7/min	8/min	6.4σ	*0.68**
Practice AU25 NFEs	7/min	6/min	5.7σ	*0.58**
Practice AU26 NFEs	4.2/min	4.0/min	4.6σ	*0.79**
Practice AU28 NFEs	0/min	0/min	0.0σ	*1.00**
Practice AU10 MFEs	0.9/min	0.8/min	2.0σ	*0.62**
Practice AU12 MFEs	5/min	4/min	5.3σ	*0.78**
Practice AU14 MFEs	0.9/min	0.6/min	1.2σ	*0.95**
Practice AU15 MFEs	3/min	5/min	4.6σ	*0.93**
Practice AU17 MFEs	9/min	15/min	11.7σ	*0.56**
Practice AU20 MFEs	7/min	9/min	9.1σ	*0.69**
Practice AU23 MFEs	17/min	12/min	15.6σ	*0.53**

Practice AU25 MFEs	8/min	6/min	6.6σ	*0.58**
Practice AU26 MFEs	4.2/min	4.0/min	4.1σ	*0.97**
Practice AU28 MFEs	0.084/min	0.085/min	0.4σ	*0.32**
Baseline to practice deviation of AU10 NFEs	-0.6/min	0.7/min	1.8σ	*< 0.01**
Baseline to practice deviation of AU12 NFEs	1/min	3/min	3.8σ	*0.01**
Baseline to practice deviation of AU14 NFEs	-1/min	1/min	1.8σ	*< 0.01**
Baseline to practice deviation of AU15 NFEs	-1/min	2/min	2.9σ	*< 0.01**
Baseline to practice deviation of AU17 NFEs	-1/min	5/min	6.4σ	*< 0.01***
Baseline to practice deviation of AU20 NFEs	1/min	4/min	5.1σ	*0.01**
Baseline to practice deviation of AU23 NFEs	1 /min	8/min	8.2σ	*< 0.01***
Baseline to practice deviation of AU25 NFEs	4/min	6/min	6.1σ	*0.07**
Baseline to practice deviation of AU26 NFEs	2/min	4/min	5.0σ	*0.08**
Baseline to practice deviation of AU28 NFEs	-0.1/min	0.0/min	0.1σ	*0.88**
Baseline to practice deviation of AU10 MFEs	-0.5/min	1.0/min	2.1σ	*< 0.01**
Baseline to practice deviation of AU12 MFEs	2/min	3/min	3.7σ	*0.03**
Baseline to practice deviation of AU14 MFEs	-1/min	1/min	2.4σ	*< 0.01**
Baseline to practice deviation of AU15 MFEs	-1/min	4/min	4.5σ	*< 0.01**
Baseline to practice deviation of AU17 MFEs	-1/min	15/min	17.0σ	*< 0.01**
Baseline to practice deviation of AU20 MFEs	2/min	9/min	1.0σ	*< 0.01**
Baseline to practice deviation of AU23 MFEs	3/min	12/min	14.7σ	*0.02**
Baseline to practice deviation of AU25 MFEs	4/min	6/min	6.8σ	*0.14**
Baseline to practice deviation of AU26 MFEs	2/min	4/min	4.4σ	*0.03**
Baseline to practice deviation of AU28 MFEs	-0.1/min	0.1/min	0.5σ	*0.01**
Training AU10 NFEs	0.7/min	0.8/min	0.9σ	*0.30**
Training AU12 NFEs	1/min	2/min	1.9σ	*0.16**
Training AU14 NFEs	1.0/min	0.8/min	1.4σ	*0.85**
Training AU15 NFEs	3/min	2/min	2.8σ	*0.75**
Training AU17 NFEs	6/min	4/min	3.9σ	*0.16**
Training AU20 NFEs	3.7/min	3.6/min	4.1σ	*0.98**
Training AU23 NFEs	6.6/min	7.1/min	5.4σ	*0.55**
Training AU25 NFEs	6/min	8/min	4.9σ	*0.26**
Training AU26 NFEs	3/min	4/min	4.4σ	*0.50**
Training AU28 NFEs	0.02/min	0.05/min	0.1σ	*0.45**
Training AU10 MFEs	1.0/min	1.1/min	1.2σ	*0.91**
Training AU12 MFEs	2.3/min	2.8/min	3.1σ	*0.30**
Training AU14 MFEs	1/min	2/min	1.4σ	*0.79**
Training AU15 MFEs	4.2/min	3.5/min	4.1σ	*0.25**
Training AU17 MFEs	13.6/min	14.4/min	9.6σ	*0.95**
Training AU20 MFEs	7/min	5/min	5.4σ	*0.20**
Training AU23 MFEs	11/min	13/min	11.5σ	*0.73**

Training AU25 MFEs	6.8/min	7.2/min	5.1σ	*0.63**
Training AU26 MFEs	3.9/min	4.3/min	3.5σ	*0.94**
Training AU28 MFEs	0.05/min	0.04/min	0.1σ	*0.71**
Baseline to training deviation of AU10 NFEs	-1/min	0/min	0.9σ	*0.43***
Baseline to training deviation of AU12 NFEs	-0.3/min	-0.0/min	1.8σ	*0.65**
Baseline to training deviation of AU14 NFEs	-0.2/min	-0.7/min	1.6σ	*0.23**
Baseline to training deviation of AU15NFEs	-0.3/min	0.0/min	2.2σ	*0.81**
Baseline to training deviation of AU17 NFEs	0.7/min	0.5/min	3.9σ	*0.87***
Baseline to training deviation of AU20 NFEs	1.5/min	1.9/min	3.8σ	*0.47**
Baseline to training deviation of AU23 NFEs	0/min	3/min	5.2σ	*0.07***
Baseline to training deviation of AU25 NFEs	3/min	6/min	5.2σ	*0.01***
Baseline to training deviation of AU26 NFEs	1/min	3/min	3.4σ	*0.08**
Baseline to training deviation of AU28 NFEs	-0.04/min	0.01/min	0.2σ	*0.20**
Baseline to training deviation of AU10 MFEs	-0.4/min	-0.3/min	2.0σ	*0.80**
Baseline to training deviation of AU12 MFEs	0.0/min	0.5/min	3.0σ	*0.52**
Baseline to training deviation of AU14 MFEs	-0.3/min	-0.5/min	2.1σ	*0.92**
Baseline to training deviation of AU15 MFEs	-0.3/min	0.4/min	4.3σ	*0.66**
Baseline to training deviation of AU17 MFEs	1/min	2/min	10.6σ	*0.70***
Baseline to training deviation of AU20 MFEs	1.8/min	2.4/min	5.8σ	*0.88**
Baseline to training deviation of AU23 MFEs	-4min	3/min	13.5σ	*0.05***
Baseline to training deviation of AU25 MFEs	2/min	5/min	5.4σ	*0.12**
Baseline to training deviation of AU26 MFEs	1/min	3/min	3.7σ	*0.13***
Baseline to training deviation of AU28 MFEs	0.0/min	-0.1/min	0.2σ	*0.14**
Recall AU10 NFEs	4/min	5/min	5.4σ	*0.54**
Recall AU12 NFEs	3.97/min	3.93/min	4.5σ	*0.51**
Recall AU14 NFEs	4/min	5/min	4.5σ	*0.39**
Recall AU15 NFEs	3.96/min	4.11/min	4.7σ	*0.77**
Recall AU17 NFEs	2.7/min	3.4/min	4.5σ	*0.54**
Recall AU20 NFEs	4.6/min	7.1/min	5.5σ	*0.03**
Recall AU23 NFEs	5.6/min	6.4/min	4.8σ	*0.33**
Recall AU25 NFEs	6/min	7/min	5.6σ	*0.88**
Recall AU26 NFEs	4/min	5/min	5.1σ	*0.51**
Recall AU28 NFEs	0.02/min	0.00/min	0.1σ	*0.33**
Recall AU10 MFEs	4/min	3/min	5.2σ	*0.55**
Recall AU12 MFEs	4/min	3/min	5.0σ	*0.55**
Recall AU14 MFEs	5/min	4/min	5.5σ	*0.43**
Recall AU15 MFEs	4/min	3/min	4.2σ	*0.85**
Recall AU17 MFEs	8/min	5/min	9.0σ	*0.13**
Recall AU20 MFEs	11/min	14/min	13.2σ	*0.73**
Recall AU23 MFEs	8.3/min	8.0/min	9.3σ	*0.81**

Recall AU25 MFEs	8/min	5/min	6.7σ	*0.02**
Recall AU26 MFEs	3.94/min	3.92/min	4.7σ	*0.71**
Recall AU28 MFEs	0.06/min	0.04/min	0.3σ	*0.98**
Baseline to recall deviation of AU10 NFEs	2/min	3/min	5.0σ	*0.35**
Baseline to recall deviation of AU12 NFEs	1.7/min	1.8/min	3.6σ	*0.72**
Baseline to recall deviation of AU14 NFEs	2.5/min	2.9/min	5.0σ	*0.75***
Baseline to recall deviation of AU15 NFEs	1.4/min	1.7/min	4.1σ	*0.51**
Baseline to recall deviation of AU17 NFEs	-1.7/min	0.2/min	6.1σ	*0.15**
Baseline to recall deviation of AU20 NFEs	3/min	5/min	5.8σ	*0.12**
Baseline to recall deviation of AU23 NFEs	-1/min	2/min	5.4σ	*0.12**
Baseline to recall deviation of AU25 NFEs	3/min	5/min	6.5σ	*0.15**
Baseline to recall deviation of AU26 NFEs	2/min	4/min	5.4σ	*0.22**
Baseline to recall deviation of AU28 NFEs	0.02/min	0.05/min	0.4σ	*0.87**
Baseline to recall deviation of AU10 MFEs	1.7/min	0.7/min	5.5σ	*0.23**
Baseline to recall deviation of AU12 MFEs	1/min	-1/min	5.3σ	*0.15***
Baseline to recall deviation of AU14 MFEs	5.5/min	-0.01/min	11.6σ	*0.16**
Baseline to recall deviation of AU15 MFEs	0.6/min	0.1/min	5.6σ	*0.91**
Baseline to recall deviation of AU17 MFEs	-4/min	-8/min	14.2σ	*0.31***
Baseline to recall deviation of AU20 MFEs	7/min	10/min	14.0σ	*0.58**
Baseline to recall deviation of AU23 MFEs	-6/min	1/min	15.4σ	*0.28**
Baseline to recall deviation of AU25 MFEs	3.5/min	2.3/min	6.9σ	*0.59**
Baseline to recall deviation of AU26 MFEs	2/min	3/min	5.5σ	*0.50**
Baseline to recall deviation of AU28 MFEs	-0.08/min	-0.09/min	0.4σ	*0.69**