# Towards a symmetrical definition of QoE: An Evaluation of Emotion Semantics in Augmented Reality Training


*Eoghan Hynes
*Dept. of computer and software engineering*
*Technological University of the Shannon, Midlands, and Midwest*
Dublin road, Athlone, Co. Westmeath, Ireland
e.hynes@research.ait.ie

Ronan Flynn
*Dept. of computer and software engineering*
*Technological University of the Shannon, Midlands, and Midwest*
Dublin road, Athlone, Co. Westmeath, Eire
ronan.Flynn@tus.ie

Brian Lee
*Software Research Institute*
*Technological University of the Shannon, Midlands, and Midwest*
Dublin road, Athlone, Co. Westmeath, Eire
blee@ait.ie

Niall Murray
*Dept. of computer and software engineering*
*Technological University of the Shannon, Midlands, and Midwest*
Dublin road, Athlone, Co. Westmeath, Eire
nmurray@research.ait.ie



***Abstract*—The current definition of quality of experience (QoE) designates delight and annoyance as diametrically opposing indicators of the degree of fulfilment of an application, service or system user's pragmatic and hedonic needs and expectations. However, these inherited emotion terms are rarely used to describe emotions of equal amounts of arousal or opposing amounts of valence in the literature. This work assesses the significance of this asymmetry to the definition of QoE by determining the utility of emotion terms to communicate the emotion component of QoE. This was done in the context of a QoE evaluation of augmented reality training instruction formats. Correlates were sought between various measures of emotional state. This included physiological ratings, facial expressions and eye gaze. Emotional state was subjectively reported using three distinct methods: self-assessment manikin questionnaire; 2D emotion space terms; and open-ended terms. Regression analysis showed multiple significant correlations between implicit and explicit metrics, but not to the emotion terms used by the participants. This calls into question the utility of such vaguely understood terms. The use of more symmetrically opposing emotions in the definition of QoE may benefit consensual interdisciplinary communication.**



***Keywords—quality of experience, emotion, augmented reality, procedure, training.***


## I. INTRODUCTION

The current definition of quality of experience (QoE) explicitly uses the terms delight and annoyance [1]. These terms are intended to represent diametrically opposing emotions in the definition of QoE. Research into the semantics of such terms rarely, if ever, attributes equal amounts of arousal or equally opposing amounts of valance to these emotions as per Fig. 1 [2]–[6]. Consensual understanding of the phenomenon under study is essential for interdisciplinary collaboration and advancement of the field of QoE research. The usage of different terms to represent emotion may allow users to communicate elements of QoE more intuitively. Evaluation of the utility of emotion terms to communicate elements of user QoE was the focus this work. This was done in the context of a QoE evaluation of augmented reality (AR) training instruction formats.

Evaluation of AR instruction formats is required to realise AR's potential as a training platform [7]. The experimental evaluation compared the influence of procedural and example instruction formats on AR trainee QoE [8]. A gender balanced sample of 60 participants [9] was procured by convenience sampling. Participants were trained on a GoCube™ [10] manipulation procedure having been assigned to one of two independent groups. The main test group (TG) used a combined text and interactive animated 3D model instruction format [11]. The control group (CG) used a text-only instruction format. The test framework recorded the participant's physiological ratings, facial expressions, and eye gaze. After the experience, the participants reported their emotional state using three methods. Firstly, they were asked to use any open-ended term of their choice to describe their post-experience emotional state. Secondly, they completed the self-assessment manikin (SAM) questionnaire [12]. Finally, they were asked to circle one label that best described their emotional state from the 2D space as per Fig. 1 [4].

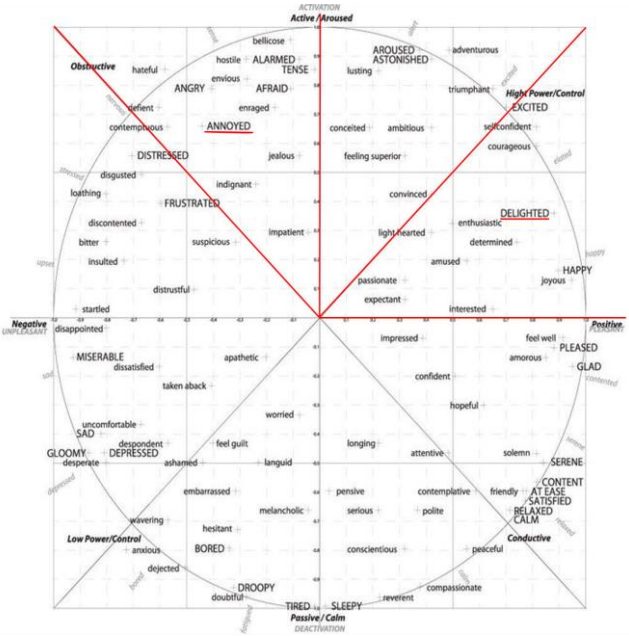


Fig.1. A comprehensively labeled 2D emotion space, adapted from [4] to highlight the asymmetrical levels of valance and arousal expressed in delight and annoyance (underlined).

Regression analysis was used to identify correlations between these different metrics. This analysis showed many significant correlations between physiological, objective and subjective metrics, but no significant correlation to the emotion terms used by the participants. This calls into question the consensual interpretation of such terms.

The remainder of this paper is structured as follows. The following related work section offers a critique of the literature relevant to this work. This is followed in Section III by a detailed description of the implemented methodology. Section IV outlines the post evaluation data analysis that was performed. A discussion of statistical results is given in Section V. Finally, the conclusions and future direction of this work are presented in Section VI.

## II. Related Work

In 1980 James A. Russell expanded on earlier works investigating the polarity of emotion semantics [2]. He used a sample of 36 participants to demonstrate that emotion terms fall meaningfully into a 2D circle in terms of positive and negative arousal and valance dimensions. The participants were instructed to position 28 emotion labels within a circle. The labels were sorted so that words at opposite sides of the circle described opposing emotions and those positioned close together were similar. Factor analysis produced the 2D emotion space containing several emotion labels positioned as in Fig. 1. A distance metric had a median correlation of $r = 0.80$ across the 36 participants and correlated to previously theorised positions with $r > 0.90$. Thus, Russell stated in [2] that the 2D circumplex model of emotion provided a convenient means for self-reporting the cognitive conceptualisation of emotion. Continued research into the 2D emotion space has resulted in the modern 2D graph used in our research as seen in Fig. 1, adapted from [4].

The SAM questionnaire was proposed by Peter J. Lang in 1985 [13] for measuring emotional response in terms of arousal, valence and dominance. This was done to simplify the complexities of Russell's semantic differential model (SDM), which was the state-of-the-art for recording explicit affect since 1974. The SDM consisted of 18 bipolar adjective pairs, each rated on a 9-point scale. Factor analysis of the scores on the three 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 was promisingly evaluated against the SDM in [13]. The authors demonstrated that the paper-based SAM questionnaires correlated with the SDM for valance, arousal and dominance with $r=0.97$, $r=0.94$ and $r=0.23$, respectively. For the two major affect dimensions (valance and arousal), SAM showed almost complete agreement with the far more complicated SDM.

Our evaluation of emotion semantics was undertaken in the context of a QoE evaluation of AR training instruction formats. Context aware interactive AR training applications can ensure correct learning by verification of instruction execution on the workpiece [14]. This stepwise verification automates self-pacing which can reduce trainee cognitive load [15]. However, cognitive load can also be impacted by the AR content including of instruction format [15]. Instructions can be presented in procedural and example formats [8]. 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. A graphical example can reduce extraneous cognitive load but may impact learning due to the development of over-dependence [8]. The cognitive effort involved in carrying out procedural instructions may benefit learning [8]. For these reasons, research is required in order to evaluate the influence of training instruction formats to fully realise AR's potential as a training platform [16]. The experimental methodology used to undertake this evaluation is described in detail in the following section.

## III. Experimental Methodology

### A. The AR training application

Participants received visuospatial aptitude training using an AR GoCube™ training application on the HoloLens 2 (HL2) AR headset [17]. The GoCube™ is an electronic version of the world-famous Rubik's Cube® puzzle. GoCube™ state was relayed to the HL2 over Wireless network. This Cube state information was used to automate instruction progression, to animate the Cube model (the example format, see Fig. 2), and to cue corrective instructions in the event of trainee errors. As such, the system provided a robust solution for the repeatable workpiece tracking required of the scientific method, whereas computer vision-based tracking has been shown to be subject to context influencing factors such as lighting, target object pose and occlusion [18]. The independent variable between the TG and CG was instruction format as described in the following sections.

### B. The example instruction format

The TG was trained using an interactive animated 3D model of the GoCube™ as an example instruction format as seen in Fig. 2. This was combined with text instructions as recommended in [11]. The model animated the direction of rotation of the Cube face for the given instruction. This included directional arrows indicating clockwise or anti-clockwise face rotation. The interactive animated 3D model represented the entire GoCube™ in its current configuration, as the trainee should be looking at it for the given instruction. The accompanying text instruction told the trainee what colour face to rotate and in what direction; either $90^{o}$ clockwise or anti-clockwise.

### C. The procedural instruction format

The CG was trained using the same suite of instructions as the TG. These instructions consisted of the text instruction only from Fig. 2. The text-based procedural instruction described the direction to rotate the Cube face for the given instruction. Magenta text colour was used to contrast the workpiece and the environment. Black is transparent by default in AR applications. The test environment was white with a wooden table. Magenta did not appear as a colour on the workpiece or in the test environment.

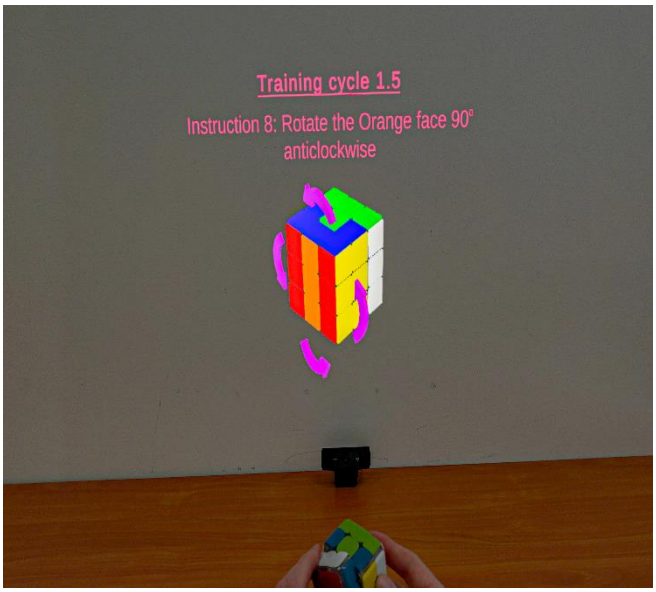

Fig. 2. The TG participant's view showing a combined text and interactive animated 3D model instruction format.

### D. Objective metric capture

The participant's manipulation of the GoCube™ was relayed to the HL2 over wireless network. This enabled real-time capture of instruction response times, quantity of errors, and durations. The participant's baseline and post-experience mental rotation abilities were recorded using the standard Vandenberg mental rotation test [19].

### E. Implicit metric capture

Our implicit data capture framework recorded eye gaze features, physiological ratings, and expressions of the lower face. The participant was seated at a table in the test laboratory where they were fitted with the HL2. The HL2 eye tracking sensors were calibrated to the participant's eyes. The eye gaze sensors on the HL2 are intended as an input medium, but their usage was adapted in this study to record user eye gaze to the HL2. The Empatica E4 sensor [20] was used to record the participant's skin temperature, blood volume pulse (BVP), heart rate (HR), interbeat interval (IBI) and electrodermal activity (EDA). The E4's photoplethysmography sensor measures BVP at 64Hz as the level of blood oxygenation measured in nano Watts (nW). HR is calculated from the BVP signal as beats per minute (bpm). IBI is calculated as the duration of time, in seconds (s), between consecutive diastolic peaks of equal amplitudes (±5 nW) in the BVP signal. EDA and peripheral skin temperature are both sampled at a rate of 4Hz. A desk-mounted C930s Logitech™ 1080p video camera [21] was used in conjunction with OpenFace facial recognition software [22] to record the participant's lower facial Action Units (AUs) [23].

### F. Explicit metric capture

Post-experience, the participants were asked to complete an affect questionnaire that consisted of three parts [24]. Firstly, participants were asked to describe their current emotional state using an emotion of their choice. Secondly, participants were asked to complete the SAM questionnaire. Finally, participants were asked to select a label from Russel's 2D emotion space [2] as per Fig. 1, that best described their post experience emotional state. They also reported their QoE, subjective task load and cognitive load in Likert scale and NASA-TLX questionnaires, respectively. The Likert scale questionnaire recorded aspects of interaction, efficiency, usability, aesthetics, utility and acceptability [25].

### G. Protocol

This evaluation consisted of ten distinct phases. These are shown in Table I. Each of these phases is described in the following paragraphs. The protocol adhered to the standards for test methodologies of head-mounted displays outlined in ITU-T P.919 [9]. This involved a self-paced between-groups study design conducted on a gender-balanced sample in a controlled testing environment, after appropriate informed consent, instruction and verification of understanding. The testing room provided control for consistent lighting conditions and noise pollution. This research was approved by the ethics committee of the Athlone Institute of Technology. Participant consent was obtained in written format. Consent and anonymised data were stored separately.

*1) Sampling and Information Sharing:* A gender-balanced sample of 60 participants was secured by convenience sampling. The sample group had an age range from 19 to 62 years old with a mean age of 32. Twelve nationalities were represented. The sample group was divided into two independent groups of 30, with an equal distribution of 15 males and 15 females. Each participant was greeted and thanked for their participation in the testing laboratory. They were then provided with a test information sheet. This explained the entire test procedure including that participants would undergo training in a GoCube™ manipulation procedure under one of the instruction conditions, and that they would be then required to recall the procedure as trained from memory. An opportunity was afforded to address any questions about the evaluation process, after which test subjects completed and signed a consent form. This information sharing phase lasted 2 minutes on average, and the signing of the consent form took 1 minute and 30 seconds on average. After giving written consent, the participant was fitted with the E4 sensor. This was done during this phase to allow the maximum time for the temperature heat flux sensor to acclimatise to the participant's skin temperature. Recording of physiological ratings began at this time. The participant then proceeded to the screening phase.

*2) Screening:* The participant was first screened for visual acuity using the standard Snellen eye test [26]. They were then screened for colour perception using digital Ishihara colour plates [27]. Following this, an interactive digital Vandenberg mental rotation test was implemented [19]. This provided a baseline of the participant's mental rotation abilities [28]. No participants were excluded during screening. The screening phase took an average of 4 minutes.

*3) Instruction and calibration:* The participant was introduced to the GoCube™ in terms of face colours and face rotation directions. The GoCube™ has six faces. Each Cube face is referenced by the tile at its centre. This is because each centre tile is bound to one face. Like the standard 3x3 Rubik's Cube®, the faces are coloured blue, green, white, yellow, red,

TABLE I. PROTOCOL PHASES

| Number | Phase |
|---|---|
| 1 | Sampling and information sharing |
| 2 | Screening |
| 3 | Instruction and calibration |
| 4 | Baseline |
| 5 | Practice |
| 6 | Training |
| 7 | Waiting |
| 8 | Recall |
| 9 | Transfer |
| 10 | Questionnaires |

and orange. The participant was instructed to rotate each Cube face in one of two ways by reference to Cube face colour. These were 90º clockwise rotation and 90º anti-clockwise rotation. The participant's understanding of this information was verified using a standard Rubik's Cube®. Then they were fitted with the HL2 which was calibrated to their eyes. This instruction and calibration phase took 4 minutes on average. The baseline phase then began.

*4) Baseline*: The start of the 5 minute baseline period was marked by the beginning of recording of eye gaze features (i.e., blink rate) using the HL2's eye tracking sensors. The raw data captured during this baseline phase was used to extract various physiological and physical baseline features as described in the following data analysis section. The deviation of these features from baseline during the task was used to create deviation features to reflect the influence of the instruction formats on the participant. After recording of baseline physiological ratings, facial AUs and eye gaze features had elapsed, the participant proceeded to the practice phase. The recording of these implicit QoE metrics continued throughout the evaluation and only ceased after the recall phase was complete.

*5) Practice:* The participant underwent a practice phase for their given instruction format. This involved following instructions to rotate each GoCube™ face 90º in both clockwise and anti-clockwise directions (i.e., 12 instructions). It had been verified that they could do this independently of the HL2 during the instruction and calibration phase. Now the goal was to verify that they could see, understand, and correctly follow instruction from the HL2. Corrective instructions were issued by the AR application in the event of user mistakes. Upon successful completion of all instructions, the participant automatically progressed to the training phase in which they were trained in a specific GoCube™ manipulation procedure. The practice phase was self-paced and lasted on average 1 minute and 9 seconds for the TG and 47 seconds for the CG.

*6) Training:* The training phase began with the GoCube™ in the solved state. Training was self-paced. Training cycles consisted of two halves, where the participant was required to action a suite of 7 instructions to and from the solved state. Total training time and number of errors were recorded to the HL2 as measures of the influence of the instruction formats on training. In the TG, the average training phase lasted 4 minutes and 36 seconds. In the CG the average training phase lasted 3 minutes and 51 seconds. The participant alerted the researcher once they were confident that they had learned the procedure as trained. The researcher then ended training by remote command to the HL2.

*7) Waiting:* As informed by the literature, a minimum of 20 seconds of workpiece-free waiting is sufficient to ensure that learned information has either been schematised into LM or retained in working memory (WM) *[29]* by means of repetition. 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 *[30]*. They performed arithmetic questions taken directly from, or adapted from *[31]* during this time. Correctly performing these equations requires WM resources, and any training not schematised to long term memory *[29]* will likely be lost during this process.

*8) Recall*: In the recall phase, the participant had to reproduce the GoCube™ manipulation procedure as trained. Number of errors, Cube face rotation intervals and total recall duration were recorded to the HL2. The recall phase was self-paced and lasted 46 seconds in the TG while in the CG it lasted 30 seconds.

*9) Transfer:* The participant re-took the standard Vandenberg mental rotations test as per baseline during the screening phase. This took 1 minute.

*10) Questionnaires:* Firstly, the participant was asked to write down an emotion that best described their emotional state. This took 20 seconds on average. They were then asked to complete the SAM questionnaire. This took 13 seconds on average. They were then asked to select one emotion label from Russel's 2D emotion space [4]. This took 48 seconds to perform on average. Following this, the participant completed a ten-statement five-point Likert scale questionnaire. This took 1 minute and 30 seconds to complete on average. There were two cognitive load questions taken from [32] on the questionnaire; one each to evaluate the amount of cognitive effort invested during the training phase and the recall phase. There were also three questions related to cognitive load (one each specific to intrinsic, extrinsic and germane cognitive load) invested during the training phase [33]. The participants then completed the NASA-TLX questionnaire, taking 3 minutes to complete on average.

## IV. DATA ANALYSIS

Post-evaluation data analysis involved time domain feature extraction from the captured data. Minimum, mean, and maximum features were extracted from the physiological data. Instruction gaze hits were used to calculate dwell times and gaze shift rates [34]. A fixation was calculated as a stationary gaze above 200ms [35]. Eye gaze below this threshold was excluded as a natural rapid eye movement known as a saccade. Gaze shift rate was normalised on a per minute basis [34]. Contiguous AU presence below 500ms was classified as a micro facial expression (MFE) [36]. Contiguous AU presence above this threshold was classified as a normal facial expression (NFE). The quantity of MFEs and NFEs recorded during varying practice, training and recall phase durations were normalised on a per-minute basis. Deviation from baseline to practice, training and recall of these features was calculated. Open-ended and 2D space emotion labels were assigned ordinal values for statistical analysis. Statistically significant differences were sought

between the independent groups to provide an insight into the influence of the different instruction formats on these features with 95% confidence level (α=0.05). Linear regression analysis between these features was undertaken to 99% confidence level (α=0.01).

## V. Results and Discussion

There was a significant difference between the genders in mental rotation baseline abilities (male: 9 rotations, female: 7 rotations) with $p = 0.04$ [37]. Instruction response times were significantly faster in the CG during practice and training with $p = 0.01$ and $p = 0.05$ respectively. Practice instruction response times are broken down by gender in Table II. In addition to Table II, CG males were faster than TG females with $p = 0.01$. Consequently, the practice phase was significantly shorter in the CG with $p = 0.04$. The CG made significantly less mistakes during the practice phase with $p < 0.01$. This result is broken down by gender in Table III. In addition to Table III, CG males made significantly less mistakes than TG females with $p = 0.04$. The CG's eye gaze dwelled significantly longer on the GoCube™ than the TG with $p = 0.03$. AU14 NFE deviation correlated to text instruction dwell in male participants with $R^2 = 0.21$, $p = 0.01$, while AU14 MFE deviation correlated to text instruction dwell in female participants with $R^2 = 0.23$, $p = 0.01$.

During recall, the CG's Cube face rotation intervals of 2.5 s were significantly shorter than the TGs of 3.4 s, with $p = 0.01$. This is broken down by gender in Table IV. In addition to the findings from Table IV, CG males were faster than TG females with $p = 0.05$. TG female recall Cube face rotation intervals correlated significantly to their training instruction response times with $R^2 = 0.46$, $p = 0.01$. However, the TG female's mental rotation baseline correlated significantly to both their training duration and recall Cube face rotation intervals with $R^2 = 0.18$, $p = 0.02$ and $R^2 = 0.13$, $p = 0.05$ respectively. This suggests that slower recall in TG females was not only influenced by training instruction format but also by their mental rotation abilities. This might suggest that female trainees might benefit from text-only instruction in terms of faster training [38], and possibly faster recall as a result. The CG female's maximum HR during recall of 90 bpm correlated to mental demand with $R^2 = 0.41$, $p = 0.01$.
Happy was the most common open-ended emotion term, being chosen by 17% of the sample, followed by excited at 13%. However, from the 2D emotion space, interested was the most chosen emotion term being chosen by 13% of the sample with happy in second place being chosen by 8%. Forty percent of the open-ended emotion terms chosen did not appear in the 2D space seen in Fig.1, with 10% of these not being regarded as emotion terms at all, perhaps due to language barriers. Once presented with the terms available in the 2D space seen in Fig. 1, only 35% of the sample whose open-ended term did appear in the 2D space persisted with their original choice. There were no significant correlations seen between either open-ended emotion terms or those chosen from the 2D emotion space to any of the other metrics captured during this evaluation, including SAM questionnaire responses.

TABLE II. MEAN PRACTICE INSTRUCTION RESPSONSE TIMES

| | **Male** | **Female** | ***p*** |
|---|---|---|---|
| **TG** | 5 s | 5 s | 0.84 |
| **CG** | 4 s | 5 s | 0.05 |
| ***p*** | 0.01 | 0.20 | |

TABLE III. MEAN PRACTICE PHASE ERRORS

| | **Male** | **Female** | ***p*** |
|---|---|---|---|
| **TG** | 0 | 1 | 0.25 |
| **CG** | 0 | 0 | 1.00 |
| ***p*** | 0.24 | 0.03 | |

TABLE IV. RECALL CUBE FACE ROTATION INTERVALS

| | **Male** | **Female** | ***p*** |
|---|---|---|---|
| **TG** | 3 s | 4 s | 0.27 |
| **CG** | 2 s | 3 s | 0.22 |
| ***p*** | 0.44 | 0.01 | |

SAM valance correlated to the rank given to NASA-TLX frustration with $R^2 = 0.24$, $p < 0.01$, while SAM dominance corelated to the rank given to performance and overall task load with $R^2 = 0.52$, $p < 0.01$ and $R^2 = 0.38$, $p < 0.01$ respectively, across both test groups. When SAM arousal (female score of 1.3, male score of 1.0, $p = 0.36$), valance (female score of 2.9, male score of 2.4, $p = 0.11$) and dominance (female score of 1.3, male score of 1.6, $p = 0.56$) results were combined into ordinal values, they correlated to deviation of AU15 and AU17 MFEs in females with $R^2 = 0.18$, $p = 0.02$ and $R^2 = 0.35$, $p < 0.01$ respectively. Ordinal SAM results also correlated to gender with $R^2 = 0.27$, $p < 0.01$. This may suggest that greater utility can be derived from consideration of emotion in terms of its constituent dimensions rather than using rather vaguely understood terms.

In summary, mental rotation abilities and training instruction response times correlated to recall Cube interaction speed in females of the TG. This suggests that female trainees may benefit from text-only instruction in terms of speed of training [38]. This in turn may even influence speed of recall from memory. There were various significant correlations seen between physiological and physical manifestations of emotion, task performance and subjective experience. However, there were no significant correlations between any of these metrics and the emotion labels chosen by the participants using either open-ended or 2D emotion space terms. This might call into question the consensual interpretation of such emotion terms. SAM questionnaire responses correlated significantly to facial expressions and gender suggesting utility for communicating emotional state.

## VI. Conclusions

This paper evaluated the significance of emotion semantics to the participants of a QoE evaluation of procedural and example instruction formats in an AR training application. Female trainees that used a text-only instruction format performed training and recall significantly faster than females that used the example format. Various significant correlations were seen between physiological and facial manifestations of emotion, objective performance and subjective experience. However, the absence of significant

correlations to the emotion designations used by the participants should call into question the utility of such vaguely understood terms. QoE requires a consistent and precise definition for the consensual development of measurement instruments and for interdisciplinary communication and collaboration. Emotion terms of utility should correlate to measurable physiological, physical or cognitive experiences of emotion. Future work will focus on identifying such correlations in various contexts. The strength of such correlations will guide meaningful discussion on the best terms to communicate the emotion component of QoE.

## ACKNOWLEDGMENT

The authors acknowledge the financial support of the Technological University of the Shannon President's Doctoral Scholarship and the Horizon Europe TRANSMIXR project (grant number: 101070109).

## REFERENCES

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

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] 'ITU-T P.919 : Subjective test methodologies for 360o video on head-mounted displays'. bit.ly/3KxfU2B (accessed Nov. 07, 2022).

[10] Particula, GoCube. https://tinyurl.com/yck4c78r (accessed Feb. 15, 2023).

[11] 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.

[12] J. D. Morris, 'Observations: SAM: The Self-Assessment Manikin An Efficient Cross-Cultural Measurement Of Emotional Response', J. Advert. Res., pp. 63–68, 1995.

[13] 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.

[14] C. Liu et al., '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.

[15] 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.

[16] S. Webel et al., '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.

[17] G. Evans et al., 'Evaluating the Microsoft HoloLens through an augmented reality assembly application', presented at the SPIE Defense + Security, Anaheim, California, United States: Degraded environments: sensing, processing, and display. Vol. 10197. International Society for Optics and Photonics, May 2017.

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

[19] S. G. Vandenberg, Allan R. Kuse, 'Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization, 1978.

[20] C. McCarthy et al., '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.

[21] 'C930s Pro HD Webcam - Logitech'. https://tinyurl.com/4fwzcb2e (accessed Feb. 15, 2023).

[22] 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.

[23] G. Donato et al., 'Classifying facial actions', IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 10, pp. 974–989, Oct. 1999.

[24] E. Hynes, 'Affect Questionnaire'. https://osf.io/4cgyd (accessed Mar. 30, 2023).

[25] I. Wechsung et al., '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.

[26] 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.

[27] Colblindor, 'Digital color blindness test', Ishihara plates, 2006. bit.ly/3M6wkjB (accessed Dec. 02, 2020).

[28] L. Hou et al., '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.

[29] 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.

[30] 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.

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

[32] F. Paas, P. Ayres, and M. Pachman, 'Assessment of Cognitive LoAd in muLtimediA LeArning theory, methods and Applications', 2008.

[33] J. Leppink et al., '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.

[34] 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.

[35] 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.

[36] W.-J. Yan et al., '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.

[37] 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.

[38] [1] S. Möller and A. Raake, Quality of Experience, Advanced Concepts, Applications and Methods. Springer, 2013.

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

[3] 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.

[4] 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.

[5] 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.

[6] 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.

[7] 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.

[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] 'ITU-T P.919 : Subjective test methodologies for 360o video on head-mounted displays'. bit.ly/3KxfU2B (accessed Nov. 07, 2022).

[10] Particula, GoCube. https://tinyurl.com/yck4c78r (accessed Feb. 15, 2023).

[11] 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.

[12] J. D. Morris, 'Observations: SAM: The Self-Assessment Manikin An Efficient Cross-Cultural Measurement Of Emotional Response', J. Advert. Res., pp. 63–68, 1995.

[13] 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.

[14] C. Liu et al., '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.

[15] 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.

[16] S. Webel et al., '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.

[17] G. Evans et al., 'Evaluating the Microsoft HoloLens through an augmented reality assembly application', presented at the SPIE Defense + Security, Anaheim, California, United States: Degraded environments: sensing, processing, and display. Vol. 10197. International Society for Optics and Photonics, May 2017.

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

[19] S. G. Vandenberg, Allan R. Kuse, 'Mental Rotations, a Group Test of Three-Dimensional Spatial Visualization, 1978.

[20] C. McCarthy et al., '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.

[21] 'C930s Pro HD Webcam - Logitech'. https://tinyurl.com/4fwzcb2e (accessed Feb. 15, 2023).

[22] 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.

[23] G. Donato et al., 'Classifying facial actions', IEEE Trans. Pattern Anal. Mach. Intell., vol. 21, no. 10, pp. 974–989, Oct. 1999.

[24] E. Hynes, 'Affect Questionnaire'. https://osf.io/4cgyd (accessed Mar. 30, 2023).

[25] I. Wechsung et al., '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.

[26] 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.

[27] Colblindor, 'Digital color blindness test', Ishihara plates, 2006. bit.ly/3M6wkjB (accessed Dec. 02, 2020).

[28] L. Hou et al., '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.

[29] 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.

[30] 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.

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

[32] F. Paas, P. Ayres, and M. Pachman, 'Assessment of Cognitive LoAd in muLtimediA LeArning theory, methods and Applications', 2008.

[33] J. Leppink et al., '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.

[34] 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.

[35] 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.

[36] W.-J. Yan et al., '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.

[37] 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.

[38] A. Duenser et al., Virtual and Augmented Reality as Spatial Ability Training Tools, vol. 158. 2006. doi: 10.1145/1152760.1152776.