Pre-publication versions of the papers

British HCI Conference 2021

Post-Pandemic HCI – Living digitally

Day 1: Tuesday 20 July 2021

Warning and apology: Please note that the formatting of papers is not perfect in this document and does not reflect how they will appear in the final conference proceedings!

Table of Contents

Paper session 1: VR/AR/XR. Session chair: Effie Law

Evaluating Visual Variables in a Virtual Reality Environment

Somnath Arjun, G S Rajshekar Reddy, Abhishek Mukhopadhyay, Sanjana Vinod, Pradipta Biswas

I^3D Lab, Indian Institute of Science Bangalore 560012, India

{somnatharjun, rajshekarg, abhishekmukh, sanjanam, pradipta}@iisc.ac.in

Large amount of multi-dimensional data can be difficult to visualize in standard 2D display. Virtual Reality and the associated 3rd dimension may be useful for data analysis; however, 3D charts may often be confusing to users rather conveying information. This paper investigated and evaluated graphical primitives of 3D charts in a Virtual Reality (VR) environment. We compared six different 3D graphs involving two graph types and five visual variables. We analysed ocular and EEG parameters of users while they undertook representative data interpretation tasks using 3D graphs. Our analysis found significant differences in fixation rate, alpha and low-beta EEG bands among different graphs and a bar chart using different sizes of columns for different data values found to be preferred among users in terms of correct response. We also found that colour makes it easier to interpret nominal data as compared to shape and size variable reduces the time required for processing numerical data as compared to orientation or opacity. Our results can be used to develop 3D sensor dashboard and visualization techniques for VR environments.

Evaluation. Visual variables. Visualization. Virtual Reality. Eye Tracking. Cognitive load.

1. Introduction

Analysing data is turning increasingly difficult as the size and complexity of datasets continue to grow every day. Using visualisation techniques for data analysis is a popular method because it exploits the human visual system as a means of communication for interpreting information. In recent times, a plethora of visualisation techniques have been developed to explore large and complex data. The rise of visualisation techniques has made the practice of evaluation of visualization techniques even more critical. A number of empirical evaluation methods for visualisation techniques have been developed in the last two decades. There has been a steady increase in evaluation methods those include human participants' performances and subjective feedback. Isenberg et al. [1] divided evaluation methods into eight categories. They reported that Qualitative Result Inspection, Algorithmic Performance, User Experience and User Performance are the most common evaluation scenarios. In this paper, we have evaluated 3D visualisation in a VR environment by comparing user performance and experience across six types of visualisation techniques. We have investigated and compared visualisations using ocular parameters and EEG (Electroencephalogram). Ocular parameters are already extensively used to explain and model visual perception [39], analyse cognitive load [18, 20, 21] and areas of interest in complex visual stimuli [22, 23]. Comparison of 2D graphs used for representing quantitative data using eye tracking device has been undertaken for evaluating user/task characteristics and finding appropriate graphs [2,3]. Drogemuller et al. [42] evaluated navigation techniques for 3D graph visualisations in VR environment. Ware and Mitchell [33] studied graph visualisation in 3D, specifically they compared 3D tubes with 2D lines to display the links in a graph. They reported that with motion and stereoscopic depth cues, skilled observers could identify paths in a 1000-node graph with an error rate less than 10% compared to 28% with 2D graphs. Although tools and techniques have been developed in a VR environment for exploring and interacting with graphs effortlessly [4,5,6], researchers have hardly explored studies that compare 3D graphs. A comparative survey of user experiences with 3D charts in a VR environment was undertaken in [7,8]. However, these studies were primarily limited to a single graph.

Visualisation can be termed as a collection of graphical objects. Ward et al. [9] state that there are eight ways in which graphical objects can encode information, i.e., eight visual variables – position, shape, size, opacity, colour, orientation, texture, and motion. These eight variables can be adjusted as necessary to maximise the effectiveness of a visualisation to convey information. Garlandini and Fabrikant [35] explored the effectiveness and efficiency of these visual variables in 2D cartography. Their results revealed that the variable size was most effective and efficient in guiding viewers, and orientation played the least role. However, researchers have not investigated and compared visual variables in 3D graphs previously. We consider the problem of comparing visual variables of 3D graphs for representing 1D numerical data. In particular, we compare variables that are used to depict numerical data - size, orientation and opacity and nominal data - colour and shape. Somnath [3] compared 2D graphs and reported that users are more comfortable using bar and area charts than line and radar charts. Extending the work in the VR environment, we also compare the 3D bar chart and area chart to find if there is any difference between them in the VR environment. The readers may be interested in knowing:

1. Which 3D graph is best in terms of correct data interpretation?
2. Which visual variable(s) is (are) easier to interpret than others?
3. Does the 3rd dimension add a value?
4. Are there differences among graph types with respect to-?
5. Ocular parameters and
6. Cognitive load while interpreting graphs.

The paper is organised as follows. We discuss the related work in Section 2 followed by user study in Section 3. Methodology is discussed in Section 4, analysis and results are discussed in Section 5 followed by discussion in Section 6. We have presented concluding remarks in Section 7.

2. Related Work

Visualisation is defined as the communication of information using graphical representations. Graphics related application demands in depth understanding of graphics primitives and their properties to communicate information. In total, there are eight ways in which graphical objects can encode information [9]. Variables such as size, orientation, and opacity [9, 36] encode quantitative data information, while colour and shape are used for visualising nominal data. Fisher et al. [34] investigated which 3D graph type was easiest to interpret among bar, pie, floating line, mixed bar/line, and layered line charts. It was revealed that information extracted from bar and pie charts were found to be more effective than others. Additionally, it was found that the participants had better information retention with pie charts than bar charts. Hitherto, researchers have either developed new methods or discussed in detail how specific approaches need to be extended for visualisation evaluation. Evaluation of visualisation is primarily based on empirical methods. In particular, empiric evaluation and the consideration of human factors are discussed in [10,11,12]. Isenberg et al. [1] identified eight evaluation scenarios. They reported that Qualitative Result Inspection (QRI), Algorithmic Performance (AP), User Experience (UE) and User Performance (UP) to be the most common evaluation scenarios. In User Performance evaluation, Livingston et al. [13] focused on time taken and errors committed to complete a task using a new technique [13]. It was found that a large number of UP studies were done with 10-15 participants [1]. Evaluation of visualisation using an eye-tracking device [3] is an example of a UP evaluation scenario. Understanding user performances and feedback includes tasks where the user must answer a set of questions after assessing the visualisation techniques [2,3]. A set of low-level analysis tasks that capture user's activities while employing visualisation for understanding data was presented in [14]. We have adopted four out of these ten analytical task questions [14] for our user study.

Cognitive measures also have an influence on a user's performance and satisfaction while working with visualisations [15, 16, 17]. Peck et al. [32] utilized fNIRS to examine how participants process bar graphs and pie charts, and cognitive loads associated with them. Their results indicated that there was no significant difference among bar graph and pie chart, and this result also correlated with the results of the NASA TLX questionnaire. Furthermore, psychologists [19] have reported a strong association between cognitive load and pupil dilation of eyes. Marshall [20] proposed a wavelet-based algorithm to detect a hike in pupil dilation corresponding to an increase in cognitive load. Gavas [21] and Duchowski [22] also estimated cognitive load from pupil dilation. Saccadic Intrusion, change in fixation duration, and blink count [23] are also used for measuring cognitive load. Prabhakar et al. [18] investigated the efficacy of various ocular parameters to estimate cognitive load and detect driver's cognitive state. They derived gaze and pupil-based metrics and proposed a machine learning model classifying different levels of cognitive states. The use of ocular parameters has also shown an impact on evaluating visualisation performance [3,30,31]. A comparative study on user experiences with 3D graphs in VR environments was undertaken in [7,8]. There are no studies reported in literature which considers ocular parameters while the user observes different visualisation techniques in a VR environment. Gaze fixations are used for identifying areas of interest in graphs [3]. Research has been conducted on identifying user gaze differences for alternative visualisations [24], task types [25] or individual user differences [26]. In [24], linear and radial versions of bar, line, area, and scatter graphs were evaluated in terms of the cognitive load induced. It was revealed that participants took more time to complete tasks with the radial versions than their linear counterparts. It was also concluded that radial graphs are most useful for finding extreme values. In this work, we investigated ocular parameters like fixation rate, saccade rate and revisit sequences while users undertake tasks in VR environment. We also investigated pupil dilation and EEG data to estimate cognitive load of participants.

3. User study

In order to investigate and compare visual variables and charts, we designed and conducted a user study with six types of visualisation techniques. Each technique displayed numerical data and nominal data using different combinations of visual variables. We considered synthetic sensor data in our study and used five different sensors: temperature, humidity, smoke, air, and light. We have three instances of each sensor, and we use the term “node” to refer to all instances of a particular type of sensor. The data type of node was nominal. In total, there are 15 data points and 5 sensor nodes. The six visualisation techniques are explained next.

3.1 Visualisation charts

We developed and used six types of charts in our study, bar-size/bar chart (BC), bar-orientation (BOR), bar-opacity (BO), shape-size (SS), shape-opacity (SO) and area chart (AC). Nodes were arranged on the x-axis and instances of each node were arranged in the z-axis for all six charts. The representation of node and real valued sensor for each chart are described next.

3.1.1. Bar-Size chart

In this technique, the nodes are represented by different colours, and the size of bars depicts a numerical value, as shown in Figure 1. The size of bars is scaled along the y-axis. The scaled value of sensor is computed using -

$$\boldsymbol{SV_{Sensor}} = \frac{\boldsymbol{RV_{Sensor}} - \boldsymbol{S_{Min}}}{\boldsymbol{S_{Max}} - \boldsymbol{S_{Min}}} * \mathbf{10},$$

where $\boldsymbol{SV_{Sensor}}$ is the scaled value of sensor (length of the bar), $\boldsymbol{RV_{Sensor}}$ is the real value of sensor, $\boldsymbol{S_{Min}}$ is the minimum value of the sensor and $\boldsymbol{S_{Max}}$ is the maximum value of the sensor.

3.1.2. Bar-Orientation chart

As before nodes are represented by different colours but numerical values of sensors are defined by the orientation of bars. Bars are oriented or rotated along the x-axis to display values of sensors. The rotation is computed using -

$$\boldsymbol{SV_{Sensor}} = \left(\frac{\boldsymbol{RV_{Sensor}} - \boldsymbol{S_{Min}}}{\boldsymbol{S_{Max}} - \boldsymbol{S_{Min}}} * \mathbf{180}\right) - \mathbf{90},$$

where $\boldsymbol{SV_{Sensor}}$ is the scaled value of the sensor (rotation of the bar), $\boldsymbol{RV_{Sensor}}$ is the real value of the sensor, $\boldsymbol{S_{Min}}$ and $\boldsymbol{S_{Max}}$ were defined as before.

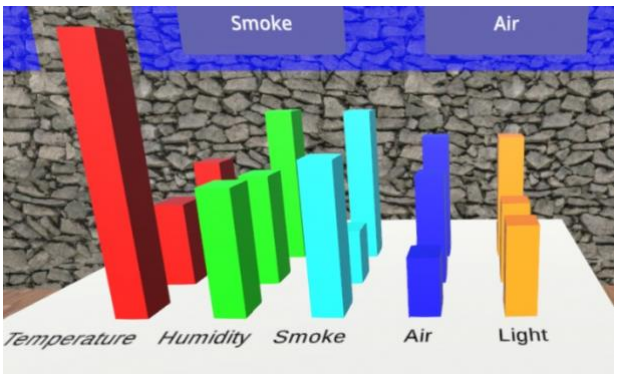


Figure 1: Bar-Size chart

3.1.3. Bar-Opacity chart

Nodes are represented by the unique bar colours, and the opacity of bars is directly proportional to the numerical value of sensors. The darker the bars, the more its value. The real value of the sensor is mapped to the opacity of the bar using the following equation -

$$\boldsymbol{SV_{Sensor}} = \frac{\boldsymbol{RV_{Sensor}} - \boldsymbol{S_{Min}}}{\boldsymbol{S_{Max}} - \boldsymbol{S_{Min}}} * \mathbf{255},$$

where $\boldsymbol{SV_{Sensor}}$ is the scaled value of sensor (opacity of bar), $\boldsymbol{RV_{Sensor}}$ is real value of sensor, $\boldsymbol{S_{Min}}$ and $\boldsymbol{S_{Max}}$ were defined as before.

3.1.4. Shape-Size chart

This visualisation technique uses a combination of shape and colour to define a node. The numerical values of the sensors are represented by the volume of the shape. The relation between sensor values and scaled values in VR environment follows the equation given below.

$$\boldsymbol{SV_{Sensor}} = \frac{\boldsymbol{RV_{Sensor}} - \boldsymbol{S_{Min}}}{\boldsymbol{S_{Max}} - \boldsymbol{S_{Min}}} * \mathbf{10},$$

where $\boldsymbol{SV_{Sensor}}$ is scaled value of sensor (size of shape), $\boldsymbol{RV_{Sensor}}$ is real value of sensor, $\boldsymbol{S_{Min}}$ and $\boldsymbol{S_{Max}}$ were defined as before.

3.1.5. Shape-Opacity chart

In this technique, nodes are represented by a combination of shape and colour. Numerical values are defined by the opacity of the shape, as shown in figure 2. The real value of the sensor is mapped to the opacity of the bar using the following equation.

$$SV_{Sensor} = \frac{RV_{Sensor} - S_{Min}}{S_{Max} - S_{Min}} * 255,$$

where SV_{Sensor} is scaled value of sensor (opacity of shape), RV_{Sensor} is real value of sensor, S_{Min} and S_{Max} were defined as before.

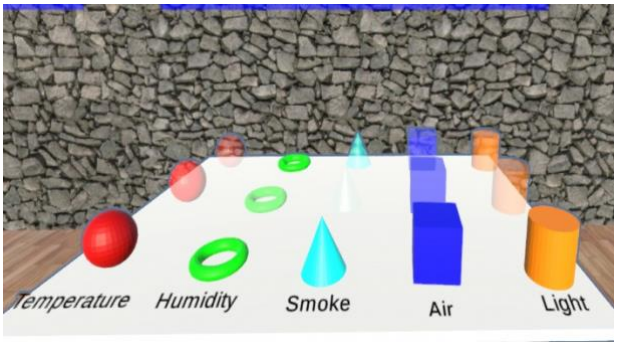


Figure 2: Shape-Opacity chart

3.1.6. Area chart

Sensors are represented by planes in the chart and each sensor has a unique colour. The values of sensors are depicted by the peaks of planes. Each plane is scaled along the y-axis. Relation between sensor value and peak of the plane is given by the following equation.

$$\boldsymbol{SV_{Sensor}} = \frac{\boldsymbol{RV_{Sensor}} - \boldsymbol{S_{Min}}}{\boldsymbol{S_{Max}} - \boldsymbol{S_{Min}}} * \mathbf{10}.$$

3.2 Materials

We used htc vive pro eye [37] with an inbuilt eye-tracker and refresh rate of 90hz to collect gaze-based data and pupil diameter (accuracy 0.5º of visual angle). We have also used emotiv insight eeg tracker [38] with 5 dry electrodes and sampling rate of 128 samples per second (sps) to collect eeg data. Our computer architecture consists of an intel core i5 processor and nvidia 2070 graphics card.

3.3 Participants

We collected data from 17 participants with an average age of 28 years (male:15 and female: 2) recruited from our university. We took appropriate ethical approval from university ethics committee for conducting the experiment. Participants were tested for visual acuity and all had 20/20 vision.

3.4 Design

We designed and set up a VR environment scene using the unity 3d game engine. The scene consists of a visualisation chart and a set of 4 questions. The VR scene is shown in figure 3. We set questions based on low-level tasks by Amar et al. [14]. The four questions that participants were requested to answer are explained below.

Q1: which node has the highest range?

Participants were asked to compare ranges of five sensor nodes and report the highest value among them.

Q2: find the node with the maximum and minimum average values?

Participants were first asked to guess the average value of each sensor node across its three instances. From these five estimated average values of five sensor nodes, participants were requested to report the sensor node with the maximum and minimum value. The process involves first browsing through y and z-axes to guess average and then comparison across x-axis.

Q3: which sensor has its average value nearest to humidity sensor?

Participants were asked to approximate the average value of each sensor as before. We then requested them to report the sensor node whose average value is closest to the average value of the humidity sensor.

Q4: sort the average values of sensors in descending order.

After estimating each sensor's average value as before, participants were asked to sort those values in descending order.

For example, in Figure 1, the temperature sensor has the highest difference between the maximum and the minimum value (range). After estimating the average value of all sensors, we can notice that the temperature sensor has the maximum average value, and the smoke sensor has the minimum average value. The air sensor's average value is closest to the average value of the humidity sensor. It may be noted that although sensors measure different physical variables, but their values were normalized in the rendering.

Figure 3: Virtual Reality scene

3.5. Procedure

Initially, participants were tested for their visual acuity and allowed the trial if they had 20/20 vision. Then they were briefed about the aim of the study and shown a virtual walkthrough of the environment. We calibrated the hand controller and eye tracker for each participant separately and proceeded with the trail when they could select the target, and the proprietary eye-tracking software indicated the calibration to be successful. We instructed participants to use the VR headset for ten minutes to get accustomed to the VR scene. Participants were instructed to move around the scene using a teleport button on the VR hand controller. When participants were comfortable with the scene, we asked them to start the task by wearing both EEG tracker and HTC Vive Pro Eye. Participants were then requested to observe the visualisation chart and answer four questions.

4. Analysis methodology

This section describes different algorithms used for calculating gaze-based metrics and cognitive load from ocular parameters. We calculated fixation rate, saccade rate and revisit sequences from eye gaze points. We also filtered the pupil dilation signal from the eye tracker using a low pass filter. The algorithms to calculate these metrics are described in the following sections.

4.1 Fixation and saccade rate

We calculated fixation rate and saccade rate by detecting fixations and saccades from gaze direction data using the velocity threshold fixation identification method (I-VT) [29]. I-VT is a velocity-based method that separates fixation and saccade points based on their point-to-point velocities. I-VT then classifies each point as a fixation or saccade based on a simple velocity threshold. If the point's velocity is below the threshold, it becomes a fixation point, otherwise it becomes a saccade point. We then calculated fixation and saccade rate as the number of fixations and saccades per second [18]. We calculated velocity in terms of visual angle i.e., degrees per second in order to render gaze velocity independent of the image and screen resolutions. This calculation is based on the relationship between the eye position in 3D space in relation to the stimuli plane and the gaze positions on the stimuli plane. The angle is calculated by taking the direction vector of two consecutive sample gaze points. The angle is then divided by the time between the two samples to get the angular velocity. The velocity threshold parameter is set to 40^{o}/sec [27]. The pseudo code for the I-VT method is shown in Table 1.

Table 1: Pseudocode for the I-VT algorithm

Calculate the angle between two consecutive points.
Calculate angular velocity by dividing the angle with the time between the two sample points.
Label each point below velocity threshold as a fixation and others as a saccade.
Return fixations and saccades.

4.2 Revisit sequences

A sequence refers to an ordered collection of focused nodes without repetitions. For example, A-B-C is a sequence, but A-A-B-C-C-C is not a sequence, where A, B and C are focused nodes. Revisit sequences provide information about how many times a participant scanned through a sequence [28]. This metric allows us to examine graphs that were repeatedly observed. We investigated three types of revisit sequences – sequences of lengths 3, 4 and 5. We also analysed two parameters of revisit sequences: (i) number of unique sequence and (ii) total revisit sequences. Unique revisit sequence is the distinct sequence for one graph that repeats itself. Total revisit sequences calculate all repetitions of every unique sequence. For the sequence of length three, if repetition is more than 3, the sequence is valid. For the sequence of length four, if repetition is more than 2, the sequence is valid. We did not consider revisits of the sequence of length five as there were less revisits for graphs. Figure 4 shows two unique sequences of length 3.

S1: Temperature – Humidity – Smoke

S2: Smoke – Air – Light

The pseudo code for the revisit sequence is shown in Table 2.

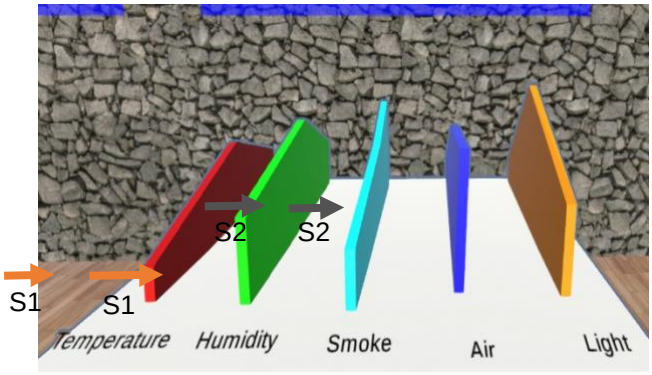


Figure 4: Two sequences of length 3

Table 2: Pseudocode for finding revisit sequences.

Find all nodes of the graph that participants were interested in and represent them into an array of sequential nodes.
Create a new sequence by the elimination of repetitive node placed in succession in a sequence.
Find unique sequences of length 3, 4 and 5 from the newly created sequence.
Calculate repetitions for each unique sequence.
Return number of unique sequences and the total number of revisits

4.3 Low pass filter of pupil (LPF)

Sudden hike in pupil dilation is related with change in cognitive load [20]. We divided the pupil dilation data into sections of 100 samples and subtracted mean from the raw data. We used a Butterworth lowpass filter with a cut off frequency of 5 Hz [40] and added the magnitude of the filtered data using a running window of size 1-sec with 70% overlap. This algorithm uses a conventional filtering technique in Digital Signal Processing (DSP), which uses time domain difference equations to filter the signal.

4.4 EEG data

We used EmotivBCI software [38] to monitor EEG signals and recorded data streams from EEG headset. The EmotivBCI software automatically calculates power signal for five EEG bands, we considered alpha, low beta, high beta and theta bands. We removed outlier from raw EEG data using inner fence.

5. Results

For all analyses, we calculated average values of parameters from all responses for all participants. We prepared tables of 6 columns corresponding to each type of graph and 17 rows corresponding to 17 participants. In all subsequent column graphs, the size of the column indicates average value while the error bar indicates standard deviation. We drew outline rectangles over columns which are statistically significantly different from each other.

We analysed the percentage of correct responses from the user for each chart. We analysed gaze-based metrics like fixation rate and saccade rate. We then calculated two parameters of revisit sequences and processed EEG data for further analysis. We analysed these parameters statistically for all participants across six charts. For statistical analysis we first undertook a Kolmogorov-Smirnov test for normality check. We then undertook Friedman test if data were not normally distributed. The following subsections explain each parameter used for the analysis and results.

5.1 User responses

The percentage of correct answers for each chart is calculated as

$$\boldsymbol{percentage} = \frac{\boldsymbol{number\ of\ correct\ answers}}{\boldsymbol{total\ number\ of\ questions}} * \mathbf{100}$$

Number of correct answers and total number of questions are calculated across all participants. We found that bar-size and bar-opacity are two charts that have highest percentage of correct answers. The comparison of percentage of correct answers across six charts is shown in Figure 5. We then carried out Wilcoxon Signed-Rank test between each pair of charts for correct answers. We found that BC is significantly different ($p<0.05$) from BOR, SO, SS and AC is significantly different ($p<0.05$) from SS. We also found that BO is significantly different ($p<0.05$) from SO and SS.

5.2 Total task duration

We measured the average time taken to complete the task for each chart. We observed that bar-size has the lowest average time and bar-orientation has the highest. As this parameter does not include user responses it would be inappropriate to evaluate

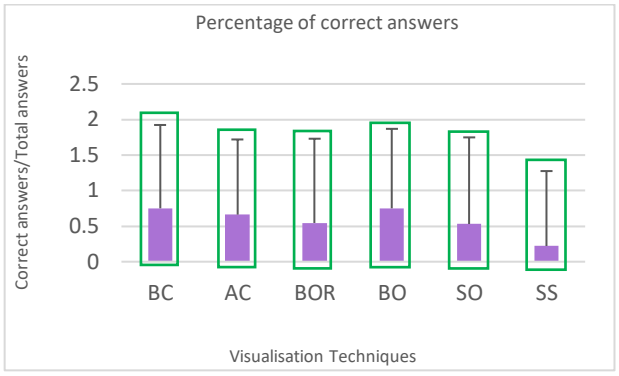


Figure 5: Percentage of correct answers

charts using only this parameter. For example, a chart with high task duration might perform better in user responses. The best-case scenario would be a high percentage of correct answers and low task duration. To mitigate this issue, we considered correct user responses along with total task duration. We refer this parameter as accuracy per unit time (APT) and is calculated as

$$APT = \frac{number\ of\ correct\ answers}{total\ task\ duration}$$

We found that APT of bar-size is the highest and shape-size is the lowest as depicted by Figure 6. We further undertook Friedman test for the average task duration of each participant. We found significant difference between means of charts (Chi square (5) = 15.354, $p<0.05$). We then carried out Wilcoxon Signed-Rank test between each pair of charts. We found BC and SS are significantly different from AC and BOR.

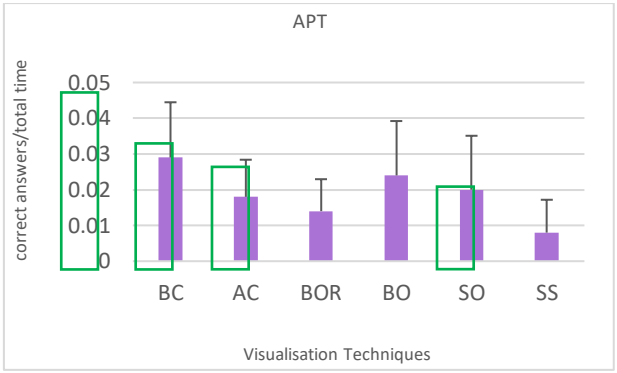


Figure 6: APT across all charts

5.3 Fixation and saccade rate

We calculated fixation and saccade rate for all participants across six charts. Bar-opacity has the lowest fixation rate but highest saccade rate. Bar-size has the highest fixation rate and area chart has the lowest saccade rate. Fixation and saccade rate would give information about total fixations during the task which includes movement of participant around the scene and answering questions. To investigate how long user focused only on visualisation chart we analysed fixation and saccade rate on chart. Fixation and saccade rate across all charts are shown in Figure 7 and 8, respectively. Bar-brightness chart has the highest fixation and saccade rate, while bar-orientation chart has the lowest. We then undertook Friedman test for the fixation and saccade rate of each participant during the entire task. We found significant difference between means of charts for fixation rate (Chi square (5) = 12.714, $p<0.05$) and saccade rate (Chi square (5) = 14.214, $p<0.05$). We then carried out Wilcoxon Signed-Rank test between each pair of charts for fixation and saccade rate. We found that BC and BO are significantly different ($p<0.05$) from AC and BOR for fixation rate. We further noticed that BO is significantly different ($p<0.05$) from BC, AC, BOR and SS.

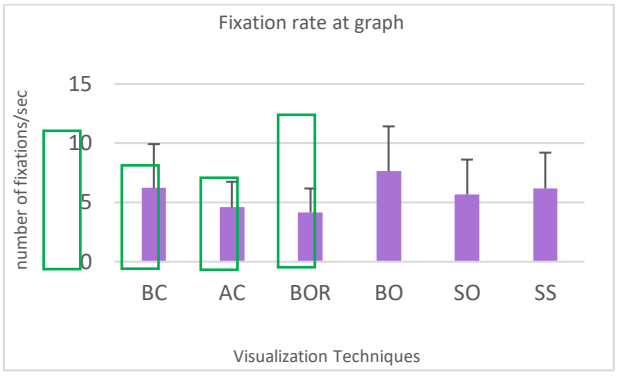


Figure 7: Fixation rate across all charts

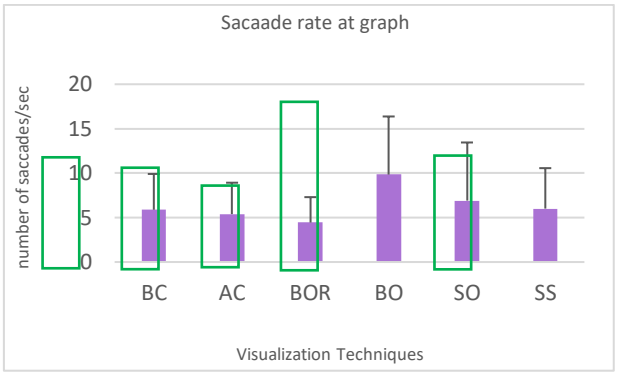


Figure 8: Saccade rate across all charts

5.4 Revisit sequences

Parameters of revisit sequences that we analysed are the number of unique sequences and the total number of revisits. A high number of unique sequences and total revisits would signify more combinations and repetitions. This would denote that participant was repeatedly scanning and focusing on the chart. We found that the bar-size chart has the lowest average unique sequences for sequences of length three and four while for sequences of length five shape-opacity chart has the lowest value. The bar-size chart also has the lowest total revisits for sequences of length three and four. Figures 9 and 10 show the number of unique sequences and total revisits for sequences of length three across all charts. We undertook Wilcoxon Signed-Rank test between each pair of charts for unique sequences and total revisits and did get significant difference ($p>0.05$).

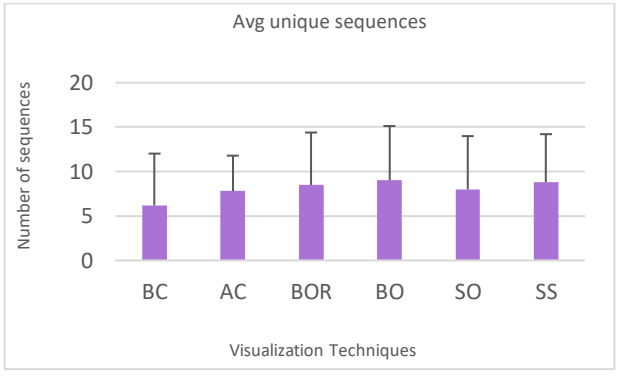


Figure 9: Unique sequences of length three sequences

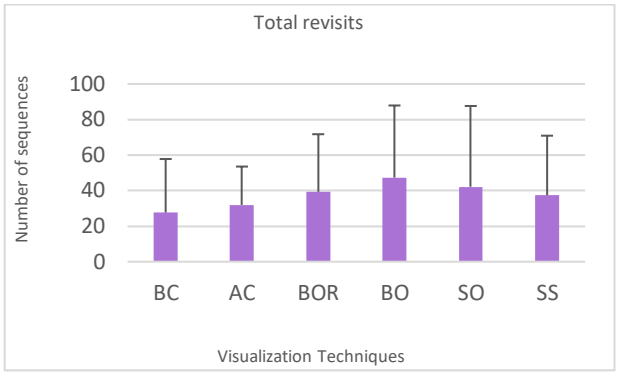


Figure 10: Total revisits of length three sequences

5.5 Analysis of pupil dilation

We undertook Friedman test on the output of LPF of the left and right pupil across all charts and found no significant difference ($p>0.05$) between the means of charts. Furthermore, we completed the Wilcoxon Signed-Rank test between each pair of charts and found no significant difference. We observed that pupil dilation ranges from 2.75 mm to 6.68 mm.

5.6 EEG data analysis

A Friedman test was undertaken on alpha, theta, low beta (Figure 11), and high beta bands of EEG. We did not get significant difference for any EEG band. We then undertook the Wilcoxon Signed-Rank test between each pair of charts for four EEG bands.

Alpha band: We got significant difference ($p<0.05$) between bar-size chart and area chart.

Theta band: We observed that bar-size chart is significantly different ($p<0.05$) from area chart and bar-opacity chart.

Low beta band: We observed that bar-size chart is significantly different ($p<0.05$) from area chart and bar-opacity chart. We got significant difference ($p<0.05$) between bar-orientation and bar-opacity charts.

We found that bar-size chart is significantly different ($p<0.05$) from the area chart in alpha, theta and low beta bands of EEG. We did not get any significant difference in the high beta band.

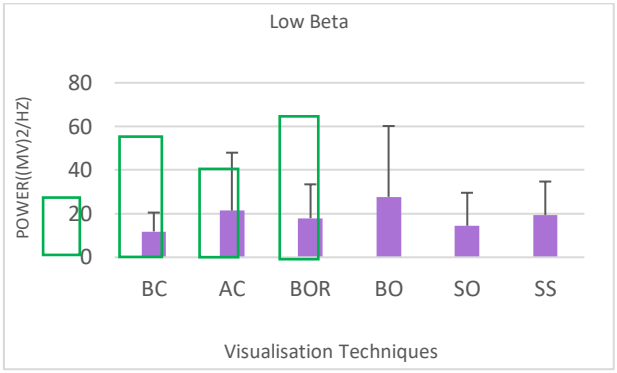


Figure 11: Average Low beta

5.7 Z-axis analysis

To analyse the effect of the 3rd dimension on participants, we separately investigated coordinates of gaze points. It would help us identify the impact of three axes on saccadic eye movements. We have considered all consecutive gaze points that form saccades. We calculated the absolute differences of coordinates between every two successive points. This calculation is based on L1 norm, which is the sum of the absolute differences of coordinates between two points. For example, if points P1: <x1, y1, z1> and P2: <x2, y2, z2> form a saccade, then the absolute differences of their coordinates are |x1-x2|, |y1-y2|, |z1-z2|. We then undertook the Friedman test on the computed absolute differences for every chart. We found that the absolute differences of coordinates are significantly different ($p<0.05$) for all charts (Table 3). Furthermore, we undertook the Wilcoxon Signed-Rank test between each pair of coordinates for six charts. We found that the x-axis and z-axis are significantly different

from the y-axis for all charts. We also analysed the movement of saccades along three axes. We calculated the average distance covered along the three axes during saccadic eye movement.

Table 3: Friedman test on differences of axes.

Bar size	*Chi square (2) = 25.765, p<0.05*
Area chart	*Chi square (2) = 22.706, p<0.05*
Bar orientation	*Chi square (2) = 25.529, p<0.05*
Bar opacity	*Chi square (2) = 20.235, p<0.05*
Shape opacity	*Chi square (2) = 20.588, p<0.05*
Shape size	*Chi square (2) = 20.588, p<0.05*

Figure 12 shows average distance of bar chart during saccade movement along all three axes. The distances are normalized from 0 to 1.

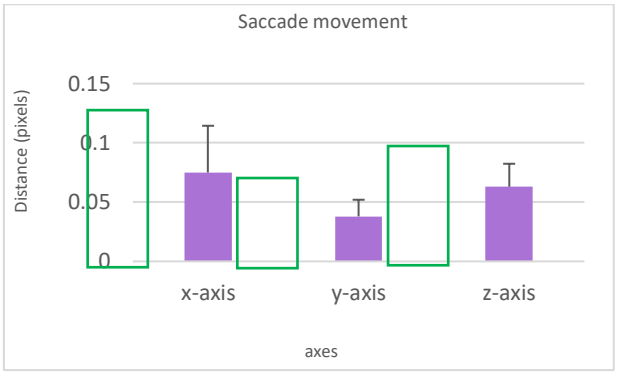


Figure 12: Average distance of bar chart

5.8 Comparisons of visual variables

As mentioned in Section 1, we have considered five visual variables in our study. Variable size, opacity, and orientation represent numerical data, while colour and shape depict nominal data. We investigated variables for each data type as discussed in the sub-section below. We compared three parameters (APT, fixation rate and saccade rate) of each variable. We also compared revisit sequences between variables.

5.8.1. Visual variables for nominal data

We divided variables representing nominal data into following two categories -

Category1: Nominal data is represented by colour.

Category2: Nominal data is represented by both colour and shape.

Bar-Size, Bar-Orientation and Bar-Opacity charts fall under category1, while Shape-Size and Shape-Opacity charts fall under category2. We calculated the average value of three parameters for three charts in category1 and two charts in category2. We observed a difference in user performance between category1 and category2. APT is higher for category1 than category2, while fixation and saccade rate is lower for category2. The number of revisits is lower for category1 in the sequence of length 3 and 5 but higher in length 4.

5.8.2. Visual variables for numerical data

We divided variables representing numerical data into following three categories –

Category1: Numerical data is represented by size.

Category2: Numerical data is represented by orientation.

Category3: Numerical data is represented by opacity.

Bar-Size and Shape-Size charts fall under category1, Bar-Opacity and Shape-Opacity charts fall under category3, while Bar-Orientation is a category2 chart. The average value of three parameters for charts in category1 and category2 are computed. We then compared these parameters among the charts of three categories and noticed a difference in user performance between all three groups. The fixation and saccade rate are lower for category2 than the other two categories. However, in terms of task duration, accuracy is higher for category1. The number of sequences is higher in category3 for sequences of length 3, 4 and 5.

6. Discussion

Our results showed that accuracy per unit time is higher for size and colour than other variables. We further observed that variable size and colour have a smaller number of fixations, saccades, and total revisits. Furthermore, our results showed a difference in cognitive load between size, opacity, and brightness. We can infer from these results that the cognitive load of participants is less when size is used to represent numerical data and colour is used to depict nominal data. In addition, from results of our analysis, we noticed that cognitive load while using a bar chart is less to an area chart. Finally, we looked back at four questions that we had raised in Section 1 –

Q1: Which 3D graph is best in terms of correct data interpretation?

We observed from Figure 5 that both bar-opacity and bar-size are similar in terms of correct data interpretation. We then noticed that bar-size's accuracy per unit time is higher than other charts and requires least number of revisits. We can infer from these results that bar-size chart is best in terms of correct data interpretation.

Q2: Which visual variables(s) is (are) easier to interpret than others?

We found that colour makes it easier to interpret nominal data as compared to shape. Performance of variable colour is higher in two gaze-based metrics and task duration as compared to shape in terms of accuracy. The size variable reduces the time required for processing numerical data as compared to the other two variables. However, size is similar to orientation for the other two ocular parameters (fixation and saccade rates). The size variable also performs favourably in terms of the count of revisit sequences. In addition, from our results we observed that opacity is worst in terms of correct data interpretation among three variables. We further observed that bar chart has lower cognitive load than area chart.

Q3: Does 3rd dimension add value to the visualisation?

We also observed that the addition of the 3rd dimension to the visualisation affects the performance of participants. We noticed that the movement of saccades along the z-axis is more than the movement along y-axis but less than the movement along x-axis, as shown in Figure 12. Moreover, the movement of saccades along axes were significantly different, as described in Table 6. This conveys that the movement along all axes are important and offers new information to participants.

Q4: Are there differences among graph types with respect to - ocular parameters, and cognitive load while interpreting graphs.

Notably, significant differences were observed among certain chart types with respect to ocular parameters. For example, bar-opacity and bar-orientation are different in terms of the fixation rate, as shown in Table 4. Similarly, significant differences were noticed among six pairs of charts concerning cognitive load. However, we found significant difference only between area chart and bar-size in all the three bands measured.

Beta band, especially in the sensory motor areas, are related to motor movements. A high value of power in the low beta band signifies low cognitive load [41]. We observed that bar-size has the highest value and bar-opacity has the lowest value in the low beta band. It indicates that bar-size chart is incurring less motor action and cognitive load.

Limitations and Future Work

This study evaluated six different graph types involving five different visual variables. The study design and analysis did not investigate interaction effects among chart types and visual variables. We were limited by time

and resource in terms of availability of participants and a repeated measure design with 2 types of graphs and five variables would increase duration of the experiment as well as required more participants than reported presently. A future work will limit the number of variables and analyse interaction effect.

Our sampling strategy did not measure participants' familiarity with different 2D graphs and the bar graph may found to be easier to interpret as participants were more familiar to it than area graph. However, it may be noted that our study involved three different types of bar graphs and the results related to visual variables are still useful for a single type of graph.

In the study design, we utilized all three axes to display data points and their values and users found to use both saccades and vergance eye gaze movements to browse through graphs. Future work will separately analyse saccades and vergance and report their proportions while interpreting 3D graphs.

For EEG analysis, we used a low-cost EEG headset and so did not analyse high frequency signals like Gamma band, future work will investigate ergonomic issues involving donning both a VR Headset and EEG cap and try to use an EEG device with more electrodes than the Emotiv Insight model.

Application

We have developed a VR model of a smart factory and set up visualization graphics at the locations of IoT nodes to embed real-time sensor readings on the virtual layout (Figure 13). We used the Unity 3D game engine and its modelling tool, Probuilder. The twin served as a three-dimensional illustration of the physical space whose dimensions were accurately mapped to the twin. Furthermore, the furniture and other objects in the physical space were also replicated in the virtual world. To improve the virtual environment's photorealism, baked global illumination was used, which entails computing the lighting behaviour and characteristics beforehand and storing them as texture files; this technique also reduces the computational load present in real-time global illumination. Additionally, Physically Based Materials or PBR were used as they physically simulate real-life materials' properties such that they accurately reflect the flow of light and thereby achieve photorealism. We deployed the twin on a Virtual Reality (VR) setup, specifically, the HTC Vive Pro Eye, since VR allows for immersive and interactive virtual walkthroughs. Users can browse through the virtual set up using 3D glass and as they touch any of the visualization, it provides both visual and haptic feedback based on sensor readings. We integrated ambient light sensor (BH1750) and, temperature and humidity sensor (DHT22) to show real-time visualization of data stream(s) in VR setup. Both sensors provide digital output. The BH1750 Sensor has a built-in 16-bit A2D converter and output unit is lux. The DHT22 sensor provides temperature in celcius and humidity as relative percentage. Sensors are interfaced to the VR machine through their respective wireless module(s). After establishing a peer-to-peer connection, individual wireless module communicates with VR machine using UDP protocol at a frequency of 1 Hz. A video demonstration of the system can be found at https://youtu.be/FX8zfQE5GF8

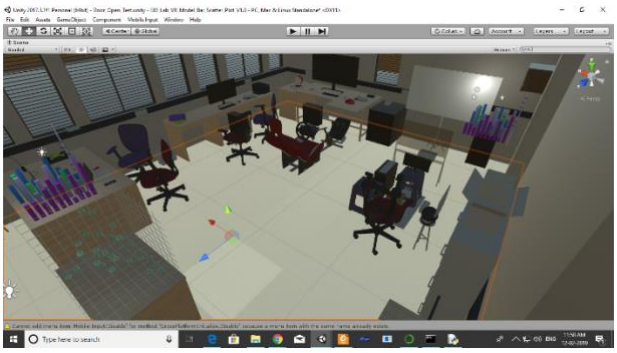


Figure 13. 3D Sensor Dashboard in a Digital Twin

7. Conclusion

This paper compared six different types of 3D graphs with respect to users' subjective and objective feedback. We analysed speed-accuracy trade off in users' response with respect to representative graph interpretation tasks. We also recorded and analysed ocular parameters and EEG to investigate eye gaze movement patterns and cognitive load while interacting with 3D graphs. A bar chart with different size of columns for different values of data points found out to generate most accurate response and least cognitive load among users.

8. References

1. Isenberg, T., Isenberg, P., Chen, J., Sedlmair, M., & Möller, T. (2013). A systematic review on the practice of evaluating visualization. IEEE Transactions on Visualization and Computer Graphics, 19(12), 2818-2827.
2. Steichen, B., Wu, M. M., Toker, D., Conati, C., & Carenini, G. (2014, July). Te, Te, Hi, Hi: Eye gaze sequence analysis for informing user-adaptive information visualizations. In International Conference on User Modeling, Adaptation, and Personalization (pp. 183-194). Springer, Cham.
3. Arjun, S. (2018, July). Personalizing data visualization and interaction. In Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization (pp. 199-202).
4. Huang, Y. J., Fujiwara, T., Lin, Y. X., Lin, W. C., & Ma, K. L. (2017, April). A gesture system for graph visualization in virtual reality environments. In 2017 ieee pacific visualization symposium (pacificvis) (pp. 41-45). IEEE.
5. Capece, N., Erra, U., & Grippa, J. (2018, July). Graphvr: A virtual reality tool for the exploration of graphs with htc vive system. In 2018 22nd international conference information visualisation (iv) (pp. 448-453). IEEE.
6. Erra, U., Malandrino, D., & Pepe, L. (2019). Virtual reality interfaces for interacting with three-dimensional graphs. International Journal of Human–Computer Interaction, 35(1), 75-88.
7. Sullivan, P. A. (2016). Graph-based data visualization in virtual reality: a comparison of user experiences.
8. Ware, C., & Franck, G. (1994, October). Viewing a graph in a virtual reality display is three times as good as a 2D diagram. In Proceedings of 1994 IEEE Symposium on Visual Languages (pp. 182-183). IEEE.
9. Ward, M. O., Grinstein, G., & Keim, D. (2010). Interactive data visualization: foundations, techniques, and applications. CRC Press.
10. Andrews, K. (2006, May). Evaluating information visualisations. In Proceedings of the 2006 AVI workshop on BEyond time and errors: novel evaluation methods for information visualization (pp. 1-5).
11. Carpendale, S. (2008). Evaluating information visualizations. In Information visualization (pp. 19-45). Springer, Berlin, Heidelberg.
12. Chen, C., & Czerwinski, M. P. (2000). Empirical evaluation of information visualizations: an introduction. International journal of human-computer studies, 53(5), 631-635.
13. Livingston, M. A., Decker, J. W., & Ai, Z. (2012). Evaluation of multivariate visualization on a multivariate task. IEEE transactions on visualization and computer graphics, 18(12), 2114-2121.
14. Amar, R., Eagan, J., Stasko, J.: Low-Level Components of Analytic Activity in Information Visualization. In: Proc. of 2005 Symp. on Information Visualization, pp. 15–21 (2005)
15. Conati, C., Maclaren, H.: Exploring the role of individual differences in information visualization. In: Proc. of the Working Conf. on Advanced Visual Interfaces, pp. 199–206 (2008)
16. Velez, M.C., Silver, D., Tremaine, M.: Understanding visualization through spatial ability differences. In: IEEE Visualization, VIS 2005, pp. 511–518 (2005)
17. Toker, D., Conati, C., Carenini, G., Haraty, M.: Towards adaptive information visualization: On the influence of user characteristics. In: Masthoff, J., Mobasher, B., Desmarais, M.C., Nkambou, R. (eds.) UMAP 2012. LNCS, vol. 7379, pp. 274–285. Springer, Heidelberg (2012)
18. Prabhakar, G., Mukhopadhyay, A., Murthy, L., Modiksha, M., Sachin, D., & Biswas, P. (2020). Cognitive load estimation using ocular parameters in automotive. Transportation Engineering, 2, 100008.
19. Palinko, O., Kun, A. L., Shyrokov, A., & Heeman, P. (2010, March). Estimating cognitive load using remote eye tracking in a driving simulator. In Proceedings of the 2010 symposium on eye-tracking research & applications (pp. 141-144).
20. Marshall, S. P. (2002, September). The index of cognitive activity: Measuring cognitive workload. In Proceedings of the IEEE 7th conference on Human Factors and Power Plants (pp. 7-7). IEEE.
21. Gavas, R., Chatterjee, D., & Sinha, A. (2017, October). Estimation of cognitive load based on the pupil size dilation. In 2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC) (pp. 1499-1504). IEEE.
22. Duchowski, A. T., Krejtz, K., Krejtz, I., Biele, C., Niedzielska, A., Kiefer, P., ... & Giannopoulos, I. (2018, April). The index of pupillary activity: Measuring cognitive load vis-à-vis task difficulty with pupil oscillation. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1-13).
23. Biswas, P., & Prabhakar, G. (2018). Detecting drivers' cognitive load from saccadic intrusion. Transportation research part F: traffic psychology and behaviour, 54, 63-78.
24. Goldberg, J., Helfman, J.: Eye tracking for visualization evaluation: reading values on linear versus radial graphs. Inf. Vis. 10, 182–195 (2011)
25. Iqbal, S.T., Bailey, B.P.: Using eye gaze patterns to identify user tasks. Presented at the The Grace Hopper Celebration of Women in Computing (2004)
26. Toker, D., Conati, C., Steichen, B., Carenini, G.: Individual user characteristics and information visualization: connecting the dots through eye tracking. In: Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, pp. 295–304 (2013)
27. Olsen, A., & Matos, R. (2012, March). Identifying parameter values for an I-VT fixation filter suitable for handling data sampled with various sampling frequencies. In proceedings of the symposium on Eye tracking research and applications (pp. 317-320)
28. Farnsworth, B. (2021). 10 Most Used Eye Tracking Metrics and Terms - iMotions. Retrieved 24 April 2021, from https://imotions.com/blog/10-terms-metrics-eye-tracking/#revisits
29. Salvucci, D. D., & Goldberg, J. H. (2000, November). Identifying fixations and saccades in eye-tracking protocols. In Proceedings of the 2000 symposium on Eye tracking research & applications (pp. 71-78).
30. Ziemkiewicz, C., Crouser, R. J., Yauilla, A. R., Su, S. L., Ribarsky, W., & Chang, R. (2011, October). How locus of control influences compatibility with visualization style. In 2011 IEEE Conference on Visual Analytics Science and Technology (VAST) (pp. 81-90). IEEE.

31. Green, T. M., & Fisher, B. (2010, October). Towards the personal equation of interaction: The impact of personality factors on visual analytics interface interaction. In 2010 IEEE Symposium on Visual Analytics Science and Technology (pp. 203-210). IEEE.
32. Peck, E. M. M., Yuksel, B. F., Ottley, A., Jacob, R. J., & Chang, R. (2013, April). Using fNIRS brain sensing to evaluate information visualization interfaces. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 473-482).
33. Ware, C., & Mitchell, P. (2008). Visualizing graphs in three dimensions. ACM Transactions on Applied Perception (TAP), 5(1), 1-15.
34. Fisher III, S. H., Dempsey, J. V., & Marousky, R. T. (1997). Data visualization: Preference and use of two-dimensional and three-dimensional graphs. Social Science Computer Review, 15(3), 256-263.
35. Garlandini, S., & Fabrikant, S. I. (2009, September). Evaluating the effectiveness and efficiency of visual variables for geographic information visualization. In International Conference on Spatial Information Theory (pp. 195-211). Springer, Berlin, Heidelberg.
36. Roth, R. E. (2017, March 6). Visual Variables.https://onlinelibrary.wiley.com/doi/abs/10.1002/9781118786352.wbieg0761
37. The professional-grade VR headset | VIVE Pro United States. Vive.com. (2021). Retrieved 7 May 2021, from https://www.vive.com /us/product/vive-pro/.
38. Insight Brainwear® 5 Channel Wireless EEG Headset | EMOTIV. EMOTIV. (2021). Retrieved 7 May 2021, from https://www.emotiv. com/insight/.
39. Biswas P and Robinson P, Evaluating the design of inclusive interfaces by simulation, Proceedings of the ACM International Conference on Intelligent User Interfaces (IUI) 2010
40. Onorati, F., Barbieri, R., Mauri, M., Russo, V., & Mainardi, L. (2013, July). Reconstruction and analysis of the pupil dilation signal: Application to a psychophysiological affective protocol. In 2013 35th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC) (pp. 5-8). IEEE.
41. 5 Types of Brain Waves Frequencies: Gamma, Beta, Alpha, Theta, Delta - Mental Health Daily. Mental Health Daily. (2021). Retrieved 10 May 2021, from https://mentalhealthdaily.com /2014/04/15/5-types-of-brain-waves-frequencies-gamma-beta-alpha-theta-delta/.
42. Drogemuller, A., Cunningham, A., Walsh, J., Cordeil, M., Ross, W., & Thomas, B. (2018, October). Evaluating navigation techniques for 3d graph visualizations in virtual reality. In 2018 International Symposium on Big Data Visual and Immersive Analytics (BDVA) (pp. 1-10). IEEE.

What do mobile AR game players complain about?: A qualitative analysis of mobile AR game reviews

Misbahu S. Zubair

School of Psychology and Computer Science, University of Central Lancashire, Preston, United Kingdom

mszubair@uclan.ac.uk

Augmented Reality (AR) combines real and virtual objects, provides opportunities for real-time interaction and provides accurate registration of 3D virtual and real objects. Mobile AR creates unique opportunities for gameplay unrestricted by the screen size of mobile devices with interactions possible between players, game objects and the real world. Since the launch of Pokémon Go in 2016, AR gaming has gone mainstream, several commercial mobile AR games have been launched, and researchers have conducted several studies on the motivations, intentions, and experiences associated with playing AR games. However, most studies in this area have focused on Pokémon Go, and have not included issues reported by players. This paper presents a study conducted with the aim of understanding issues, specifically those associated with the use of AR in games, that mobile AR game players face. User reviews for 10 popular commercial mobile AR games that utilise AR in all significant gameplay activities were gathered and analysed using Thematic Analysis to identify 11 themes of issues. This study adds to the body of knowledge and understanding of issues facing mobile AR game players and includes commercial mobile AR games other than Pokémon Go in the research on AR games.

Mobile Games. Augmented Reality. User Reviews. Qualitative Analysis. Thematic Analysis.

1. Introduction

Augmented Reality (AR) combines real and virtual objects, provides opportunities for real-time interaction and provides accurate registration of 3D virtual and real objects (Azuma, 1997). It can be utilised to enhance the user's perception of the real world and help in performing real and serious tasks, as well as to create unique, interactive and immersive experiences such as games. AR depends on capable devices with required displays, processors, input and tracking devices (Kesim & Ozarslan, 2012). Headsets/Head-Mounted Displays (HMD) (e.g. Microsoft's HoloLens 2) that independently meet these requirements currently exist; however, there are some HMDs that require the processing abilities of mobile phones and external input devices, these are considered cheaper alternatives to stand alone AR HMDs. Most modern mobile devices themselves are AR capable devices, and can be used without HMDs for mobile AR experiences.

Mobile AR frees games from the limits of the screen size of mobile devices and allows players to interact with the world (objects and locations) around them, thereby providing unique opportunities for "borderless" gaming anywhere anytime (Wetzel et al., 2011). Mobile AR games can be independent of the player's location, placing content locally to the player e.g. using a Marker to position and track the AR scene, although Markerless AR experiences are now very common (Oufqir et al., 2020). They can also be loosely coupled to certain locations i.e. played at different locations; or contextual i.e. strongly tied to the area they are played in (Wetzel et al., 2011).

Although mobile AR games existed before the launch of Pokémon Go, It took the release of the game in 2016 for most mainstream consumers to be introduced to the concept of AR and AR mobile games (Laine, 2018). Pokémon Go can be considered as a contextual, location-based, free-to-play mobile game created by Niantic Inc based on a Japanese media franchise. The game uses AR to allow players playing on their mobile devices to capture virtual pocket monsters that are augmented to their real environment. The nature of the environment determines the nature of the Pokémon found in the surroundings. Within two months of its release, it was downloaded more than 500 million times and won multiple game awards the same year (Hamari et al., 2019). An update by Niantic in December 2016 announced that players of the game have caught a combined total of more than 88 billion Pokémon, while collectively walking more than 8.7 billion kilometres (Niantic - The Pokémon GO Team, 2016).

What followed the Pokémon Go hype was an increase in the demand for AR experiences as AR mobile consumers began increasing. According to AR Insider, the number of active mobile AR users worldwide as of 2020 was 598 million, and this number is expected to grow to 1.73 billion by 2024 (AR Insider, 2021). The increased demand for AR experiences on mobile led to the launch of several other commercial mobile AR games, including those also based on existing franchises e.g. Harry Potter: Wizards Unite, Angry Birds AR: Isle of Pigs and Minecraft Earth.

Another consequence of the Pokémon Go hype is the effect it had on research on AR games and AR game players. Although research studies on AR mobile games were being conducted long before Pokémon Go, they were mostly focused on educational games (Furió et al., 2015; Zarzuela et al., 2013), and the potentials, opportunities and applications of AR for other serious (Angelopoulou et al., 2012; Botella et al., 2011). The launch and popularity of Pokémon Go inspired new research studies with aims including identifying the motivation for

playing AR games (Alha et al., 2019; Bueno et al., 2020; Zsila et al., 2018), the attitudes and intentions of mobile AR game players (Hsiao et al., 2019; Rauschnabel et al., 2017), and player experiences and engagement in AR games (Pyae et al., 2017; Pyae & Potter, 2016). However, little research on the issues faced and reported by players of mobile AR games has been conducted both before and after the launch of Pokémon Go. There is also a heavy focus on Pokémon Go when researching commercial mobile AR games. While this is not surprising due to its unparalleled popularity and impact, the one-sided research effort has led to the exclusion of other mobile AR games from this research.

Additionally, studies focused on Pokémon Go have limitations that could impact the generalizability of their findings. Respondents in these studies may have been affected by the Pokémon Go hype either negatively or positively; factors found to affect Pokémon Go players' behaviour, experience and opinions such as nostalgia, recreation and outdoor activities (Rauschnabel et al., 2017; Zsila et al., 2018) may not be present in other mobile AR games; finally, AR is only utilised in Pokémon Go for capturing pocket monsters and not in other significant gameplay activities, in fact, the game can be played without AR all together.

This study aims to understand issues, specifically those associated with the use of AR in games, that mobile AR game players face. This will be done by analysing the user reviews for popular commercial mobile AR games that utilise AR in all significant gameplay activities. In doing so, this study takes advantage of rich publicly available data provided by a large and diverse group of mobile AR game players that provides insights into their experiences.

The findings of this study add to the body of knowledge on the understanding of issues facing mobile AR game players and include commercial mobile AR games other than Pokémon Go and their players in this area of research.

2. Related Work

2.1 Mobile AR Games

Early research on mobile AR games mostly focused on the applications and benefits of AR in mobile games, especially for educational serious purposes. Some of the applications that have been studied include subject-specific learning (Furió et al., 2015; Zarzuela et al., 2013), improving social interaction and collaboration (Koceski & Koceska, 2011), recycling (Juan M et al., 2011), tourism (Angelopoulou et al., 2012; Etxeberria et al., 2012; Rodrigo et al., 2015), rehabilitation and therapy (Botella et al., 2011; Garcia & Navarro, 2014).

Several studies on mobile AR games' player experiences (Pyae et al., 2017; Pyae & Potter, 2016), including identifying the motivation for playing AR games (Alha et al., 2019; Bueno et al., 2020; Zsila et al., 2018), the attitudes and intentions of mobile AR game players (Hsiao et al., 2019; Rauschnabel et al., 2017), have been conducted since the launch and success of Pokemon Go brought mobile AR games out of research labs into mainstream usage. However, there is still little research involving experiences of players of other commercial mobile AR games, and on challenges and issues faced by players of mobile AR games.

2.2 Analysis of User Review

Application marketplaces such as Google's Play Store, Apple's App Store and Steam allow users to leave reviews and ratings for applications. Ratings allow users to assign a quantitative value based on their satisfaction with the app e.g. using a 5-star rating system. Reviews, on the other hand, are qualitative and serve several purposes such as giving feedback to developers, informing other users or potential users, reporting bugs, and even requesting new features (Di Sorbo et al., 2017; Maalej & Nabil, 2015). While a review is not required to rate an app, a rating is required to review an app in most application marketplaces (Mojica Ruiz et al., 2016).

Reviews are now considered rich sources of crowdsourced information; they have been collected and analysed in research studies to extract information that can aid in improving applications (Panichella et al., 2015), and in understanding user experiences and issues (Khalid et al., 2015).

The application(s) whose reviews are analysed depend on the aim of the study and can be selected based on popularity rankings (Khalid et al., 2015), searching the marketplace with relevant keywords based on the research study (Frie et al., 2017; Tan et al., 2020), utilising a precompiled list of relevant applications (Saoane Thach & Phuong Nam Phan, 2019; Thach, 2018), utilising category created by the market place with relevant applications (Fagernäs et al., 2021), randomly (Iacob & Harrison, 2013) etc. The reviews for the selected application(s) are then gathered usually using a scraping script. Existing studies have used review sample sizes ranging from hundreds (Faric et al., 2019) to millions (Hoon et al., 2012).

Statistical quantitative analysis and machine learning approaches are popular methods of analysing reviews due to the large number of reviews that are usually available for popular applications. These approaches have been successfully used to address several research objectives, for example, to identify: the relationship between

aspects within reviews and ratings (Guzman & Maalej, 2014; Huebner et al., 2018), reported bugs (Gao et al., 2018; Panichella et al., 2015), relationships between review length and rating (Vasa et al., 2012), review sentiments and the vocabulary used to express sentiment (Hoon et al., 2012), and retrieve feature requests (Iacob & Harrison, 2013).

However, not all research objectives can be addressed using quantitative methods. Studies aimed at understanding the context and not just identifying concepts such as experiences, opinions and issues and perceptions, take a qualitative approach to review analysis (Faric et al., 2019; Frie et al., 2017; Saoane Thach & Phuong Nam Phan, 2019; Thach, 2018). Due to the potentially large number of reviews available, a sampling approach is usually needed to decide a subset from the complete set of reviews for qualitative analysis. Examples of such sampling approaches found in the literature include selecting the top 10 most recent reviews for each app (Faric et al., 2019) or taking reviews within a certain time range (Saoane Thach & Phuong Nam Phan, 2019; Thach, 2018). This results in a more manageable sample size that can be manually analysed by researchers using, for example, thematic analysis (Faric et al., 2019; Frie et al., 2017; Tan et al., 2020; Thach, 2018).

3. METHODOLOGY

3.1 Data Collection

Ethical approval was sought and granted by the University of Central Lancashire's Ethics Review Panel before the start of data collection (SCIENCE 0117 CA). Data collection then began with identifying popular mobile AR games that utilise AR in all significant gameplay activities. This was done by searching Google Play Store apps using the search term "augmented reality game" searched in March 2021. The search was performed using a Google Chrome incognito browser window not connected to a Google account to avoid search results affected by an account's preferences. The search returned 250 results which were reviewed using a developed set of inclusion criteria to ensure AR games with a significant number of user downloads and reviews are selected for the study. The first 10 games that met these criteria were selected for the study. To meet the criteria, a game within the search result must:

have been downloaded at least 100,000 times,

have been reviewed at least 1000 times,

utilise AR(Azuma, 1997) in all its main gameplay activities.

Criteria 'i' and 'ii' were checked by reviewing a game's information on its Play Store page. Games that met both criteria were downloaded and played to test if they met criteria 'iii'. Review articles, review and gameplay walkthrough videos were used to check for criteria 'iii' in games that required equipment e.g. markers and HMDs to be used with a mobile device. By following these inclusion criteria, games that only use AR in a single mode or as an additional feature, only overlay images on camera view, have a low number of downloads or user reviews, were all excluded. The games selected and their descriptions, in brief, are presented in Table 1.

The complete set of reviews for the 10 selected games were downloaded using a python review scrapping script and saved as excel spreadsheets. To ensure the anonymity of reviewers, the scrapping script was customised to save only the review text, the star rating, and the thumbs-up count for each review and discard identifying information such as username and profile picture. In total, 36,231 reviews were saved.

Table 1: Selected games and their descriptions in brief.

Game	Title	Description
G1	Angry birds AR: isle of pigs	AR instalment in the angry birds franchise. Players destroy pigs and their structures using slingshots. They can walk around structures to find weak elements, identify different angles for the best accuracy. Markerless and requires no extra equipment.
G2	Five nights at freddy's AR: special delivery	AR instalment in the five nights at freddy's franchise. Players turn around in their real environment to find and confront malfunctioning animatronics to survive these horrors come to life.

		Markerless and requires no extra equipment.
G3	Ghosts 'n guns AR	The player shoots at ghosts that emerge from a portal placed in the player's environment. Markerless and requires no extra equipment.
G4	Hero vision iron man AR experience	The player plays as iron man and shoots at enemies. Marker-based; requires the purchase of a set that includes goggles to hold the player's device, an iron man mask to hold the goggles over the player's face, a set of markers, and scannable infinity stones.
G5	Kazooloo AR	The player fights enemies emerging from the kazoolo game board Marker-based; requires the purchase of a kazoolo game board.
G6	Knightfall AR	Strategy game where the player defends a castle against an invasion. Markerless and requires no extra equipment.
G7	Minecraft earth (early access)	Ar instalment in the minecraft franchise. The player explores, collect resources, builds and survives. Markerless and requires no extra equipment.
G8	Pulimurugan AR game	Based on the movie titled 'pulimurugan'. The player fights a tiger. Marker-based; requires a 10 rupee indian currency note, preferably ones released in years 2014, 2015, 2016
G9	Star wars™: jedi challenges	Ar instalment in the star wars game franchise. The player plays as a jedi and can take on several challenges including lightsaber battles. Marker-based; requires the star wars: jedi challenges gear (lenovo mirage ar headset, lightsaber controller, and tracking beacon)
G10	Tablezombies augmented reality	The player plays as a shooter on a rescue chopper with the objective of stopping zombies from reaching a survivor base. Marker-based; requires a marker that can be accessed online and printed.

3.2 Sampling Reviews

Since the aim of the study is to understand issues that reviewers complained about, a qualitative approach to data analysis is more appropriate. Therefore a sampling approach had to be developed, as analysing all 36,231 reviews qualitatively will be almost impossible. Similar studies have analysed only the most recent reviews (Faric et al., 2019), or reviews within a particular period (Saoane Thach & Phuong Nam Phan, 2019; Thach, 2018).

However, these approaches do not address the sampling problem for app store mining (Martin et al., 2015) as they may miss out on reviews with relevant information from excluded periods.

The approach taken in this study ensures the most relevant reviews are included in the data to be analysed by selecting reviews based on their 'helpfulness' rather than their creation date. Review 'helpfulness' is used to measure the "utility or diagnosticity" of reviews as voted by users (Karimi & Wang, 2017). Play Store records the helpfulness of reviews as a thumbs-up count i.e. positive difference between thumbs-up and thumbs-down received by a review. To ensure that helpful reviews are chosen across all possible ratings, the top 10 most helpful reviews for each rating were chosen for each selected game i.e. 10 most helpful reviews with 5 stars, 10 most helpful reviews with 4 stars, 10 most helpful reviews with 3 stars, 10 most helpful reviews with 2 stars and 10 most helpful reviews with 1 star. In cases where multiple reviews have the same helpfulness i.e. thumbs-up count, the most recent review is prioritised for selection. The final sample was made up of 500 reviews, made up of 50 reviews per game. This approach was taken to ensure the selection of a sample that includes the most relevant reviews across all rating groups across all selected games.

3.3 Data Analysis

Thematic analysis, "a method for identifying, analysing and reporting patterns (themes) within data" (Braun & Clarke, 2006) was used to identify and analyse patterns in user complaints within reviews as used in similar studies (Faric et al., 2019; Frie et al., 2017; Tan et al., 2020; Thach, 2018). Content Analysis, which is used to explore textual data to determine trends and patterns of words used, their frequency and relationships (Vaismoradi et al., 2013) was also considered. Thematic Analysis was chosen since the aim of this study is not to prioritise or count issues, but to identify them and understand the context in which they occur.

The phases provided by Braun & Clarke (2012) were followed in ensuring the flexibility of the method is not abused and a systemic analysis of data was conducted. Data analysis started with the researcher reading all sampled reviews to gain familiarity, the researcher took notes and made comments about reviews found to be interesting and their associated games. Then the researcher coded the reviews using an inductive approach i.e. with an open mind labelled interesting reviews or segments of reviews with labels describing their content (Braun & Clarke, 2012). On completing coding, codes were reviewed to identify overlaps and patterns to construct themes. To ensure that the themes constructed truly reflect the content of the complete data set, the complete set of all 36,231 reviews was searched using keywords and key phrases from coded reviews and themes to identify the existence of reviews that could validate and strengthen them. The keywords and search phrases used include verbatim words found to be common amongst coded data and words assigned by the researcher (e.g. synonyms of verbatim words) to improve the chances of finding relevant reviews. For example, the verbatim search keywords used for the Dizziness and Location themes include "dizzy" and "outside" respectively, while the researcher assigned keywords include "sick" and "outdoor" respectively. This process led to the validation of existing themes (e.g. Location, Dizziness), the extension of other themes (e.g. Extra Equipment), the construction of new themes from existing codes previously categorised as miscellaneous (e.g. Accessibility, Device Utilisation) and the construction of an entirely new theme (Gameplay). A complete list of all themes their description, in brief, is provided in Table 2. This validation process also served to reduce the impact of a possible sampling bias (Martin et al., 2015). Finally, the coded reviews or review segments for each theme were analysed to identify those that provide an accurate narrative of what is embedded in the complete data set, these were selected and are presented with each theme in the section that follows.

Table 2: Constructed themes and their descriptions in brief.

Theme	**Description**
Guidance	Lack of guidance on setting up and playing games.
Dizziness	Feeling dizzy as a result of moving around while playing.
Location	Having to play outside or in large spaces only.
Accessibility	Facing accessibility barriers to gameplay.
Gameplay	Poor utilisation of ar in improving gameplay.
Plane detection	Difficulty detecting planes to place the game scene.
Tracking	Difficulty tracking the scene or game objects.

Battery drain	Battery consumption becomes high when playing the game.
Overheating	The device becomes very hot when playing the game.
Device utilisation	Poor utilisation of the player's device and its capabilities.
Extra equipment	Issues associated with the need for such equipment, cost, availability, compatibility with devices and reusability

4. FINDINGS

The themes constructed through thematic analysis are provided in this section. Each theme is briefly discussed and examples of coded reviews or review segments are also provided. It should be noted that quoted reviews have been minimally amended to correct spelling and grammar errors, remove emojis, and to preserve the anonymity of reviewers by removing sensitive information without changing their intent (Nicholas et al., 2017).

4.1 Guidance

Findings highlighted reviews expressing frustration over the complexity of setting up and playing some mobile AR games with no guidance:

If only there were more instructions on what size of surface to use, what kind of lighting is needed, or if the playfield is scalable to what is available. (Game: G6)

"What is this game? There is absolutely no instruction or tutorial, even in the beginning. I have absolutely no idea what I am supposed to do. (Game: G7)

4.2 Dizziness

Some AR games demand a certain degree of physicality to play, this is usually in the form of utilising a players movement in their physical environment as a mechanic in the game. Some reviewers of these games reported a feeling of dizziness as a result of moving around while playing:

The only bad thing is that for players like me who get dizzy easily, if we play for more than 10 minutes we get really dizzy from looking and spinning around. Other than that, really awesome game. really recommend trying it out. (Game: G2)

Makes me dizzy but it's still fun as long as you ignore annoying stuff like in-app purchases. (Game: G7)

4.3 Location

Some of the games reviewed require players to be outdoors to complete part of the game loop. However, playing outdoors is not always ideal as shown by the reviews below:

So far it has been alright. However, it uses so much battery and data that it is impractical to play outside without using battery packs. And having a really large data package. We only played a bit outside and it jumped my phone data a gig. Very reluctant to try again but we will. (Game: G7)

Also, It's too cold outside to play, who's idea was it to launch at the start of winter? I'll try again in spring when I can play it (Game: G7)

Another issue found to be reported by reviewers associated with location is the size requirements that need to be met for some games to be playable:

A little confusing at first, but really fun. It's hard to play in small spaces, you have to be in an open area, standing. (Game: G2)

It is very fun when it works. The adventures are too big to play comfortably anywhere but an empty field and the motion tracking is terrible and the adventures end up sliding around constantly. (Game: G7)

4.4 Accessibility

AR games may utilise player movement as a mechanic, this could lead to accessibility issues for players with mobility issues as revealed by a review shown below:

When I got this there was no indication that you're expected to constantly move around the piece of cardboard you place on the floor. So, if you're in a wheelchair or have problems walking simply forget this thing. (Game: G5)

In addition to player movement mechanics, other features such as flashing lights and images, especially in games that require headgear, can also lead to accessibility issues to those that are sensitive:

If you are epileptic or sensitive to flashing lights this game is NOT for you. I am dizzy after less than a minute of trying to play this game. When you click to collect items the screen flashes a lot very bright and very quickly with animations. I am not able to play this game at all. Please be careful. (Game: G7)

Finally, some reviews complained about the visibility of objects in the game complaints caused by the scale of the scenes:

There are lots of kinks to work out such as the flat surface since it's hard for my phone to find. And when one is found, I sometimes find it hard to see on my phone when it's too far. (Game: G1)

It's hard to see the enemy, everything is so small. (Game: G6)

4.5 Gameplay

Some reviewers found some of the games analysed lacking in terms of gameplay, despite their use of AR. This can be seen in the following reviews:

Awkward and confusing. An AR haunted house idea is pretty cool. But this doesn't go beyond turning around. (Game: G2)

Nice innovation, but after prolonged use it gets boring. There's not much to do other than shoot. (Game: G3)

The graphics are nice but the gameplay is boooooring. (Game: G8)

4.6 Plane detection

Some reviewers complained about being unable to detect a surface to "place" their games, for example:

Can't even start a game because the camera cannot detect a flat surface. When it does find a surface, the stage jumps randomly off-screen, then crashes most of the time. (Game: G1)

Can't even get it to recognize any of the flat surfaces in my room. Floor, table, counter, bed, nothing. (Game: G6)

4.7 Tracking

There were also complaints about tracking the scene while playing the game. This was more common in games that required player movement. For example:

When I turn to shock the animatronic, It constantly stays to the left of my screen no matter where I'm facing, making it impossible to hit. I just can't play this. (Game: G2)

Really well made AR game for android, though it gets out of position when you move too fast. But the experience is really nice… (Game: G3)

It's a great idea. However, the game freezes to recalibrate when you move too close, too far, too fast, or away from the page a bit. (Game: G10)

4.8 Battery drain

Several reviewers reported having the batteries of their devices drained as a result of playing the reviewed AR games, for example:

I liked the game, it was fun, although the battery drain is high and AR would stop working every now and then so the ghosts would hit me and I couldn't aim at them. (Game: G3)

Cool game, but it drains a lot of battery. (Game: G10)

4.9 Overheating

There were also complaints about devices overheating during gameplay. Reviewers reported having to stop playing after a little while due to this issue and in some cases becoming concerned, for example:

Makes my device hot enough to slowly cook an egg. (Game: G6)

I wanted to give the game a try, but my device got that hot I felt it was going to explode or something. It looks like it might be a good game but maybe it shouldn't run on mobile. I don't want to risk losing my phone. (Game: G9)

4.10 Device utilisation

There were complaints from users with devices that they considered "low-end" on the performance of games on those classes of devices, for example:

Nearly impossible to play on lower-end devices, haywires are pretty much instant death, camera tracking doesn't work properly…; if you have a compatible device lower than [device], DON'T BOTHER! (Game: G2)

There were also complaints from reviewers who own “high end” about issues they believed should not be occurring on devices with specs as good as theirs, for example:

The gameplay is very nice but the problem is when playing the game my mobile heats up. I tried other devices and noticed the same thing, they get overheated. Even when handling bigger applications my mobile didn't heat up this much. So I hope you can resolve the problem in your next update (Game: G1)

The premise is really good and I've seen more than flattering gameplay but for some reason even though my phone is a [device] that should be more than enough for this game it crashes as soon as the presentation for Chica and Foxy in the opening and doesn't go any further. (Game: G2)

4.11 Extra Equipment

Games that require the use of extra equipment, including markers, received several complaints in their reviews associated with the following:

4.10.1. Use

Several reviewers complained about the need for equipment with attached costs in the games that required them, for example:

CAN YOU MAKE THE APP TO USE GOOGLE CARDBOARD? It would be better if you do so. (Game: G4)

Forces you to buy things and if you don’t have them you can’t play. Should be optional for you to have the toys. (Game: G5)

I rarely give 1 star but I’m disappointed that I need to buy expensive gear for this. It would be better if I could just Chromecast the game onto my tv and use my phone as a lightsabre. (Game: G9)

4.10.2. Compatibility

Complaints about compatibility issues were made by several reviewers who own extra equipment but were unable to use them with their devices, for example:

Won't connect with my phone, even though it's on the compatible list. I've uninstalled and reinstalled and it doesn't work. It's a waste of money. (Game: G9)

My phone isn’t connecting to my lightsabre because it "isn’t compatible" even though it fits in the headset and has Bluetooth capability. (Game: G9)

4.10.3. Cost

In addition to complaints on the requirement to use extra equipment, some reviewers complained about the associated cost implications, whilst still expressing their interest in the game. Examples are provided below:

This app is fantastic but the set is very expensive. (Game: G4)

The game is fun but $180 is a very steep price to pay to play it. (Game: G9)

4.10.4. Availability

In some cases, reviewers willing to purchase extra equipment were unable to do so due to lack of availability in certain regions and in online stores, for example:

I saw a review video on this and was impressed and enjoyed the demo version. Unfortunately, the boards were not readily available in Canada at that time. (Game: G5)

There were also complaints about the availability of markers that did not have to be purchased, for example:

Awesome game but make it compatible for new 10rs notes, it's hard to find 2016 edition 10rs notes. (Game: G8)

Link for the marker image does not work...it fails to download every time. (Game: G10)

4.10.5. Reusability

What else can my equipment be used for? This was a question that was found in several reviews associated with required extra equipment. Examples are shown in the reviews below:

It is an awesome game. First, I bought the board game and then downloaded the app. It is really awesome. But I want to ask one thing, do the board and the app become a waste if the game ends? (Game: G5)

This is an amazing game, I love it, my one problem is that I beat everything on it, and so now it just sits on the shelf collecting dust. (Game: G9)

5. Discussion

This study analysed the reviews of 10 popular mobile AR games to understand the issues reviewers complained about.

Although mobile AR games are becoming more and more popular, AR is still somewhat novel to a lot of users and its interaction methods and practices are still evolving (Ghazwani & Smith, 2020). This means that not all players will be able to intuitively set up and play mobile AR games, and several players may struggle to do so without clear and appropriate guidance and instructions. That is why guidelines for mobile AR games and applications have recommended providing help, documentation and training to uses (Tuli & Mantri, 2020).

Several research studies on the popularity and motivations associated with Pokémon Go have identified outdoor play and exercise as important factors that affect the opinions of players e.g. Rauschnabel et al. (2017) and Zsila et al. (2018). However, findings from this study have shown that these factors may not be favoured by all players. In fact, some users prefer to play indoors for several reasons including weather conditions, access to Wi-Fi, access to a charging port and personal preference; and therefore do not prefer contextual mobile AR games or games that require a large space to play that may be difficult to find indoors. Additionally, although some reviewers enjoyed the exercise provided by AR games that require player movement, others preferred to use minimal physical effort in playing as proposed by Ko et al.'s (2013) usability principles for mobile AR applications, others still reported a feeling of dizziness as a result of moving around while playing. Dizziness has been observed in participants of studies on the use of AR applications, and although it has been found to occur more frequently in participants HMDs it is also experienced by participants using mobile devices (Moro et al., 2021).

Accessibility was also found to be impacted by the need to move around while playing excludes players with mobility issues. Other accessibility issues found by this study include the use of flashing lights and visibility issues that sometimes result from the low scaling of game scenes.

Although AR has the potential to allow for the creation of games with gameplay, it is sometimes used in games based on the assumption that it will automatically improve it and not because it adds nothing to the gameplay (Wetzel et al., 2011). Some reviewers felt the same way about their reviewed games, especially after playing for a while of getting used to the AR novelty of viewing 3D objects in their real environment.

It is safe to say that the most common complaints encountered were those associated with plane detection issues and tracking issues, which are challenging issues associated with AR in general. Tracking issues, specifically, have been reported in other AR application domains by other researchers (Palmarini et al., 2018; Qian et al., 2019; Sanna & Manuri, 2016). Based on the findings of this study, these issues made games “unplayable” either by preventing players from detecting a surface to set up the game scene or having tracking issues that affect the scene, game objects and players positioning and scale. Similarly, Mulloni et al. (2012) found tracking issues caused users to stop using AR browsers due to frustration. Player behaviours that can cause these issues as found by Radu et al. (2017) include: moving the camera so it is not able to view the marker(s), covering the device camera, and aiming too close or away from the marker (for marker-based AR). Another cause of these issues could be the lighting condition of the players environment or the texture of the plane (in the case of plane detection).

Two other important technical issues were found in user reviews: overheating and battery drain. These have been previously identified as challenges facing the implementation of mobile AR (Chen et al., 2018). Some reviewers pointed the finger at poor optimisation of games, while others were not surprised by the occurrence of the issues given the nature of the processing required by AR games. Research has proved that AR games can cause overheating and battery drain in mobile devices dues to factors like camera usage (Kang et al., 2019) and the high processing demands (Qiao et al., 2019); this means that even well optimised mobile AR games could be facing these issues. This is unfortunately the state of the AR and mobile technology presently, and so high-end mobile devices are also likely to face such issues.

When it comes to mobile AR games that require the use of extra equipment including markers, this study found reviewers to complain about the need for the equipment. Availability, high cost, compatibility and the lack of reusability of the equipment. Several reviewers thought equipment that required purchase should be optional and that all mobile AR games should be playable with just a mobile device and nothing else required. While the cost of the extra equipment associated with the reviewed games is low compared to the cost of standalone AR devices such as Microsoft’s HoloLens, it should be noted that the population of mobile game players is mostly made up of individuals that do not spend on games is significantly larger than that of those that do. A study by AppsFlyer (2016) found only 3.5% of gamers spend money in games and paid for games make up less than 38% of mobile revenue (Civelek et al., 2018). Therefore, it is not surprising that this group of users find the costs of the equipment

high, were frustrated when their devices could not use the equipment and disliked the fact that the equipment have limited use.

For markers that are freely available to access, there were complaints about their availability, in one game (G10) most complaints were about a broken link, while in another (G8) complaints were about access since the marker is a currency available only in a single country. There were also availability complaints about other equipment due to lack of stock in certain countries and regions, or lack of stock altogether. Ensuring the availability of web links, the use of universally available markers and making games only available to regions where equipment can be accessed could be used to resolve these issues.

5.1 Recommendations

While some of the challenges identified by this study can only be resolved by advances in technology (both hardware and software) e.g. tracking, plane detection and battery drain, some challenges can be avoided or mitigated when designing mobile AR games with the present technology. Therefore, design recommendations for avoiding or mitigating some of the issues identified by this study are provided below:

Provide clear guidance and instructions for setting up and playing the game. This should be presented in a way that is clear to all players, including those not familiar with AR games and AR technology in general.

Include warnings in games with flashing lights and fast-moving images; also Include warnings of dizziness in games that require quick and frequent movement.

Where possible, design breaks games that require movement, especially quick and frequent movement, to allow players to rest.

Where possible, consider the player location's weather conditions when providing game objectives/missions in location-based mobile AR games.

Where possible, design games to have, at least, levels or modes that are playable without the purchase and use of extra equipment.

6. Conclusion

This study collected and analysed reviews of popular mobile AR games on Google's Play Store with the aim of understanding issues associated with the use of AR in games that users complain about. Each game's most helpful reviews across all ratings were analysed using thematic analysis to find themes that makeup patterns in user complaints, then the complete set of reviews were searched using relevant keywords from coded reviews to validate the themes constructed. This resulted in the construction of 11 themes of user complaints namely. While most of the issues identified and discussed in this study have been reported in AR games and other AR application domains, issues such as accessibility, device utilisation and those related to extra equipment have not been reported widely by other studies on mobile AR games. Given the challenges that come with the use of AR in mobile games, this study recommends that designers and developers only utilise AR if it improves the gameplay of a game and not just because of the novelty effect it will have on players.

A limitation of this study is that the analysis of reviews was conducted by a single researcher. This raises the question of coding reliability and highlights the impossibility of conducting checks such as inter-coder reliability checks. However, thematic analysis as described by (Braun & Clarke, 2012) can be performed by a single researcher and favours inductive flexible theme development through immersive and repeated engagement with the data over the agreement on codes between multiple researchers (Terry et al., 2017). The use of both sampled reviews and the complete set of reviews allowed this study to construct strong themes through immersive and repeated engagement with the data that identified both issues and their contexts; this would not have been entirely possible if only the sampled data was used.

Based on the lessons learnt from this study, it is recommended that approaches that allow the use of sampled data to ease qualitative analysis and also utilise the complete data set to identify missing information should be utilised in qualitative research of user reviews.

As future work, reviews from a larger set of games from both Google's Play Store and Apple's App store will be analysed both qualitatively, to understand reported issues, and quantitatively, to include all reviews in the analysis thus identifying finer aspects of mobile AR games that reviewers complain about.

7. References

Alha, K., Koskinen, E., Paavilainen, J., & Hamari, J. (2019). Why do people play location-based augmented reality games: A study on Pokémon GO. *Computers in Human Behavior*, *93*, 114–122. https://doi.org/10.1016/j.chb.2018.12.008

Angelopoulou, A., Economou, D., Bouki, V., Psarrou, A., Jin, L., Pritchard, C., & Kolyda, F. (2012). *Mobile Augmented Reality for Cultural Heritage* (pp. 15–22). https://doi.org/10.1007/978-3-642-30607-5_2

AppsFlyer. (2016). *The State of In-App Spending: Global & Regional Benchmarks, 2016*.

AR Insider. (2021). *How Big is the Mobile AR Market?* https://arinsider.co/2021/02/23/how-big-is-the-mobile-ar-market-2/

Azuma, R. T. (1997). A Survey of Augmented Reality. *Presence: Teleoperators and Virtual Environments*, *6*(4), 355–385. https://doi.org/10.1162/pres.1997.6.4.355

Botella, C., Breton-López, J., Quero, S., Baños, R. M., García-Palacios, A., Zaragoza, I., & Alcaniz, M. (2011). Treating cockroach phobia using a serious game on a mobile phone and augmented reality exposure: A single case study. *Computers in Human Behavior*, *27*(1), 217–227. https://doi.org/10.1016/j.chb.2010.07.043

Braun, V., & Clarke, V. (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa

Braun, V., & Clarke, V. (2012). Thematic analysis. In *APA handbook of research methods in psychology, Vol 2: Research designs: Quantitative, qualitative, neuropsychological, and biological.* (pp. 57–71). American Psychological Association. https://doi.org/10.1037/13620-004

Bueno, S., Gallego, M. D., & Noyes, J. (2020). Uses and Gratifications on Augmented Reality Games: An Examination of Pokémon Go. *Applied Sciences*, *10*(5), 1644. https://doi.org/10.3390/app10051644

Chen, H., Dai, Y., Meng, H., Chen, Y., & Li, T. (2018). Understanding the Characteristics of Mobile Augmented Reality Applications. *2018 IEEE International Symposium on Performance Analysis of Systems and Software (ISPASS)*, 128–138. https://doi.org/10.1109/ISPASS.2018.00026

Civelek, I., Liu, Y., & Marston, S. R. (2018). Design of Free-to-Play Mobile Games for the Competitive Marketplace. *International Journal of Electronic Commerce*, *22*(2), 258–288. https://doi.org/10.1080/10864415.2018.1441755

Di Sorbo, A., Panichella, S., Alexandru, C. V., Visaggio, C. A., & Canfora, G. (2017). SURF: Summarizer of User Reviews Feedback. *2017 IEEE/ACM 39th International Conference on Software Engineering Companion (ICSE-C)*, 55–58. https://doi.org/10.1109/ICSE-C.2017.5

Etxeberria, A. I., Asensio, M., Vicent, N., & Cuenca, J. M. (2012). Mobile devices: a tool for tourism and learning at archaeological sites. *International Journal of Web Based Communities*, *8*(1), 57. https://doi.org/10.1504/IJWBC.2012.044682

Fagernäs, S., Hamilton, W., Espinoza, N., Miloff, A., Carlbring, P., & Lindner, P. (2021). What do users think about Virtual Reality relaxation applications? A mixed methods study of online user reviews using natural language processing. *Internet Interventions*, *24*, 100370. https://doi.org/10.1016/j.invent.2021.100370

Faric, N., Potts, H. W. W., Hon, A., Smith, L., Newby, K., Steptoe, A., & Fisher, A. (2019). What Players of Virtual Reality Exercise Games Want: Thematic Analysis of Web-Based Reviews. *Journal of Medical Internet Research*, *21*(9), e13833. https://doi.org/10.2196/13833

Frie, K., Hartmann-Boyce, J., Jebb, S., Albury, C., Nourse, R., & Aveyard, P. (2017). Insights From Google Play Store User Reviews for the Development of Weight Loss Apps: Mixed-Method Analysis. *JMIR MHealth and UHealth*, *5*(12), e203. https://doi.org/10.2196/mhealth.8791

Furió, D., Juan, M.-C., Seguí, I., & Vivó, R. (2015). Mobile learning vs. traditional classroom lessons: a comparative study. *Journal of Computer Assisted Learning*, *31*(3), 189–201. https://doi.org/10.1111/jcal.12071

Gao, C., Zeng, J., Lyu, M. R., & King, I. (2018). Online app review analysis for identifying emerging issues. *Proceedings of the 40th International Conference on Software Engineering*, 48–58. https://doi.org/10.1145/3180155.3180218

Garcia, J. A., & Navarro, K. F. (2014). The Mobile RehApp™: an AR-based mobile game for ankle sprain rehabilitation. *2014 IEEE 3nd International Conference on Serious Games and Applications for Health (SeGAH)*, 1–6. https://doi.org/10.1109/SeGAH.2014.7067087

Ghazwani, Y., & Smith, S. (2020). Interaction in Augmented Reality. *Proceedings of the 2020 4th International Conference on Virtual and Augmented Reality Simulations*, 39–44. https://doi.org/10.1145/3385378.3385384

Guzman, E., & Maalej, W. (2014). How Do Users Like This Feature? A Fine Grained Sentiment Analysis of App Reviews. *2014 IEEE 22nd International Requirements Engineering Conference (RE)*, 153–162. https://doi.org/10.1109/RE.2014.6912257

Hamari, J., Malik, A., Koski, J., & Johri, A. (2019). Uses and Gratifications of Pokémon Go: Why do People Play Mobile Location-Based Augmented Reality Games? *International Journal of Human-Computer Interaction*, *35*(9), 804–819. https://doi.org/10.1080/10447318.2018.1497115

Hoon, L., Vasa, R., Schneider, J.-G., & Mouzakis, K. (2012). A preliminary analysis of vocabulary in mobile app user reviews. *Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12*, 245–248. https://doi.org/10.1145/2414536.2414578

Hsiao, K.-L., Lytras, M. D., & Chen, C.-C. (2019). An in-app purchase framework for location-based AR games: the case of Pokémon Go. *Library Hi Tech*, *38*(3), 638–653. https://doi.org/10.1108/LHT-09-2018-0123

Huebner, J., Frey, R. M., Ammendola, C., Fleisch, E., & Ilic, A. (2018). What People Like in Mobile Finance Apps. *Proceedings*

of the 17th International Conference on Mobile and Ubiquitous Multimedia, 293–304. https://doi.org/10.1145/3282894.3282895

Iacob, C., & Harrison, R. (2013). Retrieving and analyzing mobile apps feature requests from online reviews. *2013 10th Working Conference on Mining Software Repositories (MSR)*, 41–44. https://doi.org/10.1109/MSR.2013.6624001

Juan M, C., Furió, D., Alem, L., Ashworth, P., & Cano, J. (2011). ARGreenet and BasicGreenet: Two mobile games for learning how to recycle. *19th International Conference in Central Europe on Computer Graphics, Visualization and Computer Vision, WSCG 2011 - In Co-Operation with EUROGRAPHICS, Full Papers Proceedings*, 25–32.

Kang, S., Choi, H., Park, S., Park, C., Lee, J., Lee, U., & Lee, S.-J. (2019). Fire in Your Hands: Understanding Thermal Behavior of Smartphones. *The 25th Annual International Conference on Mobile Computing and Networking*, 1–16. https://doi.org/10.1145/3300061.3300128

Karimi, S., & Wang, F. (2017). Online review helpfulness: Impact of reviewer profile image. *Decision Support Systems*, *96*, 39–48. https://doi.org/10.1016/j.dss.2017.02.001

Kesim, M., & Ozarslan, Y. (2012). Augmented Reality in Education: Current Technologies and the Potential for Education. *Procedia - Social and Behavioral Sciences*, *47*, 297–302. https://doi.org/10.1016/j.sbspro.2012.06.654

Khalid, H., Shihab, E., Nagappan, M., & Hassan, A. E. (2015). What Do Mobile App Users Complain About? *IEEE Software*, *32*(3), 70–77. https://doi.org/10.1109/MS.2014.50

Ko, S. M., Chang, W. S., & Ji, Y. G. (2013). Usability Principles for Augmented Reality Applications in a Smartphone Environment. *International Journal of Human-Computer Interaction*, *29*(8), 501–515. https://doi.org/10.1080/10447318.2012.722466

Koceski, S., & Koceska, N. (2011). Interaction between players of mobile phone game with augmented reality (AR) interface. *2011 International Conference on User Science and Engineering (i-USEr)*, 245–250. https://doi.org/10.1109/iUSEr.2011.6150574

Laine, T. (2018). Mobile Educational Augmented Reality Games: A Systematic Literature Review and Two Case Studies. *Computers*, *7*(1), 19. https://doi.org/10.3390/computers7010019

Maalej, W., & Nabil, H. (2015). Bug report, feature request, or simply praise? On automatically classifying app reviews. *2015 IEEE 23rd International Requirements Engineering Conference (RE)*, 116–125. https://doi.org/10.1109/RE.2015.7320414

Martin, W., Harman, M., Jia, Y., Sarro, F., & Zhang, Y. (2015). The App Sampling Problem for App Store Mining. *2015 IEEE/ACM 12th Working Conference on Mining Software Repositories*, 123–133. https://doi.org/10.1109/MSR.2015.19

Mojica Ruiz, I. J., Nagappan, M., Adams, B., Berger, T., Dienst, S., & Hassan, A. E. (2016). Examining the Rating System Used in Mobile-App Stores. *IEEE Software*, *33*(6), 86–92. https://doi.org/10.1109/MS.2015.56

Moro, C., Phelps, C., Redmond, P., & Stromberga, Z. (2021). HoloLens and mobile augmented reality in medical and health science education: A randomised controlled trial. *British Journal of Educational Technology*, *52*(2), 680–694. https://doi.org/10.1111/bjet.13049

Mulloni, A., Grubert, J., Seichter, H., Langlotz, T., Grasset, R., Reitmayr, G., & Schmalstieg, D. (2012). Experiences with the Impact of Tracking Technology in Mobile Augmented Reality Evaluations. *Proc. MobileHCI Workshop on Mobile Vision and HCI (MobiVis) 2012*, *2*. http://data.icg.tugraz.at/~dieter/publications/Schmalstieg_237.pdf

Niantic - The Pokémon GO Team. (2016). *200,000 trips around the Earth!* https://pokemongo.nianticlabs.com/post/milestones/?hl=en

Nicholas, J., Fogarty, A. S., Boydell, K., & Christensen, H. (2017). The Reviews Are in: A Qualitative Content Analysis of Consumer Perspectives on Apps for Bipolar Disorder. *Journal of Medical Internet Research*, *19*(4), e105. https://doi.org/10.2196/jmir.7273

Oufqir, Z., El Abderrahmani, A., & Satori, K. (2020). *From Marker to Markerless in Augmented Reality* (pp. 599–612). https://doi.org/10.1007/978-981-15-0947-6_57

Palmarini, R., Erkoyuncu, J. A., Roy, R., & Torabmostaedi, H. (2018). A systematic review of augmented reality applications in maintenance. *Robotics and Computer-Integrated Manufacturing*, *49*, 215–228. https://doi.org/10.1016/j.rcim.2017.06.002

Panichella, S., Di Sorbo, A., Guzman, E., Visaggio, C. A., Canfora, G., & Gall, H. C. (2015). How can i improve my app? Classifying user reviews for software maintenance and evolution. *2015 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, 281–290. https://doi.org/10.1109/ICSM.2015.7332474

Pyae, A., Mika, L., & Smed, J. (2017). Understanding Players' Experiences in Location-based Augmented Reality Mobile Games. *Extended Abstracts Publication of the Annual Symposium on Computer-Human Interaction in Play*, 535–541. https://doi.org/10.1145/3130859.3131322

Pyae, A., & Potter, L. E. (2016). A player engagement model for an augmented reality game. *Proceedings of the 28th Australian Conference on Computer-Human Interaction - OzCHI '16*, 11–15. https://doi.org/10.1145/3010915.3010960

Qian, J., Ma, J., Li, X., Attal, B., Lai, H., Tompkin, J., Hughes, J. F., & Huang, J. (2019). Portal-ble: Intuitive Free-hand Manipulation in Unbounded Smartphone-based Augmented Reality. *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, 133–145. https://doi.org/10.1145/3332165.3347904

Qiao, X., Ren, P., Dustdar, S., Liu, L., Ma, H., & Chen, J. (2019). Web AR: A Promising Future for Mobile Augmented Reality—State of the Art, Challenges, and Insights. *Proceedings of the IEEE*, *107*(4), 651–666. https://doi.org/10.1109/JPROC.2019.2895105

Radu, I., Guzdial, K., & Avram, S. (2017). An Observational Coding Scheme for Detecting Children's Usability Problems in Augmented Reality. *Proceedings of the 2017 Conference on Interaction Design and Children*, 643–649. https://doi.org/10.1145/3078072.3084337

Rauschnabel, P. A., Rossmann, A., & tom Dieck, M. C. (2017). An adoption framework for mobile augmented reality games: The case of Pokémon Go. *Computers in Human Behavior*, *76*, 276–286. https://doi.org/10.1016/j.chb.2017.07.030

Rodrigo, M. M. T., Caluya, N. R., Diy, W. D. A., & Vidal, E. C. E. (2015). Igpaw: Intramuros - Design of an augmented reality game for philippine history. *Proceedings of the 23rd International Conference on Computers in Education, ICCE 2015*, 489–498.

Sanna, A., & Manuri, F. (2016). A Survey on Applications of Augmented Reality. *Advances in Computer Science : An International Journal*, *5*(1), 18–27. http://www.acsij.org/acsij/article/view/400

Saoane Thach, K., & Phuong Nam Phan, T. (2019). Persuasive Design Principles in Mental Health Apps: A Qualitative Analysis of User Reviews. *2019 IEEE-RIVF International Conference on Computing and Communication Technologies (RIVF)*, 1–6. https://doi.org/10.1109/RIVF.2019.8713753

Tan, M. L., Prasanna, R., Stock, K., Doyle, E. E. H., Leonard, G., & Johnston, D. (2020). Modified Usability Framework for Disaster Apps: A Qualitative Thematic Analysis of User Reviews. *International Journal of Disaster Risk Science*, *11*(5), 615–629. https://doi.org/10.1007/s13753-020-00282-x

Terry, G., Hayfield, N., Clarke, V., & Braun, V. (2017). Thematic Analysis. In *The SAGE Handbook of Qualitative Research in Psychology* (pp. 17–36). SAGE Publications Ltd. https://doi.org/10.4135/9781526405555.n2

Thach, K. S. (2018). User's perception on mental health applications: a qualitative analysis of user reviews. *2018 5th NAFOSTED Conference on Information and Computer Science (NICS)*, 47–52. https://doi.org/10.1109/NICS.2018.8606901

Tuli, N., & Mantri, A. (2020). Usability Principles for Augmented Reality based Kindergarten Applications. *Procedia Computer Science*, *172*, 679–687. https://doi.org/10.1016/j.procs.2020.05.089

Vaismoradi, M., Turunen, H., & Bondas, T. (2013). Content analysis and thematic analysis: Implications for conducting a qualitative descriptive study. *Nursing & Health Sciences*, *15*(3), 398–405. https://doi.org/10.1111/nhs.12048

Vasa, R., Hoon, L., Mouzakis, K., & Noguchi, A. (2012). A preliminary analysis of mobile app user reviews. *Proceedings of the 24th Australian Computer-Human Interaction Conference on - OzCHI '12*, 241–244. https://doi.org/10.1145/2414536.2414577

Wetzel, R., Blum, L., Broll, W., & Oppermann, L. (2011). Designing Mobile Augmented Reality Games. In B. Furht (Ed.), *Handbook of Augmented Reality* (pp. 513–539). Springer New York. https://doi.org/10.1007/978-1-4614-0064-6_25

Zarzuela, M. M., Pernas, F. J. D., Martínez, L. B., Ortega, D. G., & Rodríguez, M. A. (2013). Mobile Serious Game Using Augmented Reality for Supporting Children's Learning About Animals. *Procedia Computer Science*, *25*, 375–381. https://doi.org/10.1016/j.procs.2013.11.046

Zsila, Á., Orosz, G., Bőthe, B., Tóth-Király, I., Király, O., Griffiths, M., & Demetrovics, Z. (2018). An empirical study on the motivations underlying augmented reality games: The case of Pokémon Go during and after Pokémon fever. *Personality and Individual Differences*, *133*, 56–66. https://doi.org/10.1016/j.paid.2017.06.024

The Impact of Virtual Reality Nature Environments on Calmness, Arousal and Energy: a Multi-Method Study

Hildegardo Noronha and Pedro Campos

ITI/LARSyS - University of Madeira

Polo Científico e Tecnológico da Madeira

Caminho da Penteada, piso -2 9020-105 Funchal

Portugal

hildnoronha@gmail.com, pedro.campos.pt@gmail.com

Virtual Reality is the current media's epitome of Immersiveness, Presence and Suspension of Disbelief. Both research and gaming industry communities have been building on this in order to exhaustively research and explore feelings of high-adrenaline, scariness, panic and other visceral and instinctive feelings. We take the opposite approach and try to prove that Virtual Reality can also be used to induce feelings of relaxation and soothingness effectively and strongly. Therefore, it could be used to improve the mental health of people who cannot be exposed to situations that induce said feelings. In our experiments, we found that Virtual Reality can be used to induce a strong sense of Calmness and to reduce the sense of Arousal and Energy, with a high degree of significance, having an effect with short-duration exposures. We also found hints that Virtual Reality may have an effect in the circadian cycle's regulation by exposing the subjects to a virtual sunset.

Virtual Reality. Digital Wellbeing. Digital Nature.

1. Introduction

1.1. Summary

It is widely agreed that Virtual Reality (VR) is the current media's epitome of Immersiveness, Presence and Suspension of Disbelief. The technology has very appealing characteristics that takes the user experience to a whole new level. To please the vision, it offers a wide field-of-view, stereo-vision, and the ability to look and move your head anywhere inside a virtual environment (with more degrees-of-freedom than is usually possible in other media). It also offers a greater ability to manipulate virtual objects, in three-dimensional spaces, than most media, even though it is still, somewhat, underdeveloped and is, for most commercial solutions, unnatural. Finally, it enables the user to use natural locomotion to navigate the virtual environment which, together with all the previous characteristics, pulls the user into the virtual environment in a way that was never possible before on any other kind of media. All of this creates a sense of Immersiveness, Presence and Suspension of Disbelief so great in some VR experiences and games, making it so visceral, that some people completely forget that they are in a game, which triggers some extreme reactions, including an elevated sense of fear and even panic. These intense feelings have been widely explored by indie developers and is currently making its way into the AAA gaming industry, making it a contributing factor for the widespread adoption of VR. What this paper explores is the polar opposite of the spectrum. Because getting automatic fear and panic reactions out of people is actually something that can be easily achieved in other forms of media (by using, for instance, the typical "jump-scare" cliché), even though it is very intense in VR. For this reason, we explored sensations such as relaxation and that soothing and warm feeling one can get, just like a sunny day at the beach or sitting by a campfire. These feelings have been targeted before by other media, through screen savers in TV sets, computers and mobile phones, but we argue that those feelings achieve a whole new level, visceral-like, when using Virtual Reality - similar to what can be achieved for the panic and fear-like feelings.

By proving that Virtual Reality can effectively create those feelings in a heightened level, the technology can be used in many fields where the mental healthcare needs a boost, from stressful work-places to healthcare and elderly homes, it could even have beneficial effects in depressions, especially, when people are somewhat limited in the locations that can trigger those feelings (such as in big cities, prisons, medical institutions) or when they simply lack the willingness to leave their homes.

1.2. Research Questions

Our aim is to discover how effective Virtual Reality is in quickly creating soothing, relaxing, and warm feelings. For this reason, we try to answer the following research questions, taking into account a short time exposure of 1 minute:

Can Virtual Reality strongly relax people?

Can Virtual Reality strongly increase peoples' mood?

A third research question came up when the data started to be analysed and will be further explained later in the paper:

Does Virtual Reality have an influence in the circadian cycle?

2. Literature Review

Nature is the ultimate therapy. The human, as an integral part of nature, needs it, not only to survive, but also to keep its mental health (Bratman et al., 2019; Parr, 2007). With industrialization and the creation of ever growing cities, sometimes even referred to as concrete jungles, the humans are steadily losing the healing touch of nature (Bratman et al., 2019). This is why the nature tourism is on the rising with its therapeutical values (Buckley, 2020; Buckley and Westaway OAM, 2021), where people can enjoy safaris, hunting trips, cetaceans watching in order to feel closer to nature. Those experiences have, arguably, the same effect as Zoos without having animals removed from their natural habitats, even though they may disturb the animals and cause other issues (Mason, 2000). There are studies pointing that we can get some of the beneficial effects without being, directly in the nature, for instance, while by being indoors. As an example, between many, Philippot (1993), Mcsweeney et al. (2014) and several others (please check Mcsweeney's paper's references for more) explore Indoor Nature Exposure as an alternative to real nature exposure. They review several papers about different key aspects of nature that can be used to improve some aspects of quality of life and both psycho and physical health. From the studies list, we can find good effects coming from potted plants, direct sun-light, a window view. But the effects can arise from more artificial means, such as photographs and videos and even artificial imagery generated by computers.

Older studies indicate good success in inducing a range of feelings using their current media technologies. Philippot (1993) researched the capacity of film segments in, reliably and unequivocally, inducing naturally occurring emotional states on exposed subjects. They exposed the subjects to six short film segments and evaluated their responses by using three questionnaires. They found out that the films can be used to elicit emotions, in a predictable manner, in most subjects. They also found out that the Differential Emotions Scale is better at discriminating between emotional states than the Semantic Differential. Two years later, Gross and Levenson (1995) developed a set of films to elicit eight emotional states (amusement, anger, contentment, disgust, fear, neutral, sadness, and surprise). They selected clips from over 250 films and showed it to 494 English-speaking subjects and then, based on the subjects' responses, selected 2 films for each emotional state.

An evolution into Virtual Reality was just a natural step. In 2003, Plante et al. (2003) studied the possible beneficial psychological effects of doing aerobic exercise while using Virtual Reality. They concluded that Virtual Reality could enhance enjoyment, energy, while reducing tiredness, if used in such a setting. On the other hand, they discovered that Virtual Reality has the opposite effect, if used without the exercise component, by increasing tension and tiredness, and lowering the energy level. One can argue that since this is a 2003 study, the technology has evolved considerably since then and the benefits might have increased while the negative effects might have reduced or been removed altogether. Baños et al. (2005) studied how the immersion affects the sense of presence by comparing Virtual Reality to both a monitor and a projection. They, later, conducted another study (Baños et al., 2006) where they expanded Mood Induction Procedures into Virtual Reality (creating a VR-MIP) and induced different moods (sadness, happiness, anxiety and relaxation) into their experiment subjects by making changes in a Virtual Environment Park. They reported a successful induction in both sadness and happiness using the VR-MIP. Felnhofer et al. (2015) researched the emotional arousalness of Virtual Reality. They studied five emotions (joy, sadness, boredom, anger, and anxiety) by exposing their subjects to an emotionally charged Virtual Park. They found some indications that Presence does not influence emotions in Virtual Reality.

There are also links between Presence and Emotions, in Virtual Reality, as explored by Riva et al. (2007). They explore the ability to elicit emotions in Virtual Reality, like in other medias. They also try to find a relationship between Presence, a strong characteristic in Virtual Reality, and emotions. They confirm the effectiveness of the medium in triggering Anxiety and Relaxing feelings. They found a circular interaction between Presence and Emotions where one inflates the other.

The same research trend is being expanded into Augmented Reality, as demonstrated by Mehra et al. (2019) whom tried to prove the power of positive mood in the productivity of software developers through the use of Augmented Reality. They tried to improve their working environment by superimposing virtual pets and scenic features unto the real-world - their work environment.

There has been some expansion into the usage of more senses (besides vision and hearing), like demonstrated by Serrano et al. (2016) who ran some experiments using Virtual Reality coupled with touch and smell stimulation in order to induce relaxation. They tested the efficacy of mood-induction procedure in a Virtual Reality (VR-MIP).

A high sense of Presence was found and well as a statistical difference in relaxation. They also found no improvement while using smell, but the sense of touch does improve both Presence and relaxation.

Feelings in Virtual Reality have also been researched into a more therapeutically component, as demonstrated by Baus and Bouchard (2014) who reviewed an approach of running exposure therapy, specially phobias, from a Virtual Reality, which they consider to be much more expensive, to Augmented Reality, while still being effective. In Augmented Reality, Juan et al. (2006) developed and tested a prototype using Augmented Reality in order to explore the treatment of acrophobia while exploring the feeling of Presence in immersive photography. They ran parallel tests using a real-world staircase and in immersive photography. A System Usability Scale questionnaire was administered finding out that the sense of Presence was very high in their system but that there was a clear awareness of the Reality versus the Virtual Environment. This context was found to be useful in the treatment of acrophobia. Later, Botella et al. (2010) explored the utilization of Augmented Reality in the treatment of phobias, namely, cockroach phobia. They argue that in vivo exposure is the recommended treatment. They show that Virtual Reality and Augmented Reality is an effective method of treating some of those phobias and show the advantages of using Augmented Reality as a treatment. McLay et al. (2014) studied the effects of Virtual Reality PTSD treatment on Mood and Neurocognitive. They expand the results of PTSD treatment using Virtual Reality into depression and anxiety. They found significant reduction in PTSD and anxiety and significant improvement on emotional Stroop test. There was no improvement in depression nor an improvement in neuropsychologial functions. Herrero et al. (2014) induced Positive Emotions through the usage of Virtual Reality in order to treat Fibromyalgia. Their experiments found no statistical relevant improvement in pain and fatigue related values. But it did show that most of the subjects showed improvements, or no change, in their mood, with only 7.5 % showing some deterioration. They indicate that Virtual Reality is an effective method of treating acute but not chronic pain. In Mental Health, Bermúdez i Badia et al. (2018) proposed an architecture that can foster emotional regulation strategies. The system can generate procedural content based on affection and was rated pleasant by the subjects.

Emotional training can also be achieved by the usage of Virtual Reality, as demonstrated by Bosse et al. (2014) by exploring a system where the military, the law enforcement and other high stress workers can learn to regulate their own emotions. Ferrer-García and Gutiérrez-Maldonado (2011) reviewed the research into how Virtual Reality can be used to treat body image disturbances. They note the lack to published controlled studies in the subject but acknowledge the great potential of Virtual Reality as a substitute for in vivo exposure.

Nature has beneficial effects on humans. Gould van Praag et al. (2017) studied the relaxation and well-being effects of Naturalistic environments through autonomic arousal and activation. Their study reinforced the health benefits present in exposure to natural environments. But the effects are beyond health and well-being. Bratman et al. (2012) studied the effect of nature on cognitive function. They reviewed several works and proposed a system to categorize nature experiences. The beneficial effects of nature seem to cross-over to other medias. In this case, to Virtual Reality, as Browning et al. (2020) found out. They researched the effect of 6 minutes 360º videos of natural settings and found out increased levels of arousal and mood. Yu et al. (2018) expanded beyond nature and also explored urban environments. They found effects on both psychological and physiological values. They found out that the effects of environments in Virtual Reality map those of real life with fatigue levels and several bad feelings increasing in urban settings and decreasing in natural settings. The effects are strong enough to help people recover from stress and anxiety, as demonstrated by Yin et al. (2020) in a bigger 100 subjects study. Instead of pure nature, they went with a biophilic office, but the nature effect is still there.

3. Research Methods

3.1. Apparatus

3.1.1. Virtual Reality Simulation

The Virtual Reality Simulation, where the subjects are tested, was built using Unity3D 2019.4, with the High Definition Render Pipeline and the help of the built-in Terrain Tool. The SteamVR Plugin was used to handle the Virtual Reality. Assorted high quality assets from the built-in Asset Store were used to aid on the construction of the environments.

There are two environments, a beach, and a forest, described below. Each environment has two time-of-the-day, totalling four different combinations that the users can experience.

The experiments were built with as much visual quality as it was possible, given our time, human and financial constraints.

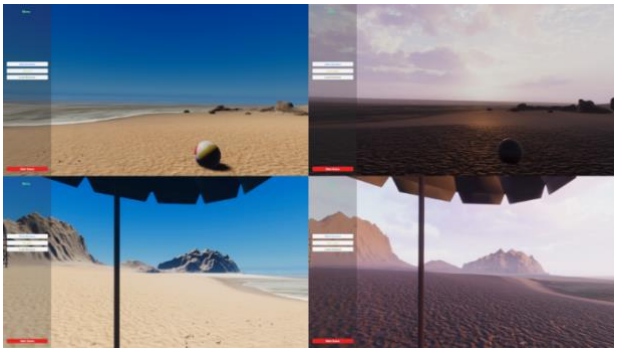


Figure 1: The Beach scenery. Midday on the left; sunset at the right

Beach - A golden-sand beach, near the ocean. Some palm trees paint the sand dunes behind the subjects. Assorted beach-related things are spread around the subject. The sound of waves can be heard coming from the sea. These sounds do not change between different times of the day. There are two time-of-the-day settings that the subjects can experience:

At midday - a very high strong sun.

At sunset - a romantic sunset.

Figure 2: The Forest scenery. Afternoon on the left; night at the right.

Forest - A calm, tree-filled, forest. The subjects are in a small sunken glade with are built nearby. You can hear the re going and a slight breeze on the trees. These sounds do not change between different times of the day. Just like in the beach, there are two time-of-the-day settings that the subjects can experience:

During the afternoon - the sun is comfortably midway in the afternoon creating some longish shadows while keeping good visibility.

During the night - the sun is long gone, giving space to a starry sky and the moon.

3.1.2. Virtual Reality Hardware

We used the HTC Vive Pro as the Virtual Reality headset. No controllers are used since there is no direct user interaction, other than looking around.

3.1.3. Sensors

We used a MUSE S to sense the level of relaxation of the user. It is a headband that uses Electroencephalography to monitor brain activity and outputs blackboxed data, such as level of relaxation. We also used a generic Heart-Rate sensor band to monitor the subjects' heart rate.

3.1.4 Computer

Due to the high requirements of the current generation Virtual Reality, we used a high-end gaming computer composed of an Intel i7-9700K CPU and a Nvidia GeForce RTX2080 Graphics Card. The rest of the computer was built with relevant matching high-end components.

3.2. Questionnaires

To complement the data gathered by the sensors, we exposed the subjects to the following standard questionnaires: 1) the AD ACL (Thayer, 1986) and 2) the SAM(Bradley and Lang, 1994). From all the questionnaires that we explored, we found these two to better evaluate the more positive and calm feelings that we are trying to study. The other related questionnaires are more focused on active and negative feelings.

We also exposed the subjects to another two non-standard, Likert-scale questionnaires - one at the beginning of the experiment and another at its conclusion.

3.2.1 Activation-Deactivation Adjective Check List (AD ACL)

The AD ACL that we used is based on Thayer's original questionnaire (Thayer, 1986). It has the original 20 adjectives randomly distributed in order not to influence the subjects' answers and to reduce coupling. The remaining of the questionnaire is standard with the standard 4 levels that the subject can feel.

3.2.2 Self-Assessment Manikin (SAM)

We used a standard SAM (Bradley and Lang, 1994) questionnaire with drawn manikins and the standard 9 levels scale but limited to Valence and Arousal.

3.2.3. Pre-Questionnaire

The pre-questionnaire tries to assess the prior level of experience of the subjects in Technology, Gaming and Virtual-Reality. It uses standard 7 points Likert-scale. All the 3 questions range from *1 - Low* to *7 - High* with a middle point at *4 - Medium*.

3.2.4. Post-Questionnaire

The post-questionnaire tries to assess how the subject feels about the subject in a subjective manner. It also uses standard 7 points Likert-scales. All the 4 questions range from *1 - Not at all* to *7 - A Lot*. The questionnaire asks the subject the following questions:

I feel that the Virtual Reality environments relaxed me.

I would enjoy spending more time relaxing in Virtual Reality environments.

I feel that, relaxing in a Virtual Reality environment could substitute a real-world alternative, in situations where the real-world alternatives are not accessible (e.g.: prisons, remote locations, big concrete cities).

I would pay to enjoy spending time relaxing in a Virtual Reality environment.

The questionnaire also asks them to sort the experiments by order of relaxation.

3.3. Experimental Protocol

3.3.1. The Experiment Script

Pre-Experiment:

The experiment starts with the subjects being asked to fill-in the pre-questionnaire followed by the evaluation of the subjects' current mood. The mood is evaluated using the AD ACL and the SAM questionnaires. This establishes a baseline, where the subject is not yet exposed to the Virtual Reality. The two sensors (Muse and Heart-rate sensor band) are then set up on the subject. The subject is, then, asked to sit quietly on a chair, facing the wall in a silent room, for 2 minutes (1 for the calibration of the MUSE S and another for the data gathering). We use this to set up another baseline, this time, for the sensors (mental relaxation and rest heart rate), again, before exposure.

Main Experiment:

The subject is exposed to each of the 4 Virtual Reality experiences random order. It starts with 1 minute of calibration of the MUSE S followed by 1 minute of exposure to the experience. After each experience, the subject is asked to, again, fill up the AD ACL and the SAM questionnaires, allowing us to assess their mood after each iteration. These 4 iterations are also experienced in the same chair, for the same 1 minute and without moving.

Post-Experiment:

In the end, the subject repeats the baseline measurements to allow us to compare the before and after of the experiment. The subject is finally asked to fill up the post-questionnaire before concluding the experiment.

Duration:

The experiment takes about 30 minutes per user. There are two short pre-questionnaires that are then followed by a baseline "empty" experience that takes a little more than 2 minutes (1 min calibration + 1 min experience). The main experiment takes (1 min calibration + 1 min experience) x 4 experiments = 8 minutes. All of this is followed by another baseline "empty" experience and the post-questionnaires. This all adds up to about 12 minutes of controlled time plus the questionnaires and overheads.

Clarifications:

Note that the subject is asked to remain still, when being evaluated, to avoid the subject's movements to affect the heart rate. This way, any change on the subject's heart rate can be attributed to a mental state, instead of a physical movement.

Also note that the order of the 4 iterations are picked from a previously built list. The list contains all the possible 24 permutations. The users are then, attributed, sequentially until all the permutations are tested. The list is, then, restarted.

Finally, the 1-minute calibration was chosen because the MUSE S was taking somewhat below 1 minute to calibrate. In order to keep the user's experiences as similar to each other as possible, we decided to calibrate for 1 minute, even if the MUSE S was done calibrating before that. As for the 1-minute Virtual Reality experiences, please check Section Limitations and Future Work.

3.4. Subjects

A total of 29 subjects participated in the experiments, 20 males and 9 females. Ages ranged from 21 to 39. 7 subjects had no to little experience with Virtual Reality, 18 had moderate experience with Virtual Reality and 4 had a lot of experience with Virtual Reality. The subjects were asked out from anyone in-campus (students, professors, staff, researchers, other), including the local university, research institutes and supporting buildings. A small number of out-of-campus subjects were also used.

4. Results

We conducted paired samples t-tests comparing the baseline, measured before the subjects are exposed to the experiment, to each of the four environment type (Beach at Noon, Beach at Sunset, Forest at Afternoon, Forest at Night). The order of the iterations is randomized to avoid any order effects. We also compared the baseline to each iteration (1st, 2nd, 3rd, and 4th), regardless of what environment it was.

We compared the baseline to the average of the 4 iterations, to the average of each environment (considering both times of the day) and, finally, against the after-exposure baseline (where relevant). We also calculated some simple averages, standard deviations, the effect power, and the effect size, where relevant. All this data is on several tables of the relevant subsections. Figure 3 shows a high-level view of the results, before and after the exposure. Figure 5 in Annex also shows a box-and-whiskers chart of the data to better clarify the data's distribution.

Figure 3: The feelings of the subjects, Before and After exposure

4.1 Activation-Deactivation Adjective Check List (AD ACL) Questionnaire

The AD ACL Questionnaire merges the adjectives into 4 feelings: *Energetic*, *Tired*, *Tension* and *Calmness*, that we use to further our research. Our Research Questions point to an increase in *Calmness*. We make no assumptions as to *Energy*, *Tired* and *Tension* other than there may be an effect. As such, we have 4 hypotheses: 1) The *Calmness* value of exposed subjects should increase; 2) The *Energy* value of exposed subjects should be different; 3) The *Tired* value of exposed subjects should be different; 4) The *Tension* value of exposed subjects should be different.

Hypothesis 1 (Calmness) is statistically significant for the Beach Midday and Sunset but not for the Forest. It has p-values of 0.009513 and 0.000024, respectively. The increase in *Calmness* is of 0.36 and 0.52, respectively, and in a scale of 1 to 4. It also an intermediate to large effect size with values of 0.529 to 0.929.

Note that the next 4 tables have the following headers: Average rating; Standard Deviation; Difference vs. Base before exposure; p-value of the t-test; Power; Effect Size.

Table 1: Calmness.

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**2.5**	**0.5**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**2.9**	**0.6**	**0.4**	**0.01**	**0.99049**	**0.53**
BEACH; SUNSET	**3.1**	**0.5**	**0.5**	**0.00002**	**0.99998**	**0.93**
FOREST; AFTERNOON	**2.6**	**0.6**	**0.1**	**0.24**	**-**	**-**
FOREST; NIGHT	**2.5**	**0.7**	**-0.0**	**0.40**	**-**	**-**

Hypothesis 2 (Energy) is statistically significant for all but the Forest at Night. It has p-values of 0.002103, 0.000043, 0.005301. The *Energetic* values always decreases ranging from 0.34 to 0.67, in a scale of 1 to 4. The effect sizes, however, are adverse for all.

Table 2: Energetic.

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**2.5**	**0.7**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**2.2**	**0.8**	**-0.3**	**0.00210**	**0.99789**	**-0.69**
BEACH; SUNSET	**1.7**	**0.6**	**-0.7**	**0.00004**	**0.99996**	**-0.88**
FOREST; AFTERNOON	**2.0**	**0.8**	**-0.5**	**0.00530**	**0.99470**	**-0.60**
FOREST; NIGHT	**2.3**	**0.8**	**-0.2**	**0.24187**	**-**	**-**

Hypothesis 3 (Tired) is statistically significant for the Beach at the Sunset. It has a p-value of 0.00078. The *Tired* value increases 0.5143, in a scale of 1 to 4. Its effect size is medium with a value of 0.7.

Table 3: Tired.

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**1.9**	**0.8**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**2.1**	**0.8**	**0.2**	**0.10784**	**-**	**-**
BEACH; SUNSET	**2.5**	**0.7**	**0.5**	**0.00078**	**0.99922**	**0.7**
FOREST; AFTERNOON	**2.2**	**0.8**	**0.3**	**0.10274**	**-**	**-**
FOREST; NIGHT	**2.0**	**0.8**	**0.0**	**0.93538**	**-**	**-**

Hypothesis 4 (Tension) is statistically significant for the Forest at Night. It has a p-value of 0.044 and an increase in *Tension* value of 0.31, in a scale of 1 to 4. It has a medium effect size of 0.747.

Table 4: Tension

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**1.3**	**0.3**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**1.3**	**0.3**	**-0.1**	**0.527**	**-**	**-**
BEACH; SUNSET	**1.2**	**0.3**	**-0.1**	**0.277**	**-**	**-0.88**
FOREST; AFTERNOON	**1.4**	**0.5**	**0.0**	**0.734**	**-**	**-**
FOREST; NIGHT	**1.6**	**0.8**	**0.3**	**0.044**	**0.956**	**0.75**

4.2. Self-Assessment Manikin (SAM) Questionnaire

The SAM questionnaire evaluates Valence and Arousal. We consider the scale to go from -4 to +4, with 0 being a neutral value. Our Research Questions point to a decrease in *Arousal*. We make no assumptions as to *Valence*. As such, we have 2 hypotheses: 1) The *Arousal* value of exposed subjects should be lower; 2) The *Valence* value of exposed subjects should be different.

Hypothesis 1 (Arousal) is statistically significant for the Beach (both times-of-day), the Forest at the afternoon and the Average of all experiments. It has p-values of 0.00073, 0.00010, 0.04547 and 0.00240. The decrease in *Arousal* values ranges from 0.97 to 1.46. All the effect sizes are adverse.

Table 5: Arousal

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**-0.3**	**2.6**	**2.6**	**-**	**-**	**-**
BEACH; MIDDAY	**-1.5**	**2.1**	**-1.2**	**0.00073**	**0.99927**	**-0.34**
BEACH; SUNSET	**-1.7**	**2.1**	**-1.5**	**0.00010**	**0.99990**	**-0.38**
FOREST; AFTERNOON	**-0.9**	**2.2**	**-0.6**	**0.04547**	**0.95453**	**-0.17**
FOREST; NIGHT	**-0.5**	**2.3**	**0.2**	**0.27068**	**-**	**-**

Hypothesis 2 (Valence) has no statistically significance in any of the experiences.

Table 6: Valence

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**2.0**	**1.5**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**2.1**	**1.6**	**0.1**	**0.69**	**-**	**-**
BEACH; SUNSET	**2.2**	**1.4**	**0.2**	**0.48**	**-**	**-**
FOREST; AFTERNOON	**1.9**	**1.5**	**-0.1**	**0.54**	**-**	**-**
FOREST; NIGHT	**1.6**	**1.5**	**-0.4**	**0.20**	**-**	**-**

4.3. Brain Relaxation

The hardware that we used - the MUSE S - has an app that outputs 3 values, measured along time: Active, Neutral, Relaxed. To merge the 3 values into 1, and since we are aiming at relaxation, we subtracted the *Active* time from the *Relaxed* time to find out the adjusted relaxed time (in 1 minute).

Our hypothesis is "The relaxed time should be higher after exposure". We achieved statistically significant results when comparing the baseline after exposure with the baseline before exposure as well as with the same comparison with the Beach. The p-values of the t-test are 0.0122 and 0.0131, respectively, with an increase of 8.79 and 9.32 seconds of relaxation in a minute of exposure. The effect sizes are small with values of 0.414 and 0.37, respectively.

Table 7: Brain Relaxation (in seconds)

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**21**	**18**	**-**	**-**	**-**	**-**
BEACH	**31**	**11**	**9.3**	**0.0131**	**0.9869**	**0.37**
FOREST	**27**	**14**	**5.9**	**0.1119**	**-**	**-**
BEACH + FOREST	**28**	**9**	**6.9**	**0.0437**	**0.9563**	**0.263**
BASE-AFTER	**30**	**15**	**8.8**	**0.0122**	**0.9878**	**0.414**

4.4. Heart rate

We take the median of the heart rate of each of the 6 measurements (before exposure, 4 virtual reality environments, after exposure). We use the median, instead the average, to filter out small spikes in the heart rate that occur due to subject movements. We then use these values to conduct the t-tests.

Our hypothesis is "The Heart-Rate of the subjects should be lower after the exposure". We were unable to prove or disprove this hypnotises with our data. The data fluctuates around no change with no statistically significant values.

Table 8: Heart Rate

Measure.	μ	σX	Δ	p	Power	E.S.
BASE-BEFORE	**73**	**10**	**-**	**-**	**-**	**-**
BEACH; MIDDAY	**73**	**9**	**0.4**	**0.35**	**-**	**-**
BEACH; SUNSET	**72**	**9**	**-0.5**	**0.29**	**-**	**-**
FOREST; AFTERNOON	**73**	**9**	**0.3**	**0.36**	**-**	**-**
FOREST; NIGHT	**74**	**9**	**0.6**	**0.25**	**-**	**-**
BASE-AFTER	**73**	**8**	**0.1**	**0.47**	**-**	**-**

4.5. Pre-Questionnaire

We did a simple average of all the subjects answers to the pre-questionnaire. The average value for Technology Experience if of 5.68, for Gaming Experience is of 4.36 and for Virtual Reality is of 3.46. The range of it goes from *1 - Low* to *7 - High* with a middle point at *4 - Medium*.

4.6. Post-Questionnaire

We did a simple average of all the subjects' answers to the post questionnaire. The subjects answered an average of 5.18 as to feeling that the Virtual Reality environments relaxed them, 5.07 as to whether they would enjoy spending more time relaxing in the Virtual Environments, 5.32 that Virtual Reality could be used as a substitute for the real-world in situations where the subject has no access to relaxing alternatives, and 3.32 as to they would pay to enjoy relaxing in a Virtual Environment. The range of it goes from *1 - Not at all* to *7 - A Lot*. The questions (Q1 through Q4) can be found in A.2.

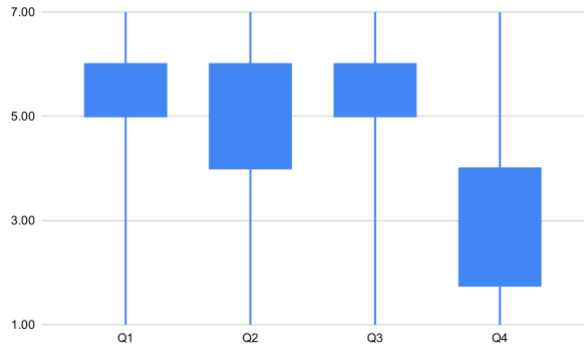


Figure 4: Answers to the Post-Questionnaire.

5. Discussion

Our tests confirmed, with a very high degree of statistically significance (best p = 0.000024; Beach at Sunset), that Nature Scenes in Virtual Reality can induce a sense of *Calmness* in the subjects who were exposed to it. The effect is considerably strong (with a power of 0.999976 and an effect size of 0.929): the subjects report an average increase in *Calmness* of 0.5214 points (Beach at Sunset) in a scale ranging from 1 to 4 (AD ACL). This effect is felt even in sessions of very short duration (1 minute). There is also a decrease in the sense of *Energy*, consistent with the current (Plante et al., 2003), with a high degree of statistically significance (p = 0.000043; Beach Sunset). This effect is even stronger than the *Calmness*, with a decrease of 0.67 points (Beach at sunset) but its effect size is, actually, adverse (negative). The *Tired* feeling is only statistically significant for the Beach at sunset (p = 0.00078) with an increase of 0.5143 points. It has an effect size of 0.7. The *Tension* feeling has no statistically significant except for the Forest at Night (p = 0.044). Its effect size is 0.747. The subjects felt uneasy

- scared even - and this can be seen in an increase of 0.3071 points in *Tension*. Some subjects even commented it directly that they were scared and/or were waiting for some kind of jump-scare. It was also rated, by most subjects, as the least relaxing of the 4 experiments. Note that the nature of the experiment was not explained to them, so they did not know that we were looking into relaxation instead of fear, and, as said before, some subjects were even expecting some kind of jump-scare.

The data supports that there is a statistically significant (best $p = 0.00010$, Beach at Sunset) decrease of 1.46 points in *Arousal*, in a scale ranging from -4 to +4, in the Beach experiment. The other scenarios also have a statistically significant reduction in *Arousal* except for the Forest at Night. All the effect sizes are adverse. The *Valence* shows no statistically significant data. Also, there seems to be little to no change in *Valence*, despite the lack of statistically significance. However, the Forest at night shows a decrease of 0.39, but not statistically significant ($p = 0.20$).

The MUSE S shows statistically significant results for both the Beach and the second baseline, after exposure, exhibiting the respective p-values of 0.013 and 0.012. It shows a considerable increase in relaxation time of 9.32 and 8.79 seconds (in 1 minute; from a baseline of 21.21 seconds before exposure). It has small effect sizes. This seems to indicate that the effect is maintained even after the users are no longer exposed. We did not, however, measured for how long this effect lasts.

The Heart-rate data did not produce any statistically significant data. There is a slight oscillation, below 1 BPM, and the p-value denies us any useful conclusions.

The subjective tests point to a higher than average feeling of *Relaxation* with 5.18 points in a 7-point Likert scale. Subjects would also like to relax more, in Virtual Reality (5.07 points). This does seem to point to the fact that the subjects feel a good level of *Relaxation* during the experiments - enough for them to feel the difference, in a conscious level. Some subjects also expressed this feeling with their words pointing to both a high level of relaxation and even sleepiness, despite being alone on a room with strangers performing an experiment. But we are aware that, this being a subjective questionnaire and a matter of opinion, our low number of answers might not be representative of the population. This leads us to the next question where they give their opinion on whether or not they feel like a system like this could substitute the reality in situations where it is not possible to experience similar environments (like prisons, remote locations, big concrete cities). This may be interpreted as a loaded question and, as such, appears after the other questions to avoid directing the subjects. The subjects did answer a little bit higher - 5.32 - but still similar to the other two questions. We built the three questions to evaluate almost the same thing with, somewhat, different phrasing. And the subjects did stay around the mid 5s.

The last question asks if the subjects were willing to pay to experience more relaxation in Virtual Reality environments. Now the value does go down, as expected, to 3.32, in the same scale. This may be interpreted as an opinion on how important they feel a system like this is, despite its capabilities to perform. We feel that 3.32 is still high enough, in light of the other answers and all the experiment results. Hence, their true feeling about relaxing in Virtual Reality, might be between the 3s and the 5s, pointing at a medium-strength subjective feeling.

6. Conclusion

Our experiments indicate a statistically significant and strong increase in *Calmness* and a decrease in *Energy*. There is also a statistically significant and strong decrease in *Arousal*. This is also confirmed by the subject data where the subjects feel above average relaxation (greater than 5 in a 7-point Likert scale with a central neutral point at 4) coupled with some verbal, free-form indications of such. All this data confirms our hypothesis that “Virtual Reality can strongly relax people”.

The same data does not provide any statistically significant information about *Valence* so we will have to withdraw any conclusions about the hypothesis that “Virtual Reality strongly increase peoples' mood”.

The hypothesis that “Virtual Reality help to regulate the circadian cycle” is, partially, confirmed. At sunset (a synchronization time for the circadian cycle), the data showed a statistically significant increase in *Tired* and a decrease in *Energy*. It also showed a statistically significant increase in *Calmness* and a statistically significant reduction in *Arousal*. These changes are also stronger in the sunset that they were for the remain of the experiments, and the 4 were felt all at the same time during the sunset. We do not make a stronger position in confirming this because we only explored a small portion of the circadian cycle. But the data does show an influence at the (virtual) sunset.

An unforeseen and unfortunate conclusion is that our subjects did not enjoy being in a Forest at night, despite it being under the full moon with a fire going nearby. There is a big decrease in *Valence*, even if not statistically significant, and a statistically significant ($p = 0.044$) moderate increase in *Tension*. The subjects corroborated this verbally. Unfortunately, this may have reduced the relaxation capabilities of the experiment as a whole, especially for the subjects who felt fear. (Boyce et al., 2000) points to about 30 lux being enough for the perception of safety, at least in city settings. In exceptional conditions, the moon can reach up to 32 lux. But typical values are way

below those values, which is below the safety threshold. Regardless, we cannot guarantee the output of the head mounted display. At this point, we can only speculate that the lux values may not be the only factor and the forest setting and/or the absence of other humans may be what is actually causing these feelings. It can even have a cultural explanation, as (Dunn and Edensor, 2020) explores.

All of this confirms that Virtual Reality with nature scenarios, can be used to effectively induce a strong feeling of relaxation in people. There are also indications of a possible effect in the circadian cycle.

7. Limitations and Future Work

7.1. Sound

We did not consider the effect of the sound. The sound, by itself can cause effects that could interfere and conflict with our results. This should be addressed in the future by either comparing the effects of sound vs no sound or even completely remove the sound from the experiment. Different sounds could also be explored.

7.2. Comparing with the real life

We did not compare the Virtual Reality against the real life. The idea was considered but was discarded for being too difficult and expensive to take users to a beach and a forest. It would also be nearly impossible to control all the variables due to external influences (including people external to the experience).

7.3. Extra props

Having the users sit on a picnic towel or in a beach stretcher, just like it is represented on the Virtual Reality environment, could have created a stronger effect. This could be interesting to explore since the haptic feedback could add to the effects of the experiences.

7.4. Users Background

We only realized how important the users' background could affect the experiment when a user noted that he really liked the beach scenario due to his childhood. Unfortunately, by this time, it was already too late to fix the experiment protocol.

7.5. More Sceneries

We studied the effects of just two natural environments. More types of natural and even artificial environments could give us different results. One could expect that some environments would be more appealing to different people with different backgrounds and experiences.

7.6. The Heart rate

We limited the users to sitting still, during the experiences. This was a limitation that we set up due to the changes that simply moving around can do to the heart rate. Unfortunately, this can also interfere with our results. On one side, the users are artificially "locked-in-place" and may feel less relaxed and can even break immersion. On the other side, moving around freely could create more variable experiences for each user which would be harder to control and measure.

7.7. One-minute experiences

The choice of 1 minute of Virtual Reality experience is an arbitrary one and we did not base it in any literature. We did, however, choose the time to make up the experience time long enough to expose the users as much as possible without taking too long and negatively affect their feelings (boredom, sleepiness, willing to quit the experiment). We ended up with the value of 1 minute by running some pre-tests and trying to find out when the user was started to get bored with the experiences. With our short, informal sample, most users felt that 2 minutes was a bit too long but were perfectly fine with a 1-minute Virtual Reality experiment.

7.8. MUSE S

The MUSE S equipment used on this study had some technical issues that hindered the experiment. Namely, the calibration was mandatory every time we would log data and it would take, up to 1 minute. We encountered some users (usually with thicker and/or longer hair) that would make the MUSE S very hard to calibrate, prolonging the session behind what should be.

7.9. Sample size and protocol refinement

We feel that the whole study could benefit from a bigger sample size. Even though we did find some interesting results, a bigger sample size could definitely make a stronger case. The experiment protocol could also benefit from a refinement with the experience we gathered on this study.

7.10. Circadian cycle and the night

We, unintentionally, found effects that are, possibly, related to the circadian cycle at the sunset and even at night. This could indicate that Virtual Reality is strong enough to even affect it and it is an interesting venue to explore in the future.

A. Annex

A.1. Data Distribution

Figure 5: Data's Distribution

From Left to Right, Top to Bottom:

a – Muse: Pre-Baseline; Post Baseline; Experiments Average; Beach; Forest.

b – Energetic: Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average.

c – Tired: Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average.

d – Tension: Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average.

e – Calmness: Pre-Baseline; Post Baseline; Experiments Average; Beach; Forest.

f – Valence: Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average.

g – Arousal: Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average.

h - Heart rate: Pre-Baseline; Beach Day; Beach Sunset; Beach Average; Forest Day; Forest Night; Forest Average; Experiments Average; Post Baseline.

A.2. Post-Questionnaire

Q1.I feel that the Virtual Reality environments relaxed me.

Q2.I would enjoy spending more time relaxing in Virtual Reality environments.

Q3.I feel that, relaxing in a Virtual Reality environment could substitute areal-world alternative, in situations where the real-world alternatives are not accessible (e.g.: prisons, remote locations, big concrete cities).

Q4.I would pay to enjoy spending time relaxing in a Virtual Reality environment.

Acknowledgments

This research was partially funded by IDERAM through grants no. M1420-01-0247-FEDER-000019 and M1420-01-0247-FEDER-00003

References

Baños, R., Botella, C., Alcañiz Raya, M., Liaño, V., Guerrero, B., Rey, B., 2005. Immersion and Emotion: Their Impact on the Sense of Presence. Cyberpsychology & behavior : the impact of the Internet, multimedia and virtual reality on behavior and society 7, 734–41. https://doi.org/10.1089/cpb.2004.7.734

Baños, R.M., Liaño, V., Botella, C., Alcañiz, M., Guerrero, B., Rey, B., 2006. Changing Induced Moods Via Virtual Reality, in: IJsselsteijn, W.A., de Kort, Y.A.W., Midden, C., Eggen, B., van den Hoven, E. (Eds.), Persuasive Technology, Lecture Notes in Computer Science. Springer, Berlin, Heidelberg, pp. 7–15. https://doi.org/10.1007/11755494_3

Baus, O., Bouchard, S., 2014. Moving from Virtual Reality Exposure-Based Therapy to Augmented Reality Exposure-Based Therapy: A Review. Front. Hum. Neurosci. 8. https://doi.org/10.3389/fnhum.2014.00112

Bermúdez i Badia, S., Quintero, L., Cameirão, M., Chirico, A., Triberti, S., Cipresso, P., Gaggioli, A., 2018. Toward Emotionally Adaptive Virtual Reality for Mental Health Applications. IEEE Journal of Biomedical and Health Informatics PP, 1–1. https://doi.org/10.1109/JBHI.2018.2878846

Bosse, T., Gerritsen, C., de Man, J., Treur, J., 2014. Towards virtual training of emotion regulation. Brain Inform 1, 27–37. https://doi.org/10.1007/s40708-014-0004-9

Botella, C., Bretón-López, J., Quero, S., Baños, R., García-Palacios, A., 2010. Treating Cockroach Phobia With Augmented Reality. Behavior Therapy 41, 401–413. https://doi.org/10.1016/j.beth.2009.07.002

Boyce, P.R., Eklund, N.H., Hamilton, B.J., Bruno, L.D., 2000. Perceptions of safety at night in different lighting conditions. International Journal of Lighting Research and Technology 32, 79–91. https://doi.org/10.1177/096032710003200205

Bradley, M.M., Lang, P.J., 1994. Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavior Therapy and Experimental Psychiatry 25, 49–59. https://doi.org/10.1016/0005-7916(94)90063-9

Bratman, G.N., Anderson, C.B., Berman, M.G., Cochran, B., de Vries, S., Flanders, J., Folke, C., Frumkin, H., Gross, J.J., Hartig, T., Kahn, P.H., Kuo, M., Lawler, J.J., Levin, P.S., Lindahl, T., Meyer-Lindenberg, A., Mitchell, R., Ouyang, Z., Roe, J., Scarlett, L., Smith, J.R., van den Bosch, M., Wheeler, B.W., White, M.P., Zheng, H., Daily, G.C., 2019. Nature and mental health: An ecosystem service perspective. Science Advances 5. https://doi.org/10.1126/sciadv.aax0903

Bratman, G.N., Hamilton, J.P., Daily, G.C., 2012. The impacts of nature experience on human cognitive function and mental health: Nature experience, cognitive function, and mental health. Annals of the New York Academy of Sciences 1249, 118–136. https://doi.org/10.1111/j.1749-6632.2011.06400.x

Browning, M.H.E.M., Mimnaugh, K.J., van Riper, C.J., Laurent, H.K., LaValle, S.M., 2020. Can Simulated Nature Support Mental Health? Comparing Short, Single-Doses of 360-Degree Nature Videos in Virtual Reality With the Outdoors. Front. Psychol. 10. https://doi.org/10.3389/fpsyg.2019.02667

Buckley, R., Westaway OAM, D., 2021. Women report that nature tourism provides recovery from psychological trauma. Tourism Recreation Research 1–5.

Dunn, N., Edensor, T., 2020. Rethinking Darkness: Cultures, Histories, Practices. Routledge.

Felnhofer, A., Kothgassner, O.D., Schmidt, M., Heinzle, A.-K., Beutl, L., Hlavacs, H., Kryspin-Exner, I., 2015. Is virtual reality emotionally arousing? Investigating five emotion inducing virtual park scenarios. International Journal of Human-Computer Studies 82, 48–56. https://doi.org/10.1016/j.ijhcs.2015.05.004

Ferrer-García, M., Gutiérrez-Maldonado, J., 2011. The use of virtual reality in the study, assessment, and treatment of body image in eating disorders and nonclinical samples: A review of the literature. Body image 9, 1–11. https://doi.org/10.1016/j.bodyim.2011.10.001

Gould van Praag, C.D., Garfinkel, S.N., Sparasci, O., Mees, A., Philippides, A.O., Ware, M., Ottaviani, C., Critchley, H.D., 2017. Mind-wandering and alterations to default mode network connectivity when listening to naturalistic versus artificial sounds. Scientific Reports 7, 45273. https://doi.org/10.1038/srep45273

Gross, J.J., Levenson, R.W., 1995. Emotion elicitation using films. Cognition and Emotion 9, 87–108. https://doi.org/10.1080/02699939508408966

Herrero, R., García-Palacios, A., Castilla, D., Molinari, G., Botella, C., 2014. Virtual Reality for the Induction of Positive Emotions in the Treatment of Fibromyalgia: A Pilot Study over Acceptability, Satisfaction, and the Effect of Virtual Reality on Mood. Cyberpsychology, Behavior, and Social Networking 17, 379–384. https://doi.org/10.1089/cyber.2014.0052

Juan, M.C., Baños, R., Botella, C., Pérez, D., Alcaníiz, M., Monserrat, C., 2006. An Augmented Reality System for the Treatment of Acrophobia: The Sense of Presence Using Immersive Photography. Presence: Teleoperators and Virtual Environments 15, 393–402. https://doi.org/10.1162/pres.15.4.393

Mason, P., 2000. Zoo tourism: The need for more research. Journal of sustainable tourism 8, 333–339.

McLay, R., Ram, V., Murphy, J., Spira, J., Wood, D.P., Wiederhold, M.D., Wiederhold, B.K., Johnston, S., Reeves, D., 2014. Effect of Virtual Reality PTSD Treatment on Mood and Neurocognitive Outcomes. Cyberpsychology, Behavior, and Social Networking 17, 439–446. https://doi.org/10.1089/cyber.2013.0383

Mehra, R., Sharma, V.S., Kaulgud, V., Podder, S., 2019. Fostering positive affects in software development environments using extended reality, in: Proceedings of the 4th International Workshop on Emotion Awareness in Software Engineering, SEmotion '19. IEEE Press, Montreal, Quebec, Canada, pp. 42–45. https://doi.org/10.1109/SEmotion.2019.00016

Parr, H., 2007. Mental health, nature work, and social inclusion. Environment and Planning D: Society and Space 25, 537–561.

Philippot, P., 1993. Inducing and assessing differentiated emotion-feeling states in the laboratory. Cognition and Emotion 7, 171–193. https://doi.org/10.1080/02699939308409183

Plante, T.G., Aldridge, A., Bogden, R., Hanelin, C., 2003. Might virtual reality promote the mood benefits of exercise? Computers in Human Behavior 19, 495–509. https://doi.org/10.1016/S0747-5632(02)00074-2

Riva, G., Mantovani, F., Capideville, C.S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C., Alcañiz, M., 2007. Affective Interactions Using Virtual Reality: The Link between Presence and Emotions. CyberPsychology & Behavior 10, 45–56. https://doi.org/10.1089/cpb.2006.9993

Serrano, B., Baños, R.M., Botella, C., 2016. Virtual reality and stimulation of touch and smell for inducing relaxation: A randomized controlled trial. Computers in Human Behavior 55, 1–8. https://doi.org/10.1016/j.chb.2015.08.007

Thayer, R.E., 1986. Activation-Deactivation Adjective Check List: Current Overview and Structural Analysis. Psychol Rep 58, 607–614. https://doi.org/10.2466/pr0.1986.58.2.607

Yin, J., Yuan, J., Arfaei, N., Catalano, P.J., Allen, J.G., Spengler, J.D., 2020. Effects of biophilic indoor environment on stress and anxiety recovery: A between-subjects experiment in virtual reality. Environment International 136, 105427. https://doi.org/10.1016/j.envint.2019.105427

Yu, C.-P., Lee, H.-Y., Luo, X.-Y., 2018. The effect of virtual reality forest and urban environments on physiological and psychological responses. Urban Forestry & Urban Greening 35, 106–114. https://doi.org/10.1016/j.ufug.2018.08.013

Can you hear the Colour? Towards a Synaesthetic and Multimodal Design Approach in Virtual Worlds

Victoria Wright and Genovefa Kefalidou

School of Informatics
University of Leicester, University Road
Leicester, LE1 7RH, UK
vrw3@leicester.ac.uk, gk169@leicester.ac.uk

Synaesthesia is a phenomenon where senses naturally combine resulting in, for example, 'seeing' music or 'hearing' colours. It is of interest in the field of Human-Computer Interaction as a way of creating new or enhanced experiences and interactions with Mixed Reality technologies. In Virtual Reality, research has mainly focused on evaluating advanced graphics and capturing immersion levels and User Experience within 'typical' and 'expected' interactions. This paper investigates how multimodal design characteristics can lay the foundations to a more 'synaesthetic' design approach in Mixed Reality to identify how 'atypical' interactions can also affect User Experience. 20 participants completed a maze activity, emotion and immersion surveys and interviews. Results suggest a significant increase in surprise, pride and inspiration and a decrease in interest and enthusiasm. The visual and audio aspects were well received by participants and the sensory elements had a positive effect on User Experience. Time perception was measured and 90 per cent of participants' time estimations were longer than the actual time. Change blindness was investigated with most participants not noticing the visual or audio changes. Finally, we discuss how this study can inform future projects which aim to implement a synaesthetic-oriented and multimodal approach in Mixed Reality design.

Multisensory. Virtual Environment. Synaesthetic-oriented approach. Change blindness. Time perception.

1. Introduction

In recent years, Mixed Reality (MR) technologies have become more advanced and more prominent in fields of healthcare (McLay et al., 2014; Striem-Amit, Guendelman & Amedi, 2012), commerce (Van Kerrebroeck, Brengman & Willems, 2017), as well as leisure, with Virtual Reality (VR) headsets such as the Oculus Rift (Oculus, 2019) and Augmented Reality (AR) games such as Pokémon Go (Niantic, 2016). In VR, research has mainly focused on evaluating advanced graphics and capturing immersion levels and User Experience (UX) of interactions that stem out of the 'typical' and 'expected'. The phenomenon Synaesthesia is of interest in the field of Human-Computer Interaction (HCI) as a way of creating new or enhanced experiences and interactions with MR technologies. This paper investigates how multimodal design characteristics can lay the foundations to a more 'synaesthetic' design approach in MR to identify how 'atypical' interactions can also affect UX in such environments.

The study was run remotely and 20 participants navigated around a series of mazes with puzzles. Emotion, immersion and presence were measured by surveys taken before and after the study. This was followed by a semi-structured interview conducted over videoconferencing about their experiences.

2. Related work

The synaesthetic-oriented approach to MR technologies is an underexplored area in HCI. The approach originates from the phenomenon Synaesthesia where people naturally combine senses resulting in, for example, being able to 'see' music or 'hear' colours among other sense combinations (Merter, 2017). The synaesthetic approach itself is a framework which combines sensory elements (Merter, 2017). Jaimes and Sebe (2005) found that multisensory VR research rarely combines the senses simultaneously. Diesendruck et al. (2010) used VR as a verification method to compare a synaesthete's month-space perception to a control group. Our study aims to combine the visual and audio aspects to create a synaesthetic-oriented UX that will inform a design framework for mixed and hybrid experiences in VR/MR.

In an attempt to design novel and gameful experiences in VE/VR/MR, existing in-person escape rooms have provided inspiration. Escape rooms are a series of puzzles solved by players in order to complete set tasks (Nicholson, 2016). Virtual escape rooms are used to perform tasks via different navigation routes and often within certain time constraints. In recent years, multisensory VR escape rooms have been established with the Hyper Reality Experience (2017) and The VOID (2019). Such attractions allow players to physically interact with a VE which has been mapped onto a real location. The player's senses are engaged by features such as a haptic feedback vest and a scent dispenser which activates at certain sections of the narrative (AWE – Augmented World Expo, 2016). Such triggers require the co-location of participants to experience these.

While there is limited research in VR for explicitly embedding 'novel' multimodal approaches, there is a good body of research that acknowledges the value of multisensory design in interventions that support healthcare and promote wellbeing. Multisensory VEs have been used to reduce stress levels (Putrino et al., 2020), have

shown potential for improved quality of life for people with dementia (Cheng, Baker & Dursun, 2019; Sánchez et al., 2013) and provided interventions for anxiety symptoms (Rajasekaran et al., 2011) and the profoundly disabled (Brooks, 2021). Although multisensory design has been embedded in VR/MR there is limited research focus on embedding or facilitating 'fused' sensory experiences, for example, synaesthetic approaches.

Synaesthesia has been used to improve creative ideation by making cards featuring sensory elements which can be combined to inspire novel HCI designs (Lee et al., 2019). The combination of touch and motion has been used to improve possibilities of interaction with mobile devices (Hinckley & Song, 2011). The association between mood and music as well as colour and music has been used to create a new form of music player (Voong & Beale, 2007).

Fire training simulation has been a significant area of VE research. Heaters and a smoke smell which both increase in intensity when approaching the fires in VR have been used to replicate the feeling of being in a burning building (Shaw et al., 2019) promoting enhanced immersion. Shaw et al. (2019) and Smith and Trenholme (2009) have participants exhibiting responses that would not map to responses in a real world fire scenario such as opening doors with smoke coming from underneath them. Smith and Trenholme (2009) also show that training is required so there is not a discrepancy in experience between those who play videogames often and those who do not when in VEs.

Multimodal interaction is another underexplored area of HCI where the multisensory elements are not 'fused'. Schifferstein (2011) created a framework for the design of multisensory experiences which we aim to expand upon in regard to MR and VEs. Colour-speech synaesthesia is based on multi-sensory perception (Bargary et al., 2009) and mirror-touch synaesthesia can affect people's perception of themselves (Maister, Banissy & Tsakiris, 2013). Multisensory perception has inspired art exhibitions (Casini, 2017) and cuisine (Spence & Youssef, 2019). It was hypothesised that a level of synaesthetic response is present in everybody (Casini, 2017; Spector & Maurer, 2013).

Perceptual phenomena such as change blindness have been used in VEs to redirect the user into taking a different path without realising it. For example, in Suma et al.'s study (2011) a participant can enter a virtual room, complete a task in it and proceed to the next room in a corridor. However, in reality, the user is entering and exiting the same room repeatedly as the location of the door is changing while they are occupied with the task. Out of the 71 participants, only one noticed the change and that was when prompted by the researcher. We aim to investigate change blindness within the context of synaesthetic-oriented approaches to see, for example, whether 'fused' UX retains this phenomenon. Are VEs that facilitate 'fused' UX more or less immersive? Do they sustain or promote change blindness? The investigation of such aspects would be valuable for applications in healthcare (supporting/training people with impairment), in transport (designing applications that assist drivers) and in fire and rescue services. Our aim is to see how the synaesthetic-oriented and multimodal approach can be implemented in Virtual Worlds (VEs and VR) and if they result in a new or enhanced (i.e. more 'fused') UX.

3. Methods

3.1. Participants

20 participants were recruited using the snowballing approach (Patrick, Pruchno & Rose, 1998) both within and outside the University of Leicester. As the synaesthetic approach should be accessible to all, it was decided to recruit a wide demographic. There were 11 men and 9 women who participated. The age range was 18-65 with an average age of 36. Out of 20 participants, seven had never played videogames while the other 13 had experience playing apps or computer games.

3.2. Research questions and hypotheses

The research questions for this study are as follows:

What effect do multimodal elements have on immersion and presence in a Virtual Environment?

What effect does a multimodal approach to a Virtual Environment have on user performance?

How does a multimodal approach contribute to perceived experience and levels of immersion in a Virtual Environment?

The first research question is addressed by a quantitative analysis of data whilst the third research question is assessed with qualitative data. Following Witmer and Singer (1998) and Berkman and Akan (2018), presence is defined as feeling present in a VE while immersion focuses on the stimuli creating a feeling of interaction and inclusion within the VE. As higher immersion levels result in higher presence levels, both concepts are being measured simultaneously in this study.

The hypotheses are as follows:

Higher levels of immersion in participants would lead to improved problem solving performances as well as faster navigation.

Higher levels of immersion would lead to higher levels of change blindness.

Participants who notice the changes would be more likely to have an enhanced UX in regard to maze narration.

For example, in the third hypothesis, if the participant notices the changes, they would be less likely to follow the narration's instructions.

3.3. Study design

The design was within-subjects so participants tested every condition with the same sensory elements. Puzzle room is abbreviated to PR.

Table 1: Independent variables

INDEPENDENT VARIABLES	LEVELS	RELATED HYPOTHESES
Colour of maze walls	White, blue or red	(i)
Narrator	Narrator or none	(ii), (iii)
Audio from windows	Birdsong or none	(i)
Location of stairs in regard to PR	None, before, after, before and after	(i)
Door in PR1	Left or right side of room	(ii)
Window view in PR3	Cloudy or sunny	(ii)
Wall colour in PR5	White or green	(ii)

Table 2: Dependent variables

DEPENDENT VARIABLES	MEASURED BY	RELATED HYPOTHESES
Immersion and presence	Questionnaire	(i), (ii), (iii)
Emotions	Questionnaire	(i), (ii), (iii)
Performance	Maze and puzzle completion times	(i)
Participants' responses to multimodality	Interviews over videoconferencing	(i), (ii)

3.4. Study materials

The emotions survey combined the Visual Analogue Scale (VAS) (Riva et al., 2007) and the Positive and Negative Affect Schedule (PANAS) (Watson, Clark & Tellegen, 1988). VAS consisted of seven emotions which the participant had to rank on a scale of one to ten. The PANAS survey included ten positive and ten negative emotions which were ranked from one to five on a Likert scale. The immersion survey was the WAS Presence Questionnaire (Witmer & Singer, 1998) which measured immersive elements and overall immersion/presence (I/P) levels. The surveys were hosted online with Jisc Online Surveys (2020).

While I/P and emotion data was collected through surveys, a log file captured all performance data and all qualitative data came from participant interviews including levels of change blindness.

The VE was hosted online and was created using the 3D modelling software Blender version 2.83 (2020). The images rendered were connected using HTML and JavaScript. To ensure consistency, participants were requested to use Google Chrome on a Windows computer or laptop.

Figure 1 is the medium fidelity VE map. Puzzle rooms are represented in blue. Landmarks are represented by 'LM' and dead-ends are shown by asterisks. 'S' is the start and 'E' is the exit of each maze which are highlighted green.

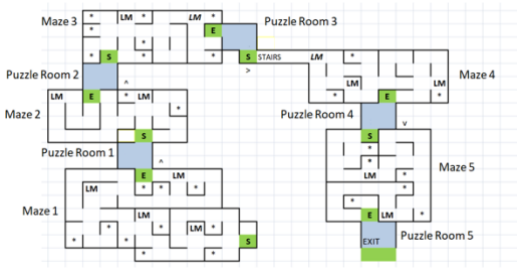


Figure 1: *Medium-fidelity map of the study*

We examined 'typical' multimodality by utilising sensory inputs/outputs such as auditory and visual triggers, while people were asked to perform certain tasks in a VE in an attempt to monitor some 'baseline' figures regarding overall UX and I/P levels and identify whether certain immersion and attention-related phenomena, for example, change blindness and time perception skewing, are present within such settings. We wished to identify whether certain types (audio/visual/tactile such as clicking buttons to perform tasks) triggered specific subjective perceptions (positive or negative). By understanding individual modalities then there can be a better insight of how/when to 'fuse' them to simulate more synaesthetic-oriented UX.

3.5. Study procedure

To begin, participants accessed a website with instructions for completing the mazes and puzzles, a link to the emotion survey and a map for the first maze. Each maze had a collection of landmarks to aid navigation although only the first maze's map was available to participants as it was the largest. To navigate around the maze the participant used either the arrow keys or directional buttons at the top of the screen. Another feature was a help button with instructions. Due to the implementation of the maze, it was not possible for the participant to change their view point.

A notable landmark in the first maze was the open window (Figure 2). As the participant approached the final area of that maze, birdsong would play and become louder as they reached the window.

Figure 2: *The open window in the first maze.*

After the first maze was completed, participants moved onto the first puzzle room (Figure 3). There were four buttons on the table (red, yellow, green and blue) which each had an associated musical note. A four button sequence was played which the participant repeated by pressing the buttons onscreen. They had to successfully repeat four sequences to progress. Each participant had the same order of puzzles. Between the second and third sequences, the room's background changed while the camera focused on the buttons. The door on the left side of the room moved to the right side to measure the participant's change blindness. After the puzzles were completed an animation played of walking upstairs to a door in a first-person view. This was included to add a vertical aspect to the navigation although it only played between certain mazes and puzzle rooms so it didn't become repetitive to the participant.

The second and third mazes were similar in design to the first except the wall colours were red and blue respectively instead of white. The second puzzle room had no changes in the background whilst the third puzzle room had a window with a view of a field on a sunny day which changed to a cloudy day overlooking a cliff.

Figure 3: *The first puzzle room before the door moves.*

The fourth and fifth mazes had landmarks, white walls and included narration which directed the participants to the exit. Participants were not made aware of this feature before completing the study. Narrators changed between mazes. The fourth puzzle room had no background changes. In the fifth puzzle room, the wall colour turned from white to green as the participant solved the puzzles. The duration of the mazes and puzzles was approximately 20-30 minutes.

After the final set of puzzles the participant completed the emotion and I/P surveys as well as downloaded the log file. A short interview with a researcher followed over videoconferencing. The interview was informal and asked about the participant's experiences with the study.

3.6. Data analysis

Qualitative and quantitative data was collected from the participants. Audio recordings of the interviews were transcribed then thematically analysed as described by Braun and Clarke (2006). NVivo 12 was used to assist the annotation of transcripts and the coding of themes. The survey responses and completion times formed the quantitative data. The data was analysed using IBM SPSS Statistics 26. Shapiro-Wilks showed that non-parametric tests were required. The two emotions surveys (before and after the study) were analysed using the Wilcoxon signed-rank test. Correlation tests used Spearman's correlation coefficient and demographics were compared using the Kruskal-Wallis test.

4. results & Discussion

The study's results have been grouped based on the hypotheses from section 3.2.

4.1. Demographics

No correlations were found between either age or gender and any levels of I/P or emotion. Compared to those who played videogames, those who didn't had a more positive response to how compelling the sense of moving around the VE was (p-value = 0.048). Non-videogame players felt more involved in the visual aspect (p-value = 0.006) and audio aspect (p-value = 0.026) of the VE. Their senses were also more engaged (p-value = 0.019).

4.2. Hypothesis 1: Higher levels of immersion in participants would lead to improved problem solving performances as well as faster navigation

Results showed there was no correlation between I/P, problem solving and navigation. Instead, I/P was effected by positive and negative UX. Results showed emotions provoked by problem solving and navigation affected participants' perception of their performance.

4.2.1. Positive UX

Thematic analysis identified the participants feeling a **sense of improvement** over time but not merely as a 'learning curve' phenomenon but also as a more enhanced 'comfort-like' feeling. The non-standard controls initially provided a challenge to participants' I/P however, by the end of the study, they felt more competent using the controls.

'I think I got better at it because I knew what I was doing. The second and third time, I thought "yeah, alright, I've got to come back a bit and then I can look and see what other doorways I've got". So I was learning. I was

learning, you know, all the time about, well, don't just assume that because you've got to go forward to do these things. Look back a bit further and see what your environment actually is and then you can work out where you're going. So, yeah, it was a learning game really.' [P2]

'I quite like the last one that wasn't the vocal one, if that makes sense. On the basis that, by then, you're getting used to it, you know. Instead of the first one, you're thinking and then by the third one you're thinking "yeah, I've got the hang of this now." And you feel more comfortable with it.' [P4]

Colour effects seem to have affected participants' perceived UX and usability (for example, reducing feelings of monotony and promoting task completion), something that can be particularly interesting within the context of synaesthetic experiences – indeed, it is well researched that colour synaesthesia can affect cognitive processes such as memory. For a recent meta-analysis on the field please see Ward, Field and Chin (2019). The different wall colours in the mazes were received positively especially by participants who had spent a long time in the first maze with its white walls; the red walls of the second maze were a relief as it showed they had succeeded in progressing. It was suggested that the VE should be more colourful.

'That was quite nice 'cause the first one was just white and I'd spent so long in that and got quite frustrated so having a variation of colour was quite- was quite welcome.' [P1]

'Change the colour. Change the colour of the stairs or change the colour of the environment' [P13]

'Give it more colour! […] It needed more colour!' [P16]

Landmarks as sensory triggers were positively received by users, especially around navigation. Some participants also expressed enthusiasm for the current design and had suggestions for a more detailed VE including a fire training simulation.

'Yeah, so I feel like if the water jug or something, like the water machine made a bit of a noise like humming or something, just different bits of sort of break up different areas of the maze.' [P11]

'I would choose the first one 'cause there were many objects, it was quite good.' [P13]

'That would be good for the mazes as well if there was a fire and you got smoke.' [P16]

4.2.2. Negative UX

A negative theme was **disorientation and confusion**. Participants who reported being lost in one of the mazes experienced this the most, often due to one of the study's conditions having a lack of landmarks and repetitive brick walls. Participants also struggled navigating due to the inability to change the camera's viewpoint to look around corners.

'So you know the walls were like the plain grey and as I got- I'm a bit- it was quite easy to get lost I suppose.' [P5]

'So I think, yeah, the hardest bits were not, you know, not being able to see around corners as it were, not being able to see behind you' [P7]

'Or there's something of note whereas when you spin round and all you see is blank walls, it's very… yeah, disorientating? Yeah, just feels like you're lost' [P15]

Another theme was **frustration** often stemming from the participant **feeling lost** within one of the conditions, often exacerbated by a feeling of lack of proficiency with the controls. Another form of frustration was the perceived amount of time spent in the mazes. The first maze in particular was seen as frustrating by participants due to its size.

'the frustration of ending up in a blind ending' [P1]

'I mean, there were definitely a couple of them where I was getting a bit frustrated. I was like "I don't know!" Like, "I'm sure I've been down here. (laughs) I don't know how to get out of this maze!"' [P5]

'Probably the first one 'cause I just spent so much time in it. And it's not a nice memory though (laughs).' [P15]

'Oh, the bloody first one. It took me ages to get out of it.' [P16]

4.2.3. Overall change in emotions

Comparing the reported levels of emotions before and after the study showed a statistically significant increase in levels of 'Surprise' (p-value = 0.016), 'Pride' (p-value = 0.033) and 'Inspired' (p-value = 0.034). In the interviews, there was a theme of surprise which was attributed to the unexpectedness of the birdsong, the narration in the fourth and fifth mazes and the participant being surprised by their ability to complete the study. There was also surprise when the participant found out about change blindness however this would not be shown statistically as

the surveys were completed before the interview. The birdsong was regarded as a pleasant surprise as both an indicator of a world outside the maze and as an indication that the maze was almost complete.

‘[…] then I found the window and I was like “Oh! I must be close to the exit now”.’ [P19]

‘[…] then the birdsong things coming louder was a really nice “Oh! This is uplifting finally”.’ [P15]

Participants were not informed about the narration before the study and some were surprised by its inclusion as it provided the correct route through the mazes. One participant believed it to be a mistake that was left in the final study due to its unexpectedness.

‘There was an issue actually, I know you’re going to go into the questions, but when I did the last two mazes, you could hear you giving directions’ [P2]

‘the first one with the verbal instructions ’cause kind of [took me?] by surprise’ [P3]

Participants who initially felt like they may not have the ability to complete the study easily expressed surprise as well as a level of pride and accomplishment afterwards.

‘[…] it worked quite well for me so I had more of a sense of “(pleased) Oh! I’ve done it!” You know, of achievement so it does get your emotions going certainly.’ [P18]

‘I think I was worried I was going to be rubbish so I was like just trying to smash through it.’ [P5]

As well as reported increases in ‘Pride’ and ‘Inspired’ for all participants, non-videogame players reported higher levels of Inspired afterwards than those who play videogames (p-value = 0.034) as well as an increase in positive emotions overall (p-value = 0.008). On the other hand, there was a decrease in reported levels of ‘Interested’ (p-value = 0.01) and ‘Enthusiastic’ (p-value = 0.03) for all participants. Participants who expressed a lack interest often compared the non-standard controls and repeated puzzles negatively to commercial videogames. The average change in statistically significant emotions was taken from the survey responses and can be seen in Figure 4.

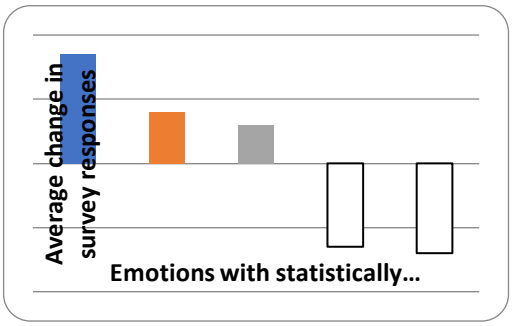


Figure 4: *Changes in emotion from the survey.*

4.3. Hypothesis 2: Higher levels of immersion would lead to higher levels of change blindness

The results showed while there was no correlation between overall I/P and change blindness, there was correlation with individual I/P questions. Furthermore, sensory elements affected I/P positively.

4.3.1. Change blindness

No participants noticed the change in the door’s location in the first puzzle room (Figure 5). Two participants noticed the change in the window’s view in the third puzzle room but only when prompted (Figure 6). Two participants noticed the change in wall colour without prompting (Figure 7). Nine participants noticed the change in narrators and two more thought there was a change but hadn’t realised there were two narrators.

There was a positive correlation between noticing the visual changes of the maze (the door, window and wall colour) and how well the participant felt they could survey the VE using vision (ρ value = 0.484, p-value = 0.031). On the other hand, there was a negative correlation between noticing the audio change (the narration) and feeling involved in the visual aspects of the VE (ρ value = -0.572, p-value = 0.008). In regard to the narration, certain participants mentioned that they stopped observing their surroundings once the narration began.

‘as soon as the voice came over at the top, I stopped using my eyes. I just really was going based on sound. I just kept clicking left, whatever the voice told me to do.’ [P11]

‘I was thinking about it afterwards and I did totally just blindly follow those instructions.’ [P7]

‘ignored what I was looking at entirely and just followed the instruction.’ [P8]

‘I wasn’t really using vision’ [P20]

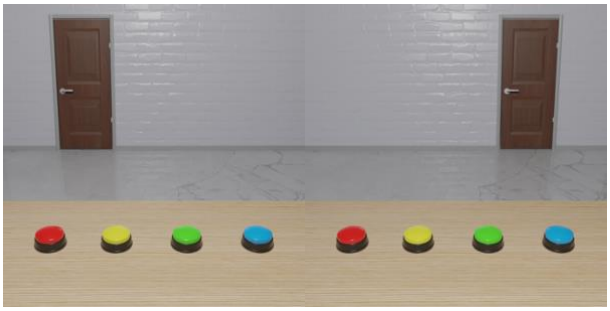


Figure 5: *The door change in the first puzzle room.*

Figure 6: *The window change in the third puzzle room.*

Figure 7: *The wall change in the fifth puzzle room.*

4.3.2. Sensory elements

A positive correlation was found between the sum of positive emotions and how involving the participants found the visual aspects (ρ value = 0.514, p-value = 0.021). Participant interviews emphasised how the landmarks around the maze both helped with navigation and made the VE more visually interesting.

‘[…] ’cause I was using those to try to pinpoint my way around.’ [P5]

‘[…] all those different bits sort of just made the environment seem a much more, instead of just a fake sort of situation, it did actually give it some sort of life to it.’ [P11]

4.3.2.1. Hearing the Song

There was also a positive correlation between the sum of positive emotions and the audio aspects being involving for the participant (ρ value = 0.5, p-value = 0.025). The audio aspects discussed were the birdsong, the narration and the sound during the puzzles. The birdsong was mentioned without prompting by 17 out of the 20 participants. Participants responded positively to the sensory element as illustrated in the quotes below. Only one participant had a negative response finding it ‘quite loud and sudden’ [P12] as the rest of the first maze had no audio cues.

‘That was quite uplifting. That was very nice. (laughs)’ [P1]

‘[…] a feeling of almost, [inaudible] more real feeling, something outside the room rather than just the background of a computer screen.’ [P8]

‘It was really warming to the senses’ [P15]

‘I think it made me feel that I’m not so much in an enclosed space now. I can see out.’ [P18]

The narration was found by participants to be memorable and responses differed depending on how challenging the participants had found the initial three mazes. Participants who had felt lost previously appreciated the additional help whilst others preferred the challenge of navigating themselves now they had a good understanding of the format of the study and its controls.

4.3.2.2. Hearing the Colour

In the puzzle rooms, participants could choose whether they used the visual or audio cues or a combination of the two in order to remember the sequences. 13 out of 20 used only the colours either by visualising them, remembering the words or making the words into an acronym. One participant wrote the sequences down. Six participants used a combination of visual and audio aspects with one participant even stating that they had begun to associate the musical notes to their respective colours in a synaesthetic-oriented manner.

‘It was like colours but kind of like it’s the colour I’d see in my head. And then I think, by the end, I’d gotten used to the sounds that were like associated with it so I was concentrating less on it and I could just like remember it.’ [P20]

Later in the interview, due to the participant’s interest in immersion, Synaesthesia (as a concept) was discussed and they expressed their surprise as they recognised that they had synaesthetic-oriented responses in the past.

‘That’s so weird, oh, my gosh! Yeah ’cause I think definitely, with different instruments, I associate different colours. Like with pianos, I probably would associate like darker colours just ’cause like how the colours of a piano usually is, whereas a guitar is more like colourful maybe? That’s really interesting.’ [P20]

4.4. Hypothesis 3: Participants who notice the changes would be more likely to have an enhanced UX in regard to maze narration

There was no correlation between noticing the changes and whether the participant followed the narrator’s instructions. Instead, experience of videogames and user expectations affected the choice to follow the narration. Perception of UX was effected by realism as well as perceived passage of time.

4.4.1 Perception

Prevalent themes were participants attempting to second-guess the purpose of the study and the VE being a study rather than a commercial videogame effecting their perception. For example, some participants trusted the narration because it was a study rather than a videogame. While changes were being made visually in the puzzle rooms, before being told about this, a few participants thought the sounds were being changed or correct sequences were being rejected.

‘I wonder whether- I’m not sure if you were deliberately changing the sounds a bit.’ [P4]

‘I did the, or at least the first time, I did the typical videogame thing of “don’t go where the person’s telling you” and, like, took a few wrong turns deliberately but, yeah, obviously then, uh, got back and was correct. I was just- I was constantly thinking, well again like too much playing videogames and puzzle games, just expecting something to be thrown in there to get in the way.’ [P6]

‘Yes, that’s something I’d do in a Dark Souls to be honest but not in this game. I was expecting the voices to be honest. […] They don’t usually just mislead you like at the final bit with no warning whatsoever. […] If it was a proper game that I was like just downloaded off the Internet or just, if I got that game on my phone from an app, I’d probably be less inclined to believe it. But I had a feeling that you made it and I didn’t think you’d intentionally fudge the results of this so.’ [P12]

‘I was also thinking that you said you can attempt it as many times as you want but I was kind of conscious like: were you secretly monitoring that? But then the instructions said that that doesn’t really matter so I disregarded that.’ [P14]

‘I thought “we’re going upstairs. We’re going higher. Is that relevant as we go higher up the building?”’ [P16]

User expectations had an effect on their response to the study in regard to a ‘feel for more’ as well as expectations on how realistic or simplistic the VE design should be. Participants sometimes felt like there was a

world outside the maze which could be seen and may be accessible through the window. Some participants also wanted a plot to increase their I/P in the VE.

‘[…] there was the kind of feeling that there was more to it than you saw.’ [P8]

‘Yeah, even if they’re silly [inaudible] a simple plot’s, it’s just a plot in general’s quite nice.’ [P12]

‘I forgot that the goal was the door ’cause that was the first maze room. So I just went to the window and [then/I?] expected like to go into the window but then I looked at the map and then I realised I had to go near the door.’ [P14]

‘Well, yeah, ’cause you don’t know what to expect, do you, so you’ve only got vague instructions. I thought that you could perhaps jump out the window or something.’ [P16]

Realism vs. Simplicity. A **realistic design** was seen by some participants as better and realistic elements increased their I/P. Breaks in realism, such as effectively teleporting between rooms, affected their I/P in a negative sense. On the other hand, some participants felt a simpler design worked well for the purposes of the study.

‘It felt more like realistic ’cause you can hear what is normal like the outdoors.’ [P3]

‘It’s nice to have something to introduce you to the- instead of just appearing at [a/the?] doorway or appearing somewhere you’re going into the room. It makes it slightly more interesting like, you know, somewhere real rather than just appearing in a room with no explanation of why you can’t go back out the door you just came in.’ [P8]

‘it was quite simple actually I suppose but it worked well.’ [P5]

‘I liked how simple it was to be honest ’cause it was easier to navigate something that’s got less distractions.’ [L12]

Lost in time. When remembering the mazes, a theme was time and how it affected which mazes were the most memorable. The first maze was remembered for how long it seemed while any mazes which surprised participants with how quick they seemed were also mentioned.

‘I thought “I couldn’t have done that that quick, could I?”’ [P18]

‘They were sort of memorable ’cause they’re so quick.’ [P8]

‘It seemed like an eternity (laughs).’ [P16]

‘The first bit was- it took me ages’ [P1]

4.4.2.1. Time perception

There was a positive correlation between the sum of positive emotions and the participant losing track of time (ρ value = 0.488, p-value = 0.029). At the end of the study, participants were asked how long they felt they had spent completing the study. The time estimated by participants was longer than the real time spent in the mazes (p-value: 0.001) with 18 out of 20 participants estimating this. There is a positive correlation between the perceived time and how well the participant felt they could examine objects from multiple viewpoints (ρ value = 0.53, p-value = 0.025).

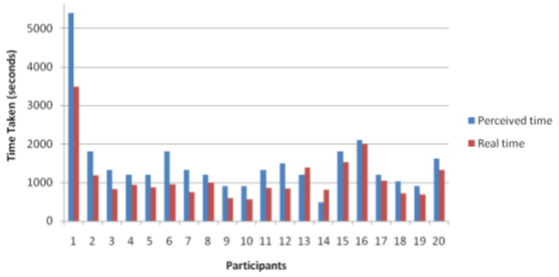


Figure 8: *Perceived time compared to real time elapsed.*

5. General discussion

Our study suggests that visual and audio aspects had a positive effect on UX both statistically and within interviews. Although the multimodal approach was only present in the puzzle rooms, one participant out of 20 showed a synaesthetic-like response between the colours and associated musical notes similar to a natural synaesthete’s association between vocal pitches and colours illustrated by Baron-Cohen, Wyke and Binnie

(1987). While the participant showed some evidence of synaesthetic-oriented tendencies, they were unaware of this before taking part in the study and were surprised when the synaesthetic traits were recognised in their own behaviour. This is a promising area for future research as if puzzle sequence completed 25 times in total can create an unexpected synaesthetic-oriented response, there is a potential for a stronger response with more repetitions or synaesthetic combinations other than visual-auditory. Also, it is important to note that designing synaesthetic-oriented approaches should be experienced by all who participate in the multisensory VEs, not just those who have Synaesthesia as suggested by Casini (2017). Multimodality can potentially support such novel design approaches for MR innovation because the sensory elements do not prerequisite the sensory 'fusion'. This allows the framework of Schifferstein (2011) to be expanded on further in the context of MR and VEs.

Contrary to the second hypothesis, there was no correlation between noticing changes and I/P. This may be because so few participants noticed the changes and the participants who didn't follow the narration were influenced either by how they would act in commercial videogames or believed it to be a distraction technique. There was also no correlation between the overall I/P scores and the changes being noticed. Despite this differing to the hypothesis, we believe this could be a positive result for using change blindness in VEs. In comparison to Suma et al. (2011), the VE in this study was relatively simple and was hosted on a website rather than being in VR. However, a maximum of two participants out of 20 noticed each change. The fact that participants didn't notice large changes in an environment with few landmarks and there was no correlation with I/P suggests that a complicated VE is not required to distract most participants. As long as the participant is focused on a task, it is likely they won't notice changes in front of them. This can be particularly useful in the design of VR training suites as it could potentially bring the cost of an application down if the environments implemented don't need to be complex to be effective.

The levels of focus shown are also illustrated with the narration in the fourth and fifth mazes. Multiple participants mentioned focusing on the audio to the extent that they ignored the visual aspect. This is likely why the change in audio was detected the most out of all the changes and why there was a negative correlation between noticing the audio and the visual aspects of the VE being involving. The percentage of people who did not notice the change in voices was 55 per cent. This result is supported by Vitevitch (2003) who had 42 per cent and then 57 per cent of participants reporting 'change deafness' over two studies where participants had to repeat words said by a voice and, for some participants, the voice changed part way through the list of words. Vitevitch also runs an additional study to check that the voices can be easily differentiated. The narrators in this study were both women in their 20s however some participants who noticed the different voices felt they had slightly different accents. In regard to following the narration, there is already research into multisensory fire training VEs (Wareing et al., 2018; Shaw et al, 2019). Using a voice to indicate the exit could be a useful addition to similar VEs as the act of leaving a building during a fire could be mapped onto participants following the narration in this goal-directed scenario.

Participants often overestimated the time spent on the study. This is contrary to research by Sanders and Cairns (2010) who found that their maze game resulted in participants underestimating the time taken. Block and Zakay (1997) found that people generally underestimate the time taken to complete a task. It is currently unknown why participants in this study overestimated the time taken. A speculation could be that the sensory triggers utilised as 'landmarks' (colour, objects, audio) had indeed an effect on time perception – however, this would need to be further examined. One participant had suggested they had become better at measuring time during lockdown (the study took place during England's second lockdown) however this is only anecdotal and research into the COVID-19 pandemic and its relation to time perception is outside the range of this study. Planned future research will include time perception as a sense in an attempt to understand why this has occurred.

There was a noticeable discrepancy in I/P scores based on participants' experience of videogames. Participants who often played videogames compared the study to commercial games and had a more negative response to the study due to this. In future work, we plan to use standard controls (arrow keys or WASD for movement, mouse for the camera) in order to map to user expectations. A tutorial would be necessary to teach non-videogame players the controls before the main section of the study to bridge the videogame playing skill gap between participants.

5.1. Implications for more 'Fused' Design

Participants discussed 'unexpected' emotional aspects promoted by combining multisensory I/O. Future designs to enhance UX for Virtual Worlds could include mixing multisensory aspects in unexpected ways (more synaesthetic-oriented) to trigger 'positive' surprises. More research is necessary to elicit what new forms for fusing sensory I/O would be more effective and indeed usable for users of Virtual Worlds. Another design research direction is to consider how to best track and transform situation awareness for 'fused' sensory environments to minimise fatigue and support more synaesthesia-oriented experiences.

6. Reflections and future work

COVID-19 impact. An online-based follow-up to this study has been planned. It aims to expand the amount and combination of senses incorporated in this study's VE (visual, auditory, time perception).

Initially, a VR study planned for summer 2020 was postponed due to the COVID-19 pandemic. The VR study featured the combination of the kinaesthetic (movement) sense with visual, olfactory and time perception senses.

The VR study was adapted to fit an online setting highlighting the challenges that VR research experiences under such crisis situations, acknowledging the constraints in implementing more extended multimodality. The kinaesthetic sense was adapted from the user physically walking along virtual corridors to navigating around a virtual maze using arrow keys. Of course, this cannot simulate the dynamics of a physical and broader kinaesthetic perception and experience but the challenges of remote VR for such study designs did not offer many alternatives. This is something that requires further discussion within the HCI community. The visual and time perception senses were simple to adapt however the olfactory aspect was replaced by audio for practicality reasons. Memorisation puzzles remained in both studies.

A website-based study was chosen due to the limitations of remote use of VR technologies. The amount of VR owners is considerably smaller than the amount of people who own an Internet connected computer. Moreover, all participants in an at-home VR study would have videogame experience which may decrease the variety of feedback. This paper's study has shown familiarity with videogames affected participants' UX.

Designing an immersive, multimodal experience without the use of typical VR technologies was an additional challenge. Initially, it seemed that only visual and auditory aspects would be possible with an at-home VE experience. As the results of this study suggest there was an increased level of I/P with these sensory elements, the follow-up online study will incorporate additional senses with the participant receiving a package of sensory props.

An additional constraint was monitoring and supporting participants in a remote study compared to being in a VR lab. A researcher was always available by email and when technical problems occurred, they took longer to solve than being in-person due to not being able to view the participant's screen or participants sometimes lacked the technical vocabulary to explain the issue. A possibility was to observe participants over videoconferencing using screen-sharing. However, this may have effected how participants interacted with the study if they felt they were being observed or judged. Moreover, participants with older hardware may not have been able to screen-share while running the study.

7. References

AWE - Augmented World Expo (2016) *Curtis Hickman THE VOID: Creating The Illusion of Reality.* Available from: https://www.youtube.com/watch?v=Ebwtq1HZJ2A (6 May 2021).

Bargary, G., Barnett, K.J., Mitchell, K.J. and Newell, F.N. (2009) Colored-speech synaesthesia is triggered by multisensory, not unisensory, perception. *Psychological Science*, *20*(5), pp.529-533.

Baron-Cohen, S., Wyke, M.A. and Binnie, C. (1987) Hearing words and seeing colours: an experimental investigation of a case of synaesthesia. *Perception*, 16(6), pp.761-767.

Berkman, M.I. and Akan, E. (2018) Presence and Immersion in Virtual Reality. *Encyclopedia of Computer Graphics and Games*, pp. 1-10.

Blender (2020) *blender.org - Home of the Blender project – Free and Open 3D Creation Software.* Available from: https://www.blender.org/ (6 May 2021).

Block, R.A. and Zakay, D. (1997) Prospective and retrospective duration judgments: A meta-analytic review. *Psychonomic bulletin & review*, 4(2), pp.184-197.

Braun, V. and Clarke, V. (2006) Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2), pp.77-101.

Brooks, A.L. (2021) Interactive Multisensory VibroAcoustic Therapeutic Intervention (iMVATi). *Recent Advances in Technologies for Inclusive Well-Being*, *196*, p.325.

Casini, S. (2017) Synesthesia, transformation and synthesis: toward a multi-sensory pedagogy of the image. *The Senses and Society*, *12*(1), pp.1-17.

Cheng, C., Baker, G.B. and Dursun, S.M. (2019) Use of multisensory stimulation interventions in the treatment of major neurocognitive disorders. *Psychiatry and Clinical Psychopharmacology*, *29*(4), pp.916-921.

Diesendruck, L., Gertner, L., Botzer, L., Goldfarb, L., Karniel, A. and Henik, A. (2010) Months in space: Synaesthesia modulates attention and action. *Cognitive neuropsychology*, 27(8), pp.665-679.

Hinckley, K. and Song, H. (2011, May) Sensor synaesthesia: touch in motion, and motion in touch. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 801-810).

Hyper Reality Experience (2017) *Hyper Reality Experience | Tick Tock Unlock.* Available from: http://hyperrealityexperience.com/ (6 May 2021).

Jaimes, A. and Sebe, N. (2005, October) Multimodal human computer interaction: A survey. In: *International workshop on human-computer interaction* (pp. 1-15). Springer, Berlin, Heidelberg.

Jisc Online surveys (2020) *Online surveys*. Available from: https://www.onlinesurveys.ac.uk/ (Accessed 6 May 2021).

Lee, C.H., Lockton, D., Stevens, J., Wang, S.J. and Ahn, S. (2019, May) Synaesthetic-Translation Tool: Synaesthesia as an Interactive Material for Ideation. In: *Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-6).

Maister, L., Banissy, M.J. and Tsakiris, M. (2013) Mirror-touch synaesthesia changes representations of self-identity. *Neuropsychologia*, *51*(5), pp.802-808.

McLay, R., Ram, V., Murphy, J., Spira, J., Wood, D.P., Wiederhold, M.D., Wiederhold, B.K., Johnston, S. and Reeves, D. (2014) Effect of Virtual Reality PTSD Treatment on Mood and Neurocognitive Outcomes. *CyberPsychology, Behavior & Social Networking*, 17(7), pp. 439-446.

Merter, S. (2017) Synesthetic approach in the design process for enhanced creativity and multisensory experiences. *The Design Journal*, 20(sup1), pp.S4519-S4528.

Niantic (2016) *Pokémon GO*. Available from: https://pokemongolive.com/en/ (6 May 2021).

Nicholson, S. (2016) The State of Escape: Escape Room Design and Facilities. *Meaningful Play* Available from: http://scottnicholson.com/pubs/stateofescape.pdf (6 May 2021)

Oculus (2019) *Oculus Rift S: VR Headset for VR-ready PCs | Oculus*. Available from: https://www.oculus.com/rift-s/ (6 May 2021).

Patrick, J.H., Pruchno, R.A. and Rose, M.S., 1998. Recruiting research participants: a comparison of the costs and effectiveness of five recruitment strategies. *The Gerontologist*, *38*(3), pp.295-302.

Putrino, D., Ripp, J., Herrera, J.E., Cortes, M., Kellner, C., Rizk, D. and Dams-O'Connor, K. (2020) Multisensory, Nature-Inspired Recharge Rooms Yield Short-Term Reductions in Perceived Stress Among Frontline Healthcare Workers. *Frontiers in Psychology*, *11*, p.3213.

Rajasekaran, S., Luteran, C., Qu, H. and Riley-Doucet, C., 2011, January. A portable autonomous multisensory intervention device (pamid) for early detection of anxiety and agitation in patients with cognitive impairments. In: *2011 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 4733-4736). IEEE.

Riva, G., Mantovani, F., Capideville, C.S., Preziosa, A., Morganti, F., Villani, D., Gaggioli, A., Botella, C. and Alcañiz, M. (2007) Affective interactions using virtual reality: the link between presence and emotions. *CyberPsychology & Behavior*, 10(1), pp.45-56.

Sánchez, A., Millán-Calenti, J.C., Lorenzo-López, L. and Maseda, A. (2013) Multisensory stimulation for people with dementia: a review of the literature. *American Journal of Alzheimer's Disease & Other Dementias®*, *28*(1), pp.7-14.

Sanders, T. and Cairns, P. (2010) Time perception, immersion and music in videogames. *Proceedings of HCI 2010 24*, pp.160-167.

Schifferstein, H.N. (2011, October) Multi sensory design. In: *Proceedings of the Second Conference on Creativity and Innovation in Design* (pp. 361-362).

Shaw, E., Roper, T., Nilsson, T., Lawson, G., Cobb, S.V. and Miller, D. (2019, April) The Heat is On: Exploring User Behaviour in a Multisensory Virtual Environment for Fire Evacuation. In: *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (pp. 1-13). ACM.

Smith, S.P. and Trenholme, D. (2009) Rapid prototyping a virtual fire drill environment using computer game technology. *Fire safety journal*, *44*(4), pp.559-569.

Spector, F. and Maurer, D. (2013) Synesthesia: A new approach to understanding the development of perception. Psychology of Consciousness: Theory, Research, and Practice, 1(S), pp. 108-129.

Spence, C. and Youssef, J. (2019) Synaesthesia: The multisensory dining experience. *International Journal of Gastronomy and Food Science*, *18*, p.100179.

Striem-Amit, E., Guendelman, M. and Amedi, A. (2012) 'Visual' acuity of the congenitally blind using visual-to-auditory sensory substitution. *PloS one*, 7(3), p.e33136.

Suma, E.A., Clark, S., Krum, D., Finkelstein, S., Bolas, M. and Warte, Z. (2011, March). Leveraging change blindness for redirection in virtual environments. In: *2011 IEEE Virtual Reality Conference* (pp. 159-166). IEEE.

The VOID (2019) *The VOID | A Virtual Reality Experience*. Available from: http://web.archive.org/web/20200304070516if_/https://www.thevoid.com/ (8 May 2021).

Van Kerrebroeck, H., Brengman, M. and Willems, K. (2017) Escaping the crowd: An experimental study on the impact of a Virtual Reality experience in a shopping mall. *Computers in Human Behavior*, 77(1), pp. 437-450.

Vitevitch, M.S. (2003) Change deafness: The inability to detect changes between two voices. *Journal of Experimental Psychology: Human Perception and Performance*, 29(2), p.333.

Voong, M. and Beale, R. (2007, April) Music organisation using colour synaesthesia. In *CHI'07 extended abstracts on Human Factors in Computing Systems* (pp. 1869-1874).

Ward, J., Field, A. P. and Chin, T. (2019). A meta-analysis of memory ability in synaesthesia. *Memory*, *27*(9), pp. 1299-1312.

Wareing, J., Lawson, G., Abdullah, C. and Roper, T. (2018, September) User Perception of Heat Source Location for a Multisensory Fire Training Simulation. In*: 2018 10th Computer Science and Electronic Engineering (CEEC)* pp. 214-218. IEEE.

Watson, D., Clark, L.A. and Tellegen, A. (1988) Development and validation of brief measures of positive and negative affect: the PANAS scales. *Journal of personality and social psychology*, 54(6), p.1063.

Witmer, B.G. and Singer, M.J. (1998) Measuring presence in virtual environments: A presence questionnaire. *Presence*, 7(3), pp.225-240.

Paper session 2: Design methods 1. Session chair: Julio Abascal

Heuristics for Course Workspace Design and Evaluation

Dorina Rajanen, Atte Tornberg and Mikko Rajanen

University of Oulu

dorina.rajanen@oulu.fi, atte.tornberg@gmail.com, mikko.rajanen@oulu.fi

Course workspace represents the interface between the teacher, the course content, and the student. Both student engagement and student learning are influenced by the course workspace in a similar way as the end-user experience is influenced by system interface. However, course workspaces are typically designed by teachers without a specific input from students, which contrasts the practice in end-user computing and instructional design where users are or should be involved in the early stages of interface design and evaluation. In this paper, we develop a heuristics framework for being applied in a participatory manner involving the student perspective in the evaluation of course workspaces. The framework has been validated on three Moodle course workspaces. The results showed that the framework and the participatory approach provided valuable insights into student experience with the course workspace, while keeping the evaluation effort manageable in terms of data analysis and interpretation. We believe that the heuristics framework and participatory approach could be valuable for teachers, academics, and practitioners that use e-learning platforms for designing course workspaces. The paper provides also examples of improvement areas that have been identified in the evaluation and highlights new research directions in this area.

Heuristic evaluation. Usability. User experience. Student experience. Online education. Student-centred teaching. Course workspace evaluation.

1. Introduction

In the era of online education, course digitalization is commonplace. One element of the course digitalization is the course workspace which represents the interface between the teacher and the course content, the teacher and the student, and the student and the course content. Both student engagement and student learning are influenced by the course workspace in a similar way as the end-user experience is influenced by system interface (see e.g., Meiselwitz & Sadera, 2008; Monari, 2005). However, course workspaces are typically designed by teachers without a specific input from students, which contrasts the practice in end-user computing or instructional design where users are or should be involved in the early stages of system and interface design and evaluation (Preece et al., 2015).

To help teachers create online course environments, there exist numerous electronic platforms, such as Moodle – one of the most popular systems of this kind (Moodle, 2017). According to Moodle developers (Moodle, 2018), the e-learning platform is founded on the pedagogical principles of *social constructivism* (see Palincsar, 1998; Lim & Chai, 2008; Windschitl, 2002), thus by its design the platform provides educators with tools for including the type of resources and activities necessary to build a course that takes into account the student learning and needs. *Social constructivism paradigm* views learning as being an iterative construction of knowledge and meaning as a result of students and teacher interactions, collaborations, and sharing, and Moodle is built upon this paradigm (Nash & Moore, 2014). However, in most cases in higher education, courses' design and implementation are the responsibility of the responsible teacher (possible involving also other teachers). However, the students are very rarely or never involved in the course or workspace design. Typically, the students' input is utilized in the next course implementation based on the feedback given, but rarely this feedback addresses the student experience with the course workspace.

In this paper, we develop a framework for evaluating course workspaces which takes into account specifically the student learning experience and the usability of the course interface. We validated the framework on three Moodle course workspaces by employing a heuristics-based evaluation approach with the aim to incorporate the student view in improving the workspace. The results showed that our evaluation approach provided valuable insights into student experience with the course workspace, while keeping the evaluation effort manageable in terms of data analysis and interpretation; thus, the framework and approach can be applied as a means for improvement of the workspace in different stages of course development. We believe that the framework and approach could be valuable for teachers, academics, and practitioners that use Moodle or other e-learning platforms. The paper thus provides also examples of improvement areas and highlights new research directions in this area of human-computer interaction.

2. Background

2.1 Electronic learning platforms evaluation

Online learning platforms are built as interfaces for teaching or supporting face to face education. Nowadays when remote education and work became a necessity, these platforms also substitute the traditional face to face teaching modality as well as the traditional social interactions for educational purposes. These platforms are built on top of learning management systems (LMSs). They provide different functionalities and views to their users and provide access every time, everywhere, thus enabling synchronous and/or asynchronous communication. The teachers can store and structure the teaching and learning materials, design and accommodate the learning assignments and the group tasks and interactions, provide feedback and grade the students, store the grades and manage the learning activities. Students can access everything that the teachers provide as learning resources including other students' work in the context of various peer tasks, engage in social interactions with their teachers and peer students, submit their work, view the grades, provide feedback and so on.

To be used effectively by teachers and learners, these systems should possess a series of qualities, and one crucial feature is *usability* (Rentróia-Bonito et al., 2008). Usability (ISO, 2010) ensures that the learning system can be used by its users (teachers and students) with effectiveness, efficiency and satisfaction towards attaining their teaching and learning goals, respectively. Still, numerous studies evaluating e-learning platforms indicate problems with usability (see e.g., Chua & Dyson, 2004). Furthermore, there is no single framework agreed upon or established to be used for the evaluation of these systems (Ardito et al. 2006; Chua & Dyson, 2004). There are numerous approaches employed such as the use of questionnaires (e.g., Kakasevski et al., 2008; Senol et al., 2014; Zaharias & Poulymanakou, 2009), observation and interviews (e.g., Ardito et al., 2006), and heuristics (e.g., Reeves et al., 2002; Zaharias & Poulymenakou, 2006). There are also diverse frameworks for defining the features to be assessed. These include standard models of usability such as ISO 9241-210 (ISO, 2010) or ISO/IEC 9126 (ISO, 1991) (see e.g., Chua & Dyson, 2004), established heuristics or models such as Nielsen's model (1993) (e.g., Senol et al., 2014), or customized models created for the purpose of the evaluation (e.g., Ardito et al., 2006; Ozkan & Koseler, 2009; Zaharias & Poulymenakou, 2006; Zaharias & Poulymanakou, 2009).

Nakamura et al. (2017) reviews existing research that evaluates the usability and user experience of LMSs and provide a summary of methods and constructs employed for evaluation. One notable finding was that while a large majority of studies evaluated the *learning factors*, these were formulated in different ways across the studies. Learning factors varied from "content relevance" to "interaction between participants", "feedback and orientation", "instructional assessment", "content organization and structure", "motivation", "support for significant learning approach", "media use", and to "collaborative learning", as well as other dimensions (Nakamura et al., 2017). Based on these findings, we formulate the first proposition regarding the evaluation of workspaces:

Proposition 1: A course workspace design can be characterized by the following learning dimensions *contents' structure, navigation*, *social interaction and collaboration*, *teacher feedback and support.*

The above learning dimensions are consistent with other findings (e.g., Rentróia-Bonito et al., 2008). Furthermore, across studies, evaluation of usability and user experience in the learning context addresses both *technical or interface* issues such as navigation, but also *pedagogical aspects* such as well-structured content and instructions to facilitate learning (e.g., Nokelainen, 2006; Rentróia-Bonito et al., 2008; Squires & Preece, 1996; Zaharias & Koutsabasis, 2012). Usability guidelines regard both the *interaction (dialogue) with the system* (menu, hyperlinks, structure) and the *presentation of the information* (clarity, visibility, colours), and both aspects will have implications to both functional and ergonomic acceptance (van Welie et al., 1999). Therefore, we formulate the second proposition.

Proposition 2: Usability of a course workspace design consists of both *the technical or interface usability* and *the pedagogical usability*.

2.2 Student- or learning-centred evaluation

Student- or learning-centred teaching is the state-of-the-art pedagogy paradigm in higher education (Postareff & Lindblom-Ylänne, 2008; Wright, 2011). The abstract and complex scientific concepts may be counterintuitive to the students and can be different from the currently held world views (Lehtinen et al., 2020; Posner et al., 1982). Therefore, merely transmitting information does not help in understanding the new concepts, and a participatory and collaborative approach to learning is more suitable. Accordingly, teaching is planned to facilitate students' learning processes, rather than to only transmit information and focus on content (Postareff & Lindblom-Ylänne, 2008). The student-centred pedagogy is an approach aligned with the *constructivist view* of teaching and learning (Prawat, 1992). The objective of student-centred teaching is that students learn to form their own knowledge and conceptions about the taught discipline, and eventually change these conceptions when new knowledge, skills, and experience are developed, acquired, and accommodated (Posner et al., 1982; Trigwell et al., 1994). In this conception of teaching, the learning process is effective when the student actively participates in the teaching-learning process (Trigwell et al., 1994). The teacher is responsible of

facilitating this learning process by structuring the educational situations and facilitating peer interactions (Trigwell et al., 1994).

In the context of online education, the socio-digital environment design features such as collaboration, topic structure, and feedback play an important role (Hypponen & Linden, 2009; Lim & Chai, 2008; Siklander et al., 2017). Usability aspects such as ease of use, visual appearance, and multimodality influence student participation and interest (Siklander et al., 2017). The time is an important resource used for students in learning (Hyppönen-Linden, 2009), thus the usability of the workspace plays an important role. An intuitive and easy to use course interface design minimizes the cognitive load of navigating the course contents and provides optimal context for learning the subject matter, avoiding situations where the learner's cognitive resources are expended on ancillary tasks of students finding their ways through the interface (Mehlenbacher et al., 2005; Squires & Preece, 1999). Therefore, we formulate the third and fourth propositions.

Proposition 3: The student-focused teaching strategy is implemented in a course interface design by ensuring that the design of social interactions, computer-mediated dialogue including tasks and assignments, and contents' structure facilitate learning and engage students in the teaching-learning process. Good practices of pedagogical usability include providing feedback, facilitating collaboration, providing collaboration opportunities, good contents' structure.

Corollary of P3: With regards to evaluation, the student-centric evaluation of workspace designs addresses pedagogical usability issues such as social interactions, teacher support and feedback, contents' structure.

Proposition 4: The student-focused strategy is implemented in a course interface design by ensuring that the design of the computer-mediated dialogue is easy to use and intuitive, and that the visual appearance and modalities are suitably designed to facilitate interest and minimize time and cognitive load.

Corollary of P4: With regards to evaluation, the student-centric evaluation of workspace designs addresses interface and technical issues such as easy to use dialogue, intuitive visual designs, easy to use navigation.

2.3 Usability evaluation

Usability is a complex, multi-dimensional, evolving concept reflecting different facets of how users perceive and experience a product, service, or system (Rajanen & Rajanen, 2020). There are various definitions of usability, and practitioners adopt one or another depending on factors such as culture, background, organizational factors, and system development and usability experience and practice (Rajanen et al., 2017). The existing, alternative definitions (e.g., Folmer & Bosch, 2004; Rajanen et al., 2017; van Welie et al., 1999) highlight different aspects of the interaction with a system such as *learnability and freedom from errors* (Nielsen 1993; Shneiderman, 1998), social and organizational contextual aspects such as *impact on the organization* (organizational usability, Hertzum, 2010), the *effectiveness, efficiency, and satisfaction* of using a system (ISO, 1998), *experiential or system-orientated attributes* (ISO, 2010; Kujala et al., 2011; McCarthy & Wright, 2004). The international usability standards define usability as being the *effectiveness, efficiency, and satisfaction* of accomplishing user's tasks in a specific context of use (ISO, 1998). The user satisfaction component of usability evolved into the *user experience concept* (Bevan 2015) with its own standard definition (ISO, 2010). When applying these abstract definitions in evaluation and design practice, usability attributes and indicators or metrics are crucial (Folmer & Bosch, 2004; Marghescu, 2009). With regards to course workspaces, we formulate Definition 1 based on ISO (1998) and identify the usability attributes accordingly in Proposition 5.

Definition 1: Usability of a course workspace is the extent to which students can use the workspace with effectiveness, efficiency, and satisfaction towards completing the course.

Proposition 5: The usability attributes of a course workspace are effectiveness, efficiency and satisfaction of using the workspace towards completing the course.

Folmer and Bosch (2004), inspired by a model of van Welie et al. (1999) define an *intermediary layer* between usability definitions and attributes on the one hand, and the design patterns and heuristics on the other hand. This intermediary layer is represented by the *usability properties* (or *usability means*) and helps in *mapping the low-level usability or user-centred design patterns, principles, and heuristics to higher-level usability attributes or goals* (Folmer and Bosch, 2004; van Welie et al., 1999). Examples of usability properties are consistency, feedback, task conformance, user control, guidance, error management (Folmer and Bosch, 2004; van Welie et al., 1999). These are embedded in design heuristics and principles and ensure that usability goals and attributes are achieved. *As the usability means are not observable in user testing, they are employed within inspection evaluation methods such as heuristic evaluation for diagnosing and improving usability* (van Welie et al., 1999). We formulate the following definitions and the sixth proposition.

Definition 2: Heuristics of a course workspace are low-level propositions that guide the design and evaluation of a course workspace.

Definition 3: Usability means or properties of a course workspace are propositions that link the low-level heuristics to high-level attributes. In other words, the usability properties are means to formulate the low-level heuristics.

Proposition 6: The usability means of a course workspace are not measurable in user testing, but can be assessed using inspection methods such as heuristic evaluation. Thus, heuristic evaluation of a course workspace is employed to evaluate whether the low-level design solutions are aligned with the higher-level attributes. For this, low-level heuristics are formulated and used in inspection as guidelines.

Among the available methods for evaluation, heuristic evaluation (HE) is particularly suitable for formative assessment, and thus useful for improving the course interface. Empirical studies comparing various evaluation methods provide evidence that HE performs well in terms of effectiveness in identifying usability problems and efficiency in using time and human resources and expertise (Davids et al., 2013; Ssemugabi & De Villiers, 2007). Ssemugabi and De Villiers (2007) compared HE with the survey method in the evaluation of a course website. Their findings showed that the two evaluation approaches were similar in performance, yielding a similar number of problems, though each method produced a considerable amount of unique results. HE found more major problems and used a smaller amount of human resources, though those were members of the academic staff while in the survey the participants were students.

Literature reviews on e-learning evaluations (Nakamura et al., 2017; Sagar & Saha, 2017; Salas et al., 2019) provide comprehensive lists of high-level attributes and low-level guidelines to be evaluated, as well as insights into the most common methods of evaluation. However, many of these attributes, such as those related to help functions, do not purport to e-learning environments with which students are already familiar. Thus, the multitude of attributes listed in previous e-learning evaluation studies are on one hand overwhelming by their amount and on the other hand just slightly relevant with regard to a course workspace evaluation. Furthermore, the vast amount of studies and frameworks can provide the same type of confusion as the general usability attributes, in that the same concepts are named differently in different studies, while different concepts get the same or similar names (Folmer and Bosch, 2004). We formulate the following research problem.

Research Problem: There is a need to develop suitable heuristics for course workspaces that map high-level attributes (effectiveness, efficiency, satisfaction) to low-level workspace design solutions (structure of contents, etc.).

2.4 Heuristic evaluation

Heuristic evaluation (HE) was developed in the early 1990s as an answer to the calls from the software industry for a discount usability evaluation method, not requiring the resources and infrastructure needed for the traditional laboratory testing and other traditional usability evaluation methods (Molich & Nielsen, 1990; Nielsen, 1993). The process of HE starts with assembling a group of usability or domain experts, who evaluate the design against a list of individual items called heuristics (Nielsen, 1993). The experts first perform the evaluation individually, and then compare and combine their findings into one list of usability problems and proposed solutions for fixing them (Nielsen & Molich, 1990). HE is a relatively cheap and flexible tool for finding usability issues in all stages of the design process and it can be also conducted by novice evaluators with good results (Nielsen, 1993). Although it is recommended that there are at least five evaluators to uncover 75% of the usability issues, it is possible for also a single evaluator to perform HE, though there will likely be less problems found (Nielsen, 1993). Often these lists of heuristics have been assembled by researchers and practitioners, who try to distil the list into manageable number of best practices to be followed or common pitfalls to be avoided. Today HE is widely used in different domains and there are many lists of heuristics for many different contexts such as software development, web development, game development, as well as online learning.

There have been developed several heuristics that target digital learning platforms, web-based e-learning applications and online courses (see e.g., Albion, 1999; Alsumait & Al-Osaimi, 2009; Ardito et al., 2004; Dringus & Cohen, 2005; Georgiakakis et al., 2005; Mehlenbacher et al., 2005; Mtebe & Kissaka, 2015; Nokelainen, 2006; Reeves et al., 2002; Squires & Preece, 1996; Squires & Preece, 1999; Tolhurst, 1992). Most of these heuristics use as basis for evaluation the Nielsen's ten principles (Nielsen, 1994) and adapt or complement them to fit the purpose of the evaluations. As a consequence, the existing heuristics exhibit a high degree of overlap, which means they will uncover similar usability problems as the Nielsen's heuristics (see e.g., Zaharias & Koutsabasis, 2011). On the other hand, the conceptual basis for the new heuristics is not clarified in most cases and the refined lists vary across the different studies as researchers try to overcome the existing gaps.

These shortcomings make it difficult to select one set of heuristics to use for a particular context, and researchers and practitioners prefer to develop new heuristics to fit their purposes. In an effort to advance the development and validation of usability

heuristics of e-learning systems, Zaharias and Koutsabasis (2011) compared two representative usability heuristics for e-learning, namely those developed by Mehlenbacher et al. (2005) and Reeves et al. (2002). It was found that the two sets uncover similar number and types of usability problems (Zaharias & Koutsabasis, 2011). However, traditional heuristics such as those related to error prevention seem to be inadequate when employed for evaluating e-learning systems, when these are built upon conventional platforms such as Moodle. In the study by Zaharias and Koutsabasis, no such usability problems were found related to error prevention. Thus, the current heuristic lists, while inclusive in covering a multitude of issues including error prevention and accessibility, may appear overwhelming to evaluators as there are very many principles to take into account. Many issues are relevant to the whole environment, service, or program rather than to the individual course design and interface. For example, the visibility of system status heuristic (Nielsen, 1994) is operationalized in terms of usability property as the means of the system to provide information about what is going on or about the success of an operation such as downloading a file (see Reeves et al., 2002). We assume that these types of heuristics, while relevant for the evaluation of a new system or when comparing two systems or platforms, are not useful to be evaluated when the aim is to **improve the usability of a course workspace, especially when the teachers are not able to modify the learning platform behaviour or functionality**. We formulate the seventh and eighth propositions.

Proposition 7: Course workspace heuristics should provide means to identify problems with the course workspace design as opposed to the learning management system design.

Proposition 8: Good heuristics are based on previously validated conceptual models that map high-level attributes to low-level design principles.

In order to answer the Research problem that we formulated above, we developed a framework for the usability evaluation of course workspaces, framework built upon the aforementioned definitions and propositions (Figure 1).

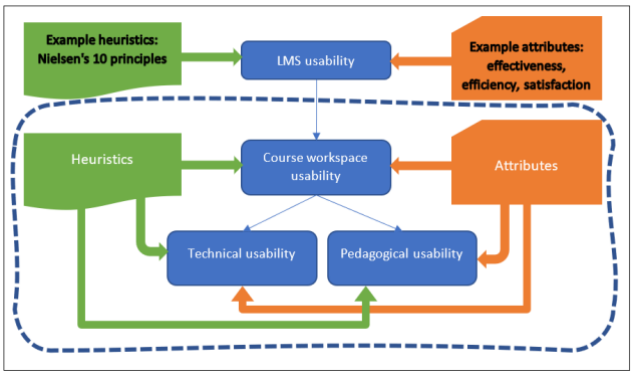


Figure 1: High-level heuristics evaluation framework for course workspace usability (the dashed line separates the general LMS usability from the course workspace usability – the focus of this paper)

Figure 1 illustrates also the relationships between LMS usability and course workspace usability. The high-level view of the framework depicted in Figure 1 highlights the role of heuristics, the two aspects of usability of a learning workspace (technical and pedagogical), and the role of usability attributes in the framework as lenses.

Heuristics Evaluation framework

In this section, we describe the framework at the structural level. The heuristics framework that we propose for evaluating the course workspace usability is largely inspired by the usability model by Doll and Torkzadeh (1988). Doll and Torkzadeh's model addresses the evaluation of the satisfaction (usability attribute level) with information systems, where the quality of information and presentation is critical for the users. This model consists of five usability dimensions, namely ease of use, content, format, accuracy, and timeliness and has been validated (Doll & Torkzadeh,1989). Thus, based on Proposition 8, we construct our framework on this model as it links high-level usability attributes (effectiveness, efficiency, satisfaction) to low-level propositions such as ("the information is clear").

The framework is structured in a layered hierarchy (Folmer & Bosch, 2004; van Welie et al., 1999), where the top level is the highest-level goal, namely the workspace usability (Figure 2). The middle layer represents the usability properties (means, criteria, or dimensions of achieving the top-level goal; Folmer & Bosch, 2004; van Welie et al., 1999). The high-level attributes, namely effectiveness, efficiency, satisfaction, are not represented in the figure to keep the model simple as these are incorporated in the definition of usability. The last layer represents the heuristics which are low-level, actionable principles or guidelines that can be evaluated and acted upon to improve the usability of the workspace.

To incorporate both technical and pedagogical usability, we adapted the original dimensions of the Doll and Torkzadeh's model as follows. Ease of use dimension was divided into *navigation* and *social interaction, as these are two factors identified as being related to learning* (see Proposition 1). In addition, we included the Learning dimension to incorporate additional pedagogical usability aspects. The framework consists of a model with 7 dimensions as follows: Interaction, Navigation, Format, Content, Accuracy, Timeliness, and Learning. Each of these dimensions is assessed through a number of items (called *heuristics*); in total 39 heuristics. Each dimension represents a different aspect of the way the course information is presented and interacted with during the course, and ultimately affects the learning experience. Thus, the framework *focuses on how information is presented and how convenient it is for the student to navigate through different pages, activities and resources (technical usability).* Furthermore, *pedagogical usability aspects that address the learning factors* are also included throughout the framework to ensure that the *content's structure, social interaction and collaboration, teacher feedback and support* facilitate and do not hinder the students' learning (Rentróia-Bonito et al., 2008; Squires & Preece, 1996; Zaharias & Koutsabasis, 2012; Zaharias & Poulymenakou, 2009).

The evaluation model is shown in Figure 2, while the heuristics are presented in Appendix. The evaluation dimensions are described below. Examples of aspects to be evaluated are provided and different dimensions are contrasted.

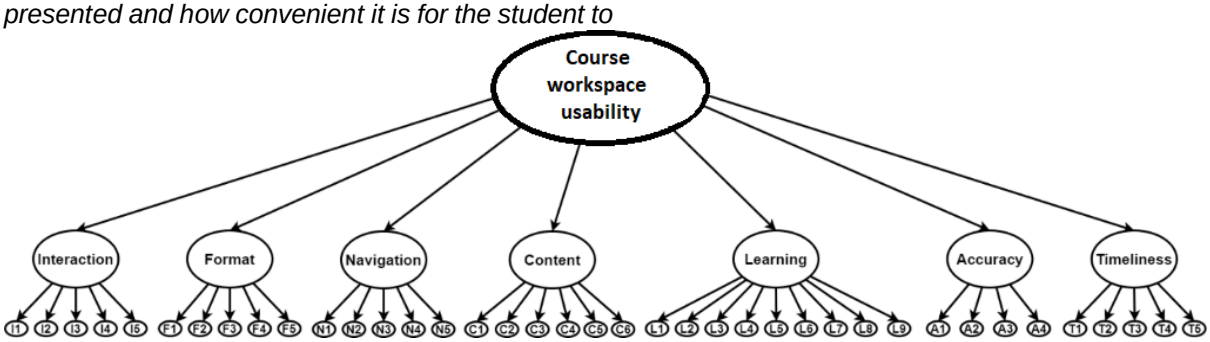


Figure 2: Evaluation framework of the usability of course workspaces

Interaction is about the ease of use of the different tools, tasks and activities of the course that involve social interaction (Doll & Torkzadeh, 1988). For example, here it is evaluated how easy it is to perform a task related to the course such as, submit files for evaluation or feedback, answer questions, or interact with the others. It is not about the level of a difficulty of a learning task, but about the ease of use of the interface or dialogue (Karat et al., 1992): how convenient it is for the student to provide the information requested, to communicate with others (teacher and students), to respond to a given request when needed. Social interaction is also present in various frameworks (see Koohang & du Plessis, 2004; Mehlenbacher et al., 2005).

Navigation is about moving around the course workspace, searching for information from it and browsing its contents (Sagar & Saha, 2017; Zaharias & Poulymenakou, 2006). This is different from Interaction in that here the evaluation is focused on the visibility and accessibility (see Reeves et al., 2002; Koohang & du Plessis, 2004) of the information available in the course workspace and the controls available. Is the information easy to find or quickly available? or are there a lot of user actions required to get to the needed/relevant information? – these are examples of aspects evaluated in this category.

Format is about evaluating the format of the course and the structure of the information presented at the higher level in the course workspace. For example, aspects such as look and feel (Reeves et al., 2002), presentation of information (Koohang & du Plessis, 2004), clarity, structure (Doll & Torkzadeh, 1988), visual design (Zaharias & Poulymenakou, 2006) of the course are evaluated. This is similar with the Navigation criterion in that a good Format would naturally lead to an easy, optimal Navigation, however, in this evaluation the focus is on the visual presentation and the structure of the information (visual layout, Karat et al., 1992), rather than the controls used for Navigation. However, the two evaluations (Navigation and Format) can sometimes lead to similar conclusions, but Format should be more informative on the learning effort and experience (presentation), while Navigation should inform more about the ease of use and the effort spent on workspace operation (users' tasks to operate the workspace).

Content is about evaluating the content provided in the course workspace (Doll & Torkzadeh, 1988). For example, students should not need multiple applications to view the content (minimize the use and effects of modes, Karat et al., 1992), and the information should be complete and well-formatted (Doll & Torkzadeh, 1988) and presented in a consistent way (Karat et al., 1992).

Accuracy is about how accurate the workspace is in that the interface is free of errors that would make a student feel frustration, confusion, and disengagement (Doll & Torkzadeh, 1988). For example, when clicking an option or a link, one would get what was expected. Accuracy is also about tracking if there are any errors in the content, for example when the information is not updated, or information about tasks is not clear or accurate.

Timeliness is about evaluating time-related aspects (Doll & Torkzadeh, 1988) of the course such as schedule, progression, and getting feedback.

Learning is about evaluating the aspects in the workspace that are related to directly supporting the student learning (Zaharias & Poulymenakou, 2006), besides the ones already covered in the previous categories. For example, the presentation of the learning objectives, the structure and variety of the learning tasks, strategies and materials (see also Squires & Preece, 1999), the appropriateness of the scheduling of the tasks and activities (Zaharias & Poulymenakou, 2006). If Timeliness evaluates whether schedules are provided and progression is enabled, Learning evaluates whether the schedules are *properly* defined from the student perspective. Social dynamics (Mehlenbacher et al., 2005) play an important role in constructivist learning, thus they are also included in this category.

Validating the framework

To validate the framework, we applied it to three course workspaces in order to identify usability issues that can be fixed by the teachers or course designers. Heuristics validation studies adopted this method most often (Hermawati & Lawson, 2016).

For validation, we have employed a participatory approach where the evaluation was carried out by a MSc student specialized in usability and user experience design and evaluation. This evaluator profile was selected according to the goal of evaluation, to incorporate both the student view (user task domain expertise) and usability expertise (design domain expertise) (Paz et al., 2018). The evaluation manager represented the pedagogical expertise. Heuristic evaluation using one evaluator with domain experience is an accepted and viable method in usability field (c.f. Hertzum & Jacobsen, 2001) and it has been successfully used in usability studies on distant learning (Erenler, 2018) and social networks (Toribio-Guzmán et al., 2016), among others. The review by Hermawati and Lawson (2016) reports on three studies that use one evaluator for the validation of the heuristics (Carvalho et al., 2009; Jiménez et al., 2013; Salvador & Assi-Moura, 2010) in two different domains health information systems and grid applications. Furthermore, Paz et al. (2018) reports on ten studies using one or two evaluators. The procedure employed in this study ensured that the evaluation was thorough thus finding the most relevant usability issues for students.

The evaluator was representative for the target user group; he had experience with Moodle from other courses, but it was first time when he was exposed to the three course workspaces to be evaluated. The evaluator had participated in one of the three courses in the past but using a different LMS, thus he was familiar with its contents to a certain extent, but he was not familiar with its current implementation and workspace in Moodle. Thus, the context of evaluation resembled a real situation where a student takes the first contact with a course workspace and implementation. Furthermore, the evaluator had knowledge and skills of usability, user experience, and interaction design acquired from his studies. This background was ideal for this type of student-centred evaluation and the existing literature indicates that students are suitable for heuristic evaluation of learning environments (Albion, 1999; Quinn, 1996). Moreover, prior to evaluation, the courses' responsible acted as the evaluation manager and supervisor. The evaluator became familiar with the chosen heuristics as these were designed, collected, assembled, and discussed in a collaborative manner together with the manager of the evaluation. Post-factum, a senior usability researcher with a long-term experience and expertise in heuristic evaluations reviewed both the heuristics and the evaluation results.

The validation process followed the protocol outlined by Reeves et al. (2002) and has been adjusted to fit the current evaluation context and needs. Furthermore, the validation protocol followed the steps outlined by Paz et al. (2018): planning, training, evaluation, discussion, and reporting. The employed protocol included steps such as planning the evaluation, the evaluator getting familiar with the evaluation task and the target course workspaces, the evaluator participating into the literature review and formulation of the heuristics. In the actual evaluation, the evaluator analysed one workspace at a time in light of the defined heuristics. If necessary, the heuristics were slightly refined and discussed with the evaluation manager. The evaluation was conducted in an independent manner, by reviewing thoroughly every page, activity, link, learning material available in the course workspace according to the defined heuristics. The evaluator rated the heuristic on a scale from 1 (poor) to 5 (excellent), similarly with the approach used in other studies (e.g., Albion 1999). Both good experiences and problems were written down in a table, where in the first column were listed the heuristics, in the second column the scores, and in the third column the suggestions for improvement, the problems or the examples of good

practices or experiences. Severity was not assessed, as the evaluation aim was to identify and fix all the problems.

The evaluation results of each course workspace were discussed with the evaluation manager. The discussion included also suggestions to improve the workspace or solve the problems. A report of each evaluation was compiled in a table with the heuristics and the evaluation results found. The reports included the suggestions for improvements.

Validation results and recommendations

The results of the validation are presented in this section by the usability dimensions, namely, interaction, format, navigation, content, learning, accuracy, and timeliness. The detailed heuristics used in evaluation are presented in Appendix.

The evaluation of the Interaction addressed aspects such as submitting assignments (I1), interaction with teachers (I2), interaction with peer students (I3), students giving course feedback (I4, I5, see Appendix). The focus was on evaluating the ease of use of the workspace regarding the above-mentioned aspects, or whether it was possible at all for the workspace to afford those tasks. Generally, all three workspaces seemed well designed with respect to Interaction heuristics. For usability to be optimal, it is important to provide a logical design of the interactions and tasks. For example, the return boxes should be logically placed so that they are easy to access by the students (I1), and the feedback forms to be easier to access and feedback to be encouraged throughout the course (I4 and I5). Furthermore, the face to face activities need to be translated into similar online activities that are easy and meaningful to do.

The evaluation of the Format addressed the following aspects: course structure (F1), course format (F2), visual elements in the course workspace (F3), consistency (F4), and clarity and meaningfulness of the course/workspace structure (F5). There were a few usability problems identified that are worth noting. Long pages with a lot of information could be divided into shorter pages, their structure be improved for example by using weeks to divide the content, and redundant information removed (issues with F1). On the other hand, relevant information for the students regarding the course structure can be directly incorporated in the workspace, rather than on a separate document or be made available in both ways to accommodate different students' needs (F2). Consistent visual patterns in presentation seem to affect the user perception of workspace usability, thus elements such as headings and font formatting should be used consistently throughout the course workspace (F3). Consistency in terms of structure and contents are also important for students; thus, if one page in the workspace differs significantly from the others, the usability can be perceived as low (F4). In cases where the course contents and learning objectives dictate the structure, then the workspace designer should ensure there is appropriate guidance and context provided so that the students are not confused or overwhelmed. However, whenever possible the consistency principle should be applied. The consistency of visual elements seems to affect also the meaningfulness, usefulness and clarity of the workspace (F5).

The evaluation of the Navigation addressed the following aspects: moving around the course workspace (N1), finding information in the workspace (N2), augmenting the workspace with hyperlinks and guiding elements (N3), and remembering things in the workspace (N4). There were pointed out usability problems regarding the navigation through a particularly long page in one of the workspaces (N1, N2). Long pages that are not structured meaningfully and demand students to scroll large portions of text seem to lower the navigation experience and increase the time to find the target information, despite its "ease of use". The use of hyperlinks to connect different pages and activities contributes to usability (N3). Page headers are important for students to navigate through the various pages and activities, and they are part of the guiding elements (N4, N5).

The evaluation of the Content addressed the: completeness of information (C1), formatting of the content (i.e., lectures, downloadable documents; C2), ease of access (C3), video services (C4), and consistency (C6). For an optimal usability, it is important that all content relevant to the students be available directly in the course workspace rather than through links to external pages or services (C1). Thus, embedded videos are preferred to linked videos; however, for having more control over the videos, some teachers prefer to have the videos linked. Usability is also influenced by consistency of lecture materials (C2), though in the case of guest lectures the originality of the layout can be beneficial to students' engagement. Easy access to downloadable resources such as .pdf files can be implemented as descriptive hyperlinks that can be open in a browser for quick scanning, rather than using file uploads that require the students to first download the file and then scan them (C3). Suitable file formats also influence the usability. for reading assignments, the pdf format is preferred, but for writing assignments where a template is required to be followed, this should be provided in an editable format (C4). Formatting of the contents, including the way the video materials are aligned in the page, can bring confusion to students that lead to additional cognitive tasks requiring allocating resources for interpretation and attention.

Thus, ensuring the presentation of the contents is consistent within a page as well as between pages and activities will increase the usability of the workspace (C5).

The evaluation of the Learning addressed the following aspects: presentation of the learning objectives (L1), sequence of the learning content (L2), guidance and support (L3), hierarchical organization that facilitates learning (L4), visual elements that enhance learning (L5), social interaction implemented in tasks and activities (L6), variety, richness and recency of resources (L7-L9). All workspaces provided a list of learning objectives, but to enhance engagement, learning objectives should be more visible and given more attention from teachers, for example, be formatted and positioned as to capture students' interest (L1). The visual means were limited in the workspace according to the evaluation (L5). In all three courses, the lectures provided various types of graphics and pictures to enhance student engagement and understanding, but more visual means are demanded in the workspace. For example, graphics and charts depicting the lectures and content, the different modalities of completing the course, animated videos or games illustrating concepts or approaches, or interactive tools that engage the students in exploration could be implemented.

Social learning occurs when peers and teachers interact and this interaction can have many forms, including peer-reviews, feedback sessions or filling a form. A real dialogic approach would be optimal for social learning, but this is not always possible, for example when students prefer to do the course independently. From the student's perspective, this independent mode that incorporates only peer-reviews may lack social interaction (L6), so more active ways of social interaction can be facilitated whenever possible also for the independent study mode. Question and answer activities are good at enhancing student experience with the course and learning through social interaction (L6).

Variation in resources, exercises, and a flexible approach where students can choose among different tasks and materials increase student's learning experience (L7, L8). Providing different media for lectures (live, video, and text or slides) increases the student learning experience (L7) and accommodates different learning styles and modes (participatory or independent study). However, care should be taken when providing too many resources for similar contents (L8), as students may feel overwhelmed with the information and spend additional time scanning through the resources. Finally, teachers must ensure the learning resources are up to date to meet the student's expectations (L9).

The evaluation of the Accuracy addressed the following aspects: precise and consistent information (A1), precise formulation of the activities (A2), accuracy of hyperlinks or of clicking options (A3), freedom from errors (A4). The accuracy of the workspaces was relatively high; the only problems found were related to some links that did not work in the *test environment* – the workspaces have been duplicated for the test, and some links were directed to pages that were not accessible to the evaluator.

The evaluation of the Timeliness addressed the following aspects: evenly scheduling of the activities (T1), timely feedback (T2), knowing what are the next tasks (T3), following own progression (T4), and updated information (T5). One issue raised by the evaluator was the time of the deadline, and accordingly deadlines set at evening are better than in the morning (T1). Feedback by the teacher was not possible to be evaluated in a test environment, but the evaluator suggested that some information regarding the schedule of the grading could be provided (T2). There was suggested that the contents and tasks in one course to be structured by weeks, so that students would know what tasks they are supposed to do next (T3). Though, a more clean and clear structure would help (see Format and Navigation), another design strategy is to use the calendar option in Moodle that points out the coming activities. Furthermore, for students is important to see how much of the required tasks they have completed. This information can be provided by enabling the Activity Completion feature for the return boxes (T4). The information about schedules and time should be up to date already when the course starts and progresses (T5).

Discussion

The current study presented a framework for evaluating course workspaces from the student perspective. The framework was built based on literature review and definitions and propositions derived from it. Furthermore, the framework built upon the conceptual model of usability described and validated by Doll and Torkzadeh's (1988). This model of user satisfaction with computing systems survived the test of time and proved to be useful for evaluating information-rich interfaces. We have incorporated relevant heuristics regarding educational systems and applications. Many of these were adapted from the Nielsen's heuristics (Molich & Nielsen, 1990; Nielsen, 1993) for software evaluation. However, we selected and adapted those heuristics that fit our evaluation goals. The framework incorporates pedagogical usability as well as the technical usability properties. We validated the framework by applying it to three workspaces using a participatory evaluation approach guided by the protocol

of Reeves et al. (2002) and summarized into the following stages: planning, training, evaluation, discussion and reporting (Paz et al., 2018).

The empirical results showed that heuristic evaluation by student provides valuable insights into student experience with the course workspace, while keeping the evaluation effort manageable in terms of data analysis and interpretation. The evaluation results were also interpreted beyond the reported rating scores and suggestions. Thus, ideas for improvement of the workspaces were provided in the results sections. The recommendations can act as checklists for teachers and course designers in future implementations and new courses.

6.1 Implications for course design practice

Though there exist numerous other frameworks and heuristics for evaluating learning environments, many of them include aspects that are out of the teachers' control such as help and errors prevention and recovery, visibility of system status (e.g., Reeves et al., 2002). Compared to other frameworks and heuristics for evaluating e-learning contexts and applications (e.g., Mehlenbacher et al., 2005), the proposed framework includes only aspects that the teacher or course designer can control, thus, the results provide recommendations that teachers can apply. The framework can also be used in introspection by teachers to design and evaluate their own course.

The participatory approach of involving a student in workspace evaluation proved to be very useful and should be considered by course designers especially when the course is fully digitalized, is a new implementation or has a complex structure. A student specialized in usability was in a double role, as evaluator and representative user as recommended in the literature (e.g., Paz et al., 2018; Sivaji et al., 2011). Depending on the available resources, more students could be involved in the heuristic evaluation with the aim to yield more views on the usability and student experience. This kind of participatory approach (see e.g., Kogi, 2006) would make the students active agents and co-creators of these courses, which would increase their level of engagement (c.f. Naylor et al., 2020).

The framework can also be used during the course. In these "live" evaluations, it is not necessary to apply all heuristics, but to concentrate on those relevant for the teacher at that specific moment. For example, a teacher might at the beginning of the course be interested on the format and navigation heuristics to make sure that complexity of course workspace does not hinder learning, while later on the focus might be on pedagogical aspects such as learning and timeliness. This kind of evaluation feedback would allow teachers to adapt and improve the course workspace as needed and overcome problems before they have an impact on students. Furthermore, the framework and heuristics can be utilized in questionnaire surveys, at the end of the course, where more students can rate the items thus contributing at improving the workspace.

6.2 Implications for research and future work

Research in the areas of e-learning and computer-assisted learning can benefit from the proposed framework. The framework can be adapted to different typologies of courses: fully digitalized, collaborative learning, face-to-face learning.

The individual heuristics can be compared with other lists of heuristics yielding new insights and systematic comparisons for future research and for evaluation or design practice. An open repository of heuristics that can be interactively explored with data-driven approaches such as text mining and visual analytics could be one of the priorities of e-learning heuristics researchers and HCI researchers in general to advance the field through dataset and participatory research contributions (see Wobbrock & Kientz, 2016).

Future research should also address relevant accessibility or universal usability aspects (Shneiderman et al., 2018) such as subtitles in the videos and access to the online learning and video materials for visually impaired. Some heuristics lists do include accessibility items to a certain extent (see Mehlenbacher et al., 2005; Zaharias & Koutsabasis, 2012), however we consider that accessibility should be addressed separately in a dedicated framework that defines the heuristics specifically in relation with specific user needs and use cases (see also Shneiderman et al., 2018).

Hybrid methods employing user tests, textual feedback, and automated evaluation methods (e.g., Adepoju & Shehu, 2014; Sivaji et al., 2013; Sivaji et al., 2017; Ivory & Hearst, 2001) can also be integrated with the proposed framework in future work.

Conclusion

Course workspace acts as the interface between the teacher and the course content, the teacher and the student, and the student and the course content. Therefore, it is vital to evaluate the course workspaces with regards to

learning experience and usability. We developed and tested a heuristics framework on three Moodle course workspaces using a participatory approach. The results showed that heuristic evaluation by a representative student with usability expertise provides valuable insights into student experience with the course workspace, while keeping the evaluation effort manageable in terms of data analysis and interpretation. The heuristic evaluation could be conducted before, during and/or after the course, and be focused on the particular heuristic criteria based on the course type and the course workspace design.

References

Adepoju, S. A., & Shehu, I. S. (2014). Usability evaluation of academic websites using automated tools. In Proc. of the 3rd International Conference on User Science and Engineering (i-USEr) (pp. 186-191). IEEE.

Albion, P. (1999). Heuristic evaluation of educational multimedia: from theory to practice. In Proceedings ASCILITE 1999: 16th Annual Conf. of the Australasian Society for Computers in Learning in Tertiary Education: Responding to Diversity (pp. 9-15). Australasian Society for Computers in Learning in Tertiary Education (ASCILITE).

Alsumait, A., & Al-Osaimi, A. (2009). Usability heuristics evaluation for child e-learning applications. In Proceedings of the 11th international conference on information integration and web-based applications & services (pp. 425-430).

Ardito, C., Costabile, M., Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T. & Rossano, V. (2006). An approach to usability evaluation of e-learning applications. Universal Access in the Information Society, 4(3), 270-283. doi:10.1007/s10209-005-0008-6

Ardito, C., De Marsico, M., Lanzilotti, R., Levialdi, S., Roselli, T., Rossano, V., & Tersigni, M. (2004). Usability of e-learning tools. In Proceedings of the working conference on Advanced visual interfaces (pp. 80-84).

Bevan, N., Carter, J., Harker, S. (2015). ISO 9241-11 revised: what have we learnt about usability since 1998? In: Kurosu, M. (ed.) HCI 2015. LNCS, vol. 9169, pp. 143–151. Springer, Cham. doi:10.1007/978-3-319-20901-2_13

Boudreau, M. C., Gefen, D., & Straub, D. W. (2001). Validation in information systems research: A state-of-the-art assessment. MIS Quarterly, 25, 1-16.

Carvalho, C. J., Borycki, E. M., & Kushniruk, A. (2009). Ensuring the safety of health information systems: using heuristics for patient safety. Healthcare Quarterly 12: 49-54.

Davids, M. R., Chikte, U. M., & Halperin, M. L. (2013). An efficient approach to improve the usability of e-learning resources: the role of heuristic evaluation. Advances in physiology education, 37(3), 242-248.

Doll, W. J. & Torkzadeh, G. (1988). The Measurement of End-User Computing Satisfaction. MIS Quarterly, 12(2), pp. 259-274. doi:10.2307/248851

Dringus, L. P., & Cohen, M. S. (2005). An adaptable usability heuristic checklist for online courses. In *Proceedings Frontiers in Education 35th Annual Conference* (pp. T2H-6). IEEE.

Erenler, T., & Hale, H. (2018). Heuristic evaluation of e-learning. *International Journal of Organizational Leadership, 7*, 195-210.

Georgiakakis, P., Papasalouros, A., Retalis, S., Siassiakos, K., & Papaspyrou, N. (2005). Evaluating the usability of web-based learning management systems. *THEMES in Education*, *6*(1), 45-59.

Hermawati, S., & Lawson, G. (2016). Establishing usability heuristics for heuristics evaluation in a specific domain: Is there a consensus?. *Applied ergonomics*, *56*, 34-51.

Hertzum, M. (2010). Images of usability. *Intl. Journal of Human–Computer Interaction, 26*(6), 567-600.

Hertzum, M., & Jacobsen, N. E. (2001). The evaluator effect: A chilling fact about usability evaluation methods. *International Journal of Human-Computer Interaction, 13*(4), 421-443.

Hyppönen, O., & Lindén, S. (2009). *Handbook for teachers: course structures, teaching methods and assessment*. Publications of Teaching and Learning Development Unit 5/2009. Helsinki University of Technology, Espoo.

ISO International Standardization Organization (1991). *ISO/IEC: 9126 Information technology – Software Product Evaluation – Quality characteristics and guidelines for their use*.

ISO International Standardization Organization (1998) *ISO 9241-11: Ergonomic Requirements for Office Work with Visual Display Terminals (VDTs) - Part 11: Guidance on Usability*.

ISO International Standardization Organization (2010). *ISO 9241-210. Ergonomics of human system interaction – Part 210*: *Human-centred design for interactive systems*.

Ivory, M. Y., & Hearst, M. A. (2001). The state of the art in automating usability evaluation of user interfaces. *ACM Computing Surveys* 33(4), 470-516.

Jiménez, C., Rusu, C., Gorgan, D., & Inostroza, R. (2013). Grid applications to process, supervise and analyze earth science related phenomena: what about usability?. In Proc. of the 2013 Chilean Conf. on Human-Computer Interaction (pp. 94-97).

Kakasevski, G., Mihajlov, M., Arsenovski, S., & Chungurski, S. (2008). Evaluating usability in learning management system Moodle. In Iti 2008-30th international conference on information technology interfaces (pp. 613-618). IEEE.

Karat, C. M., Campbell, R., & Fiegel, T. (1992). Comparison of empirical testing and walkthrough methods in user interface evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 397-404).

Kogi, K. (2006). Participatory methods effective for ergonomic workplace improvement. *Applied ergonomics, 37*(4), 547-554.

Koohang, A., & du Plessis, J. (2004). Architecting usability properties in the e-learning instructional design process. *International Journal on E-learning*, *3*(3), 38-44.

Kujala, S., Roto, V., Väänänen-Vainio-Mattila, K., Karapanos, E., Sinnelä, A. (2011). UX curve: a method for evaluating long-term user experience. *Interact. Comput. 23*(5), 473–483.

Lehtinen, E., Gegenfurtner, A., Helle, L., & Säljö, R. (2020). Conceptual change in the development of visual expertise. *International Journal of Educational Research*, *100*, 101545.

Lim, C. P., & Chai, C. S. (2008). Teachers' pedagogical beliefs and their planning and conduct of computer-mediated classroom lessons. *British Journal of Educational Technology, 39*(5), 807-828.

Marghescu, D. (2009). Usability evaluation of information systems: a review of five international standards. In: Wojtkowski, W. et al. (eds.) *Information Systems Development*, pp. 131–142. Springer, Boston. doi:10.1007/978-0-387-68772-8_11

McCarthy, J., Wright, P. (2004). Technology as experience. *Interactions 11*(5), 42–43.

Mehlenbacher, B., Bennett, L., Bird, T., Ivey, M., Lucas, J., Morton, J., & Whitman, L. (2005). Usable e-learning: A conceptual model for evaluation and design. In *Proceedings of HCI International* (Vol. 2005, p. 11th).

Meiselwitz, G., & Sadera, W. (2008). Investigating the connection between usability and learning outcomes in online learning environments. *Journal of Online Learning and Teaching*, *4*(2), 234-242.

Molich, R., & Nielsen, J. (1990). Improving a human-computer dialogue. *Communications of the ACM, 33*(3), 338-348.

Monari, M (2005). *Evaluation of Collaborative Tools in Web-Based E-Learning Systems.* Unpublished Master's Thesis, Royal Institute of Technology, Stockholm, Sweden.

Moodle (2017) Why is Moodle the world's most widely used learning platform? https://moodle.com/news/moodle-worlds-widely-used-learning-platform/, last retrieved 9.9.2020

Moodle (2018) About Moodle - Pedagogy https://docs.moodle.org/39/en/Pedagogy, last retrieved 4.9.2020

Mtebe, J. S., & Kissaka, M. M. (2015). Heuristics for evaluating usability of learning management systems in Africa. In *2015 IST-Africa Conference* (pp. 1-13). IEEE.

Nakamura, W. T., de Oliveira, E. H. T., & Conte, T. (2017). Usability and User Experience Evaluation of Learning Management Systems-A Systematic Mapping Study. In *International Conference on Enterprise Information Systems* (Vol. 2, pp. 97-108). SCITEPRESS.

Nash, S. S., & Moore, M. (2014). *Moodle Course Design Best Practices*. Packt Publishing Ltd.

Naylor, R., Dollinger, M., Mahat, M., & Khawaja, M. (2020). Students as customers versus as active agents: conceptualising the student role in governance and quality assurance. *Higher Education Research & Development*, 1-14.

Nielsen, J. (1993) *Usability Engineering*. Academic Press, Boston

Nielsen, J. (1994). Heuristic evaluation. In Nielsen, J., and Mack, R.L. (Eds.), *Usability inspection methods*. New York: John Wiley & Sons.

Nielsen, J., & Molich, R. (1990). Heuristic evaluation of user interfaces. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 249-256).

Nokelainen, P. (2006). An empirical assessment of pedagogical usability criteria for digital learning material with elementary school students. *Journal of Educational Technology & Society*, *9*(2), 178-197.

Ozkan, S. & Koseler, R. (2009). Multi-dimensional students' evaluation of e-learning systems in the higher education context: An empirical investigation. *Computers & Education, 53*(4), 1285-1296. doi:10.1016/j.compedu.2009.06.011

Palincsar, A. S. (1998). Social constructivist perspectives on teaching and learning. *Annual review of psychology*, *49*(1), 345-375.

Paz, F., Pow-Sang, J. A., & Collazos, C. (2018). Formal protocol to conduct usability heuristic evaluations in the context of the software development process. *Int. J. Eng. Technol*, *7*(2.28), 10-19.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science education*, *66*(2), 211-227.

Postareff, L., & Lindblom-Ylänne, S. (2008). Variation in teachers' descriptions of teaching: Broadening the understanding of teaching in higher education. *Learning and Instruction, 18*(2), 109-120.

Prawat, R. S. (1992). Teachers' beliefs about teaching and learning: A constructivist perspective. *American Journal of Education*, *100*(3), 354-395.

Preece, J., Sharp, H., & Rogers, Y. (2015). *Interaction design: beyond human-computer interaction*. John Wiley & Sons.

Rajanen, D., Clemmensen, T., Iivari, N., Inal, Y., Rizvanoglu, K., Sivaji, A., Roche, A. (2017) UX professionals' definitions of usability and UX – A comparison between Turkey, Finland, Denmark, France and Malaysia. In: Bernhaupt R. et al. (eds) *Human-Computer Interaction – INTERACT 2017.* LNCS vol. 10516 (16th IFIP TC.13 International Conference on HCI - INTERACT 2017). https://doi.org/10.1007/978-3-319-68059-0_14

Rajanen, M. & Rajanen, D. (2020) Usability: A Cybernetics Perspective. In *Proc. of the 6th International Workshop on Socio-Technical Perspective in IS development (STPIS'20)*.

Reeves, T. C., Benson, L., Elliott, D., Grant, M., Holschuh, D., Kim, B., . . . Loh, S. (2002). *Usability and instructional design heuristics for e-learning evaluation* (pp. 1615-1621). Association for the Advancement of Computing in Education (AACE).

Rentróia-Bonito, A., Martins, A., Guerreiro, T., & Jorge, J. (2008). Evaluating learning support systems usability: An empirical approach. *Communication & Cognition*, *41*(1), 143.

Sagar, K., & Saha, A. (2017). Qualitative usability feature selection with ranking: a novel approach for ranking the identified usability problematic attributes for academic websites using data-mining techniques. *Human-centric Computing and Information Sciences*, *7*(1), 29.

Salas, J., Chang, A., Montalvo, L., Núñez, A., Vilcapoma, M., Moquillaza, A., . . . Paz, F. (2019). Guidelines to evaluate the usability and user experience of learning support platforms: A systematic review. *Communications in Computer and Information Science, 1114*, 238-254. doi:10.1007/978-3-030-37386-3_18

Salvador, V. F. M., & de Assis Moura, L. (2010). Heuristic evaluation for automatic radiology reporting transcription systems. In 10th International Conference on Information Science, Signal Processing and their Applications (ISSPA 2010) (pp. 292-295). IEEE.

Senol, L., Gecili, H., & Durdu, P. O. (2014). Usability evaluation of a moodle based learning management system. In EdMedia+ Innovate Learning (pp. 850-858). Association for the Advancement of Computing in Education (AACE).

Shneiderman, B. (1998). *Designing the User Interface*, Addison-Wesley Publishing Company, USA.

Shneiderman, B., Plaisant, C., Cohen, M. S., Jacobs, S., Elmqvist, N., (2018). *Designing the User Interface: Strategies for Effective Human-Computer Interaction*. 6th Ed. Pearson.

Siklander, P., Kangas, M., Ruhalahti, S., & Korva, S. (2017). Exploring triggers for arousing interest in the online learning. In *International Technology, Education and Development Conference* (pp. 9081-9089).

Sivaji, A., Abdullah, A., & Downe, A. G. (2011). Usability testing methodology: Effectiveness of heuristic evaluation in E-government website development. In *Proc of the 5fth Asia Modelling Symposium* (pp. 68-72). IEEE.

Sivaji, A., Abdullah, M. R., Downe, A. G., & Ahmad, W. F. W. (2013). Hybrid usability methodology: integrating heuristic evaluation with laboratory testing across the software development lifecycle. In *Proc. of the 10th International Conference on Information Technology: New Generations* (pp. 375-383). IEEE.

Sivaji, A., Nielsen, S. F. & Clemmensen, T. (2017) A textual feedback tool for empowering participants in usability and UX evaluations. *Int. J. Human–Computer Interact. 33*(5), pp. 357–370, 2017.

Squires, D., & Preece, J. (1996). Usability and learning: evaluating the potential of educational software. *Computers & Education*, *27*(1), 15-22.

Squires, D., & Preece, J. (1999). Predicting quality in educational software. *Interacting with computers*, *11*(5), 467-483.

Ssemugabi, S., & De Villiers, R. (2007). A comparative study of two usability evaluation methods using a web-based e-learning application. In *Proceedings of the 2007 annual research conference of the South African institute of computer scientists and information technologists on IT research in developing countries* (pp. 132-142).

Tolhurst, D. (1992). A checklist for evaluating content-based hypertext computer software. *Educational Technology, 32*(3), 17-21.

Trigwell, K., Prosser, M., & Taylor, P. (1994). Qualitative differences in approaches to teaching first year university science. *Higher Education, 27*(1), 75-84.

Toribio-Guzmán, J. M., García-Holgado, A., Pérez, F. S., García-Peñalvo, F. J., & Martín, M. A. F. (2016). Study of the usability of the private social network SocialNet using heuristic evaluation. In Proceedings of the *XVII International Conference on Human Computer Interaction* (pp. 1-5).

van Welie, M., van der Veer, G. C., & Eliëns, A. P. W. (1999). Breaking down Usability. In Human Computer Interaction-Proceedings of Interact 99, 30th August-3rd September 1999, Edinburgh, Scotland (pp. 613-620). IOS Press.

Windschitl, M. (2002). Framing constructivism in practice as the negotiation of dilemmas: An analysis of the conceptual, pedagogical, cultural, and political challenges facing teachers. *Review of Educational Research*, *72*(2), 131-175.

Wobbrock, J. O., & Kientz, J. A. (2016). Research contributions in human-computer interaction. *Interactions, 23*(3), 38-44.

Wright, G. B. (2011). Student-centered learning in higher education. *International Journal of Teaching and Learning in Higher Education*, *23*(1), 92-97.

Zaharias, P. & Koutsabasis, P. (2012). Heuristic evaluation of e-learning courses: A comparative analysis of two e-learning heuristic sets. *Campus-Wide Information Systems, 29*(1), 45-60. doi:10.1108/10650741211192046

Zaharias, P. & Poulymenakou, A. (2006). Implementing learner-centred design: The interplay between usability and instructional design practices. *Interactive Technology and Smart Education, 3*(2), 87-100. doi:10.1108/17415650680000055

Zaharias, P. & Poulymenakou, A. (2009). Developing a Usability Evaluation Method for e-Learning Applications: Beyond Functional Usability. *International Journal of Human-Computer Interaction, 25*(1), 75-98. doi:10.1080/10447310802546716

Appendix. Evaluation criteria and heuristics

Heuristic	**Example references (c.f.)**
Interaction	Doll & Torkzadeh 1988 (ease of use); Mehlenbacher et al. 2005 (social dynamics)
I1: It is easy to submit assignments.	Monari 2005; Senol et al. 2014
I2: It is easy to interact with the teacher.	Monari 2005; Senol et al. 2014
I3: It is easy to interact with other students when needed, e.g. in tasks designed for this purpose.	Mehlenbacher et al. 2005; Monari 2005; Senol et al. 2014
I4: It is easy to give course feedback.	Mehlenbacher et al. 2005; Monari 2005
I5: There are ways to interact with the teachers to give and receive feedback.	Mehlenbacher et al. 2005; Monari 2005; Senol et al. 2014

Format	Doll & Torkzadeh 1988
F1: The course structure is easy to understand.	Karat et al. 1992; Koohang & du Plessis 2004
F2: The course format is clear.	Doll & Torkzadeh 1988
F3: The utilized visual elements in the course workspace are consistent.	Karat et al. 1992; Reeves et al., 2002; Senol et al. 2014; Sagar & Saha 2017
F4: Different pages in the course workspace have similar structure.	Karat et al. 1992; Sagar & Saha 2017; Zaharias & Poulymenakou 2009
F5: The structure of each page/ section in the course workspace is good (meaningful, useful) and clear.	Reeves et al., 2002; Sagar & Saha 2017
Navigation	Reeves et al., 2002; Sagar & Saha 2017
N1: It is easy to move around the course workspace.	Senol et al. 2014
N2: You can quickly find what you want from the course workspace.	Senol et al. 2014; Zaharias & Poulymenakou 2006
N3: The course workspace provides hyperlinks to things referred.	Sagar & Saha 2017
N4: It is easy to remember where you are in the course workspace	Monari 2005; Reeves et al., 2002; Senol et al. 2014;Zaharias & Poulymenakou 2006
N5: The course workspace provides guidance.	Koohang & du Plessis 2004; Reeves et al., 2002; Senol et al. 2014; Zaharias & Poulymenakou 2009
Content	Doll & Torkzadeh 1988
C1: The course workspace provides all the information needed for completion of the course.	Doll & Torkzadeh 1988; Sagar & Saha 2017
C2: The content is well formatted.	Doll & Torkzadeh 1988
C3: The content is easy to view.	Senol et al. 2014
C4: Downloadable documents are in appropriate format.	Karat et al. 1992; Sagar & Saha 2017
C5: Video content is provided through one service.	Karat et al. 1992; Sagar & Saha 2017
C6: The content format is consistent.	Karat et al. 1992; Sagar & Saha 2017; Zaharias & Poulymenakou 2009
Learning	Zaharias & Poulymenakou 2006
L1: The learning objectives are clearly presented.	Ozkan & Koseler 2009; Zaharias & Poulymenakou 2006, 2009
L2: Learning content is sequenced properly.	Zaharias & Poulymenakou 2006
L3: Learners' guidance and support is provided.	Ozkan & Koseler 2009; Zaharias & Poulymenakou 2006, 2009
L4: The hierarchical organization of the course facilitates learning.	Ozkan & Koseler 2009; Zaharias & Poulymenakou 2006

L5: The use of visual means in the workspace enhances learning.	Ozkan & Koseler 2009; Squires & Preece 1996;Zaharias & Poulymenakou 2006, 2009
L6: There are enough social learning tasks and activities implemented in the course.	Mehlenbacher et al. 2005; Zaharias & Poulymenakou 2006, 2009
L7: The course resources are varied.	Squires & Preece 1999; Zaharias & Poulymenakou 2006
L8: The course resources are plentiful.	Squires & Preece 1999; Zaharias & Poulymenakou 2006, 2009
L9: The course resources are up to date.	Ozkan & Koseler 2009; Sagar & Saha 2017; Zaharias & Poulymenakou 2009
Accuracy	Doll & Torkzadeh 1988
A1: The information in the course workspace is precise and consistent.	Doll & Torkzadeh 1988; Karat et al. 1992
A2: The activities in the course workspace are precisely formulated.	Doll & Torkzadeh 1988; Ozkan & Koseler 2009
A3: When clicking an option, you get what you expect.	Sagar & Saha 2017; Senol et al. 2014; Zaharias & Poulymenakou 2009
A4: The workspace is error free.	Sagar & Saha 2017; Senol et al. 2014; Zaharias & Poulymenakou 2006, 2009
Timeliness	Doll & Torkzadeh 1988
T1: The course activities are scheduled evenly.	Ozkan & Koseler 2009
T2: Feedback is provided timely.	Mehlenbacher et al. 2005; Ozkan & Koseler 2009
T3: It is easy to know what you are supposed to do next in the course.	Senol et al. 2014; Zaharias & Poulymenakou 2006, 2009
T4: It is easy to follow your progression in the course.	Mehlenbacher et al. 2005; Ozkan & Koseler 2009; Zaharias & Poulymenakou 2006, 2009
T5: Information provided is up to date.	Doll & Torkzadeh 1988; Ozkan & Koseler 2009; Sagar & Saha 2017

A Design Space for Memory Augmentation Technologies

Madeleine Steeds and Sarah Clinch

University of Manchester
Manchester, UK
madeleine.steeds@manchester.ac.uk, sarah.clinch@manchester.ac.uk

The pervasive nature of display technologies can enable novel ever-accessible memory aids to address deterioration caused by ageing and cognitive decline. To date, however, memory has largely been treated as a single-unit, and there has been little formal consideration of how to select the most appropriate technology for a given intervention. We build on existing domain knowledge from neuroscience and psychology to suggest a novel design space with two axes: processing level, and display modality. In particular, we consider how augmentations might intervene at a biological, cognitive or meta-cognitive level using head-mounted (private) displays, small-scale (personal) displays, larger public and semi-public displays, and with technology that bypasses the visual channels entirely (e.g. through neural stimulation or non-visual senses). We then provide examples of potential studies to explore these design areas, and discuss future directions this approach to memory augmentation may take. Consideration is also given to the ethics of memory augmentation.

HCI. Cognition. Cyberpsychology. Memory. Displays. Ethics.

1. Introduction

Technology and tools to address physical deterioration that emerges as a result of age or illness are commonplace. More recently, researchers have begun to express a similar vision for technology use to address limitations in cognitive function, including memory (Chen and Jones, 2010; Davies et al., 2015; Harvey et al., 2016; Hodges et al., 2011; Hodges et al., 2006; Iwamura et al., 2014; Le et al., 2016; Mikusz et al., 2018; Rhodes, 1997; Schmidt, 2017). Most commonly, lifelogging devices and other data sources are used to provide cues to help rehearse personal experiences, known as episodic memories (Harvey et al., 2016; Hodges et al., 2011; Hodges et al., 2006; Le et al., 2016). Other memory augmentations tackle concepts such as prospective (Rhodes, 1997) or procedural memory (Seim et al., 2014; Seim et al., 2015).

Despite addressing a variety of types of memory, these augmentations have typically focussed on in-the-moment experiences of memory, particularly for episodic memories. Other understandings of memory drawn from a variety of disciplines (including psychology, neuroscience, philosophy and sociology) are yet to be used as the foundation for memory technologies. This paper aims to formulate a design space that considers the medium through which memory augmentation is presented, and the aspect of memory being augmented. Through this design space we identify where previous memory augmentations have targeted, and where future research may take place. We posit that the design space for memory augmentation is far broader than is reflected in existing literature. Considering memory augmentation through the lens of this design space will allow for further investigation into effective memory augmentation techniques, which have previously been overlooked.

We then consider the unexplored areas of the design space, and present potential studies which would serve as a first step in addressing these research areas. We also discuss ways the design space could be expanded to allow for concepts of memory in other disciplines, as well as some ethical considerations to be made when researching and implementing human memory augmentation.

2. The Design space

We suggest two concrete dimensions as a foundation for a taxonomy of existing systems. Firstly, processing level – the conceptual level at which a given augmentation is operating:

Biological: augmentation on a neural level by impacting neurons, neurotransmitters etc.;

Cognitive: augmenting memory on a case by case basis, augmentation in the present;

Meta-cognitive: augmenting the techniques to memorise, and the monitoring of cognitive abilities.

Technology may be used to impact one of these levels, or potentially multiple levels. A survey by Madan (2014) gave an overview of approaches to augmenting memory, which focused on techniques which fit into these processing levels. Two examples affecting the biological processing of memories were given- nootropics and brain stimulation. While these are quite different techniques, the former involving the taking of pharmacological agents such as caffeine, and the latter involving the stimulation of neurons, they both impact a person's biological response to memory stimuli. Thus in the present design space, we group these together as impacting the biological level.

Madan also gives an example of augmentation on the cognitive level- that of external aids. These are aids which support cognition in the present, which in our design space represents the cognitive level. This in-the-moment augmentation is what most would consider when thinking of technological, memory augmentation.

Madan's final example is the use of mnemonics. This is a memory technique to aid cognition which people can utilise. We classify this as the meta-cognitive level. Meta-cognition is the term given to thinking about thinking. In terms of our design space, we consider it the use of technology to augment memorisation techniques, or the monitoring of cognitive abilities.

Our second axis is display modality. Whilst not all memory interventions are visual, the majority are, and displays have played a significant role to date. We therefore build on prior classifications of displays (Muller et al. 2010) to categorise this axis as follows:

Non-displays: internal (implantable) and external technology which does not provide visual feedback to the user;

Private/Head-mounted displays: technology with immersive properties which are attached to a single person;

Personal displays: smaller displays, which although potentially shareable, are designed for individual use;

Semi-Public and Public displays: technology targeting a wide audience.

We focus this design space on visual mediums as they are pervasive within HCI, however this is a limitation in terms of working with visually impaired users. While devices such as haptics are included in the non-display level (e.g. Seim et al., 2014), there is no differentiation between those external devices and implantable devices. As such it may be necessary to expand the definition of 'non-displays' to better highlight the use of other sensory displays, such as audio cues and haptics (for further discussion of this please see Section 5).

3. Trends in memory augmentation

On reviewing existing literature in memory augmentation technology, we can map these works onto our two axes, creating 12 distinct spaces for augmentation. In **Figure 1** we can see these distinct spaces, and where existing work tends to cluster within this model.

On the biological level, memory prosthetic research has begun with the aim of creating implantable devices (non-displays) to aid those with memory impairment (Solis, 2017). Recently, success has been seen in memory implants in epileptic patients (Hampson et al., 2018), although these implants were later removed and so the long-term feasibility of this approach is yet to be understood. Currently, work augmenting biological processes involves invasive procedures and does not utilise existing, readily-available technology. Similarly, work in this area has largely ignored head-mounted displays such as Virtual Reality (VR), Augmented Reality (AR) or Mixed Reality (MR), despite the technology's immersive properties which may be more stimulating than other displays.

More research has been conducted on the cognitive level. For example, Mikusz et al. (2018) used large displays around campus to support student learning of lecture content. While limitations were found regarding the extraneous variables, the study successfully utilised public displays to augment memory on a cognitive level. Likewise, Dingler et al. (2016) placed memory displays in the homes of students revising for exams which were found to encourage participants to study. These displays were in the form of tablets but due to the way they were displayed in the participant's homes, they had the potential to reach multiple occupants placing them on the intersection between a personal and semi-public display. There have also been some studies on this level utilising non-displays (specifically haptics).

The meta-cognitive level is also relatively sparse compared to the cognitive level, with few interventions targeting these skills. An exception to this is the work by Yang et al. (2020) where they use VR to aid participants in utilising the 'memory palace' (method of loci) memory technique. This is when an individual imagines a location and places the items to be remembered around the location.

Mapping these prior works into our design space [**Figure 1**] we can see the areas of augmentation which have received limited attention, and require further research. In terms of processing level, bio-logical augmentation has been largely overlooked,

***Figure 1**: A chart of the design space, populated with existing research in these areas.*

and the main contribution to this level has been invasive technologies. While the task of externally augmenting biological processes is not an easy one, if achieved, it would aid in bypassing the ethical concerns that invasive augmentations raise.

4. Exploring the design space

The design space suggests that there are two processing levels that have received limited attention: the biological and meta-cognitive. In terms of display modality, personal displays are the most represented within the research, with few studies investigating the other modalities. In this section, we will discuss some potential ways the under-represented areas of the design space could be investigated in future work.

The work in the biological level utilises technology implanted into people to increase the presence of neurotransmitters which aid in the formation of memories. To do this non-invasively, displays would need to elicit the natural production of these key neurotransmitters such as dopamine, serotonin and norepinephrine (Handra et al. 2019). If the display can show an image or give an experience which would elicit an emotion known to stimulate the release of one of these neurotransmitters, then this could augment the biological level in a non-invasive way. However, considerations would need to be made regarding the ethics of such stimulation, particularly in cases where norepinephrine is being stimulated as this is closely associated with fear memory.

The meta-cognitive level is also sparsely researched. The studies at this level have focussed on the method of loci memory technique, by creating memory palaces through VR. However this is a highly individualised set up and may not be suitable for those who experience motion sickness (a common side effect of VR). As such, meta-cognitive augmentation may lend itself to the use of semi-public and public displays. This could be done by expanding upon the study by Mikusz et al. (2018) where they used public displays to deliver memory cues to students across the university campus. However, instead of just delivering the cues, this system could be used to create a campus-wide memory palace. The displays could present the relevant information but actively tie it to the location to encourage the training of this metacognitive technique.

Despite the recent increase in accessibility of head-mounted/private displays, there is still limited work in using these for cognitive memory augmentation purposes. These displays have been suggested to give participants greater senses of immersion compared to desktop computer displays (Shu et al., 2019) and this feeling may create a better learning environment for memory than public or personal displays. Therefore, future research may wish to investigate whether existing cognitive memory aids can be transferred to a head-mounted display. AR may be particularly useful in this regard as push notifications may be presented to the user, without them having to direct their attention away from the environment. For example, an AR shopping list could help the user remember what they needed from each aisle by providing unobtrusive prompts as they navigate the shop.

5. Future Directions and Ethical augmentation

The ideas in Section 4 are just some examples of future directions for memory augmentation. However, they show the value of mapping relevant research to this design space, as we are able to generate ideas targeting novel research areas. By utilising this design space, future research may identify novel methods of memory augmentation, which will lead to a holistic picture of the ecosystems that best augment human memory. This may

then lead the design space to serve practitioners. For example, as dopamine transmission in the brain reduces with age (Bäckman et al., 2006), the practitioner may find interventions for older adults more effective if they augment the biological level to address this deficiency.

Further ideas may still come from the evolution of the design space itself. As noted in Section 2, the present design space does not distinguish within the category of non-displays, leading to invasive technologies such as neural prosthetics (e.g. Hampson et al., 2018) to be grouped with non-visual, sensory technology such as haptic devices (e.g. Kuznetsov et al., 2009). As research into those fields develops, it may be beneficial to better distinguish between these, as the applications and accessibility of such technologies vary greatly. Further to this, there is some overlap between personal displays and semi-public displays, as exemplified in the work of Dingler et al. (see Section 3). As such, increasing the level of detail explored in the display modality axis may allow for stronger distinctions to be made between these domains.

The processing levels described in the design space are also limited to interpretations from neuroscience and psychology, and as such primarily address the individual. Current understandings of memory are, however, much richer than this — spanning many disciplines (e.g. sociology, philosophy, cultural studies). Whilst these understandings are yet to be the focus of most memory augmentations, as technology develops, it may be valuable to expand the design space to encompass these interpretations of memory. This could be through the addition of new processing levels (e.g. collective memory), or through the addition of new axes. This could enable human memory augmentation on a scale previously thought unachievable, for example by augmenting group memory to aid in the standardisation of technical skills, to reduce the risk of human error.

However, the question of memory augmentation, particularly as it becomes more commonplace within society raises ethical issues. Throughout this paper we have largely overlooked the issue of ethics, beyond that of invasive technology, however memory augmentation itself raises ethical issues. Firstly, the digital divide may limit the accessibility of memory augmentation technology. This may lead to those able to afford the memory aids to have advantages over others, for example in the workplace, where technology may aid individuals in job performance metrics. As such, the goal of this technology should be for it to be universally accessible so that anyone may utilise. This is particularly important in health contexts, where augmentation technology is being used to aid those with cognitive impairment.

Secondly, the act of augmenting one's memory will undoubtedly have knock-on effects, including new potentials for harm and deliberately malicious intervention. In the case of the current design space, augmenting an individual has a limited effect but as the design space opens up to group augmentation, this effect could have widespread consequences. One such consequence could be the use of memory augmentation technology to sway public opinion of national events. If group augmentation is achieved, entire populations could be manipulated into remembering national events differently, causing threats to freedom of thought. As such, the implementations of such technologies must be done with caution and Davies et al., (2015) suggested the need for memory security, to protect memory augmentation technologies from external threats, and protect a person's memories from tampering. This would be similar to the way we use anti-virus technologies to prevent our devices from being hacked. Preventing the hacking of a person's memory would allow these augmentation technologies to be used safely and with confidence that the memories are real. Given this, when considering memory augmentation, the ethical implementation of such tools is a priority as while it has implications for preventing cognitive impairments, it would be a powerful tool if used maliciously.

6. Conclusions

This paper aimed to present a novel design space for memory augmentation technologies. Our two-axis model identifies two key design factors in the creation of technological memory interventions: the display modality and the processing level. The mapping of previous research into this model enabled us to highlight areas in which little research has been conducted, thus allowing a clear insight into the spaces where more work is required.

The proposed directions for exploring this design space highlight the ways memory could be augmented in the future, and shows the benefit of mapping memory augmentation into such a design space. Reflections on future expansions that could be made to the design space also suggest the potential for memory to be considered in an inter-disciplinary fashion. However, as we highlight, there are ethical considerations to be made to ensure the safety of such technologies. Overall, the presented design space gives clear insight into current directions of memory augmentation research, and highlights the ways this field may continue to grow.

References

Bäckman, L., Nyberg, L., Lindenberger, U., Li, S. C., & Farde, L. (2006). The correlative triad among aging, dopamine, and cognition: current status and future prospects. Neuroscience & Biobehavioral Reviews, 30(6), 791-807.

Chen, Y. and G. J. Jones (2010). Augmenting human memory using personal lifelogs. In Proceedings of the 1st augmented human international conference, New York, NY, USA, pp. 24. ACM.

Davies, N., A. Friday, S. Clinch, C. Sas, M. Langheinrich, G. Ward, and A. Schmidt (2015). Security and privacy implications of pervasive memory augmentation. IEEE Pervasive Computing 14(1), 44–53.

Dingler, T., C. Giebler, U. Kunze, T. Wundefinedrtele, N. Henze, and A. Schmidt (2016). Memory displays: Investigating the effects of learning in the periphery. In Proceedings of the 5th ACM International Symposium on Pervasive Displays, PerDis '16, New York, NY, USA, pp. 118–123. Association for Computing Machinery.

Győrödi, R., C. Győrödi, G. Borha, M. Burtic, L. Pal, and J. Ferenczi (2015). Acquaintance reminder using google glass. In 2015 13th International Conference on Engineering of Modern Electric Systems (EMES), pp. 1–4.

Handra, C., O. A. Coman, L. Coman, T. Enache, S. Stoleru, A.-M. Sorescu, I. Ghita, and I. Fulga ˘ (2019). The connection between different neurotransmitters involved in cognitive processes. FARMACIA 67(2), 193–201.

Hampson, R. E., D. Song, B. S. Robinson, D. Fetterhoff, A. S. Dakos, B. M. Roeder, X. She, R. T. Wicks, M. R. Witcher, D. E. Couture, et al. (2018). Developing a hippocampal neural prosthetic to facilitate human memory encoding and recall. Journal of neural engineering 15(3), 036014.

Harvey, M., M. Langheinrich, and G. Ward (2016). Remembering through lifelogging: A survey of human memory augmentation. Pervasive and Mobile Computing 27, 14–26.

Hodges, S., E. Berry, and K. Wood (2011). Sensecam: A wearable camera that stimulates and rehabilitates autobiographical memory. Memory 19(7), 685– 696.

Hodges, S., L. Williams, E. Berry, S. Izadi, J. Srinivasan, A. Butler, G. Smyth, N. Kapur, and K. Wood (2006). Sensecam: A retrospective memory aid. In International Conference on Ubiquitous Computing, pp. 177–193. Springer.

Iwamura, M., K. Kunze, Y. Kato, Y. Utsumi, and K. Kise (2014). Haven't we met before?: a realistic memory assistance system to remind you of the person in front of you. In Proceedings of the 5th Augmented Human International Conference, New York, NY, USA, pp. 32. ACM.

Kuznetsov, S., A. K. Dey, and S. E. Hudson (2009). The effectiveness of haptic cues as an assistive technology for human memory. In International Conference on Pervasive Computing, pp. 168–175. Springer.

Le, H. V., S. Clinch, C. Sas, T. Dingler, N. Henze, and N. Davies (2016). Impact of video summary viewing on episodic memory recall: Design guidelines for video summarizations. In Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, New York, NY, USA, pp. 4793–4805. ACM.

Madan, C. R. (2014). Augmented memory: a survey of the approaches to remembering more. Frontiers in systems neuroscience 8, 30.

Mikusz, M., S. Clinch, P. Shaw, N. Davies, and P. Nurmi (2018). Using pervasive displays to aid student recall -reflections on a campus-wide trial. In Proceedings of the 7th ACM International Symposium on Pervasive Displays, PerDis '18, New York, NY, USA. Association for Computing Machinery.

Muller, J., F. Alt, D. Michelis, and A. Schmidt (2010). ¨ Requirements and design space for interactive public displays. In Proceedings of the 18th ACM international conference on Multimedia, New York, NY, USA, pp. 1285–1294. ACM.

Reggente, N., J. K. Essoe, H. Y. Baek, and J. Rissman (2020). The method of loci in virtual reality: explicit binding of objects to spatial contexts enhances subsequent memory recall. Journal of Cognitive Enhancement 4(1), 12–30.

Rhodes, B. J. (1997). The wearable remembrance agent: A system for augmented memory. Personal Technologies 1(4), 218–224.

Schmidt, A. (2017). Augmenting human intellect and amplifying perception and cognition. IEEE Pervasive Computing 16(1), 6–10.

Seim, C., J. Chandler, K. DesPortes, S. Dhingra, M. Park, and T. Starner (2014). Passive haptic learning of braille typing. In Proceedings of the 2014 ACM International Symposium on Wearable Computers, New York, NY, USA, pp. 111–118. ACM.

Seim, C., T. Estes, and T. Starner (2015). Towards passive haptic learning of piano songs. In Proceedings of the 2015 World Haptics Conference, pp. 445–450. IEEE. Solis, M. (2017). Committing to memory: Memory prosthetics show promise in helping those with neurodegenerative disorders. IEEE pulse 8(1), 33– 37.

Shu, Y., Y.-Z. Huang, S.-H. Chang, and M.-Y. Chen (2019). Do virtual reality head-mounted displays make a difference? A comparison of presence and self-efficacy between head-mounted displays and desktop computer-facilitated virtual environments. Virtual Reality 23(4), 437–446.

Solis, M. (2017). Committing to memory: Memory prosthetics show promise in helping those with neurodegenerative disorders. IEEE pulse 8(1), 33– 37.

Yang, F., J. Qian, J. Novotny, D. Badre, C. Jackson, and D. Laidlaw (2020). A virtual reality memory palace variant aids knowledge retrieval from scholarly articles. IEEE Transactions on Visualization and Computer Graphics.

15 Usability Recommendations for Delivering Clinical Guidelines on Mobile Devices

James Mitchell, Ed de Quincey, Charles Pantin and Naveed Mustfa

Keele University and UHNM NHS Trust
j.a.mitchell@keele.ac.uk, e.de.quincey@keele.ac.uk, c.pantin@keele.ac.uk, naveed.mustfa@uhnm.nhs.uk

Local point of care clinical guidelines exist in numerous formats and cover a variety of clinical information, normally created on a national and local level. They are generally available as basic web pages, PDFs or documents. Despite widespread availability and use, accessing clinical guidelines and information can be highly inefficient and restrictive. This reflective study investigates the evaluation of a clinical guidelines mobile application in the challenging area of co-design with clinicians. It aimed to answer if the selected methods of user centred design were suitable when working with limited access to users and what design recommendations can be elicited/changed by utilising user centred design (UCD) methods to gather feedback on features and functions. Specifically, this study utilised a mixed-method UCD approach and triangulation technique (Think-aloud and idea writing, screen recording and system usability scale). This culminated into the creation of 15 recommendations for developing clinical guidelines applications for mobile devices.

User centred design. Clinical guidelines. Mobile application design.

1. Introduction

Clinical guidelines are produced by numerous organisations, health trusts and hospitals worldwide (NICE, 2020; Pantin et al., 2006). They exist in numerous formats and cover a variety of clinical information (NICE, 2020; Pantin et al., 2006). Point of care clinical guidelines (patient treatment and medical process information designed for use at the point of care) are generally available as basic web pages, PDFs or documents (NICE, 2020; Pantin et al., 2006). Clinicians require agile access to these guidelines and an efficient delivery method (Free et al., 2013; Takeshita et al., 2002). Despite widespread availability and use however, accessing clinical guidelines and information can be highly inefficient and restrictive (Burton and Edwards, 2019; Littlejohns et al., 2003).

A previous study by the authors (Mitchell et al., 2020) investigated this issue by producing and evaluating a clinical guidelines app (based on the Bedside Clinical Guidelines (BCG)) utilising user-centred design methods (Abras et al., 2004; Norman, 1986; usability.gov, 2019). The main aim of the research was to identify and evaluate suitable methods for presenting clinical guidelines on a mobile phone interface, with a focus on efficiency and usability. The results from this study were then used to create a set of recommendations for developing mobile device apps to deliver clinical guidelines.

A total of thirteen (n=13) recommendations were developed:

Cross-Platform

List view with A to Z and Categories

Basic filter

Easy access menu (such as tabbed)

Minimise manual tasks (e.g. Manual calculations)

Minimise the requirement to use other systems (if possible), e.g. if a drug dosage calculation is required, this should be available to the clinician without the need to use another app or system. This may not be possible due to security, organisational governance or limitations of technology.

Decision algorithms to be displayed in-line with the guideline information

The original ‘flowchart’ decision algorithm is provided

Minimise the number of warnings/alerts to avoid ‘alert fatigue’

Acronym use is prevalent in medicine, but not all clinicians have knowledge of acronyms. Methods to address both experts and novices should be adopted.

Warnings should be more explicit and adopt better salience for the user

Guideline sentences should be reduced

Content Pages should utilise icons/images as well as headers

A second iteration of the application was developed implementing the 13 recommendations listed (see figure 1). This paper presents the evaluation and adaption of these recommendations through UCD, as well as reflecting on the UCD process and related methods, and their appropriateness for user groups with limited availability.

A mixed-methods UCD approach has been used based on the triangulation technique (Heale and Forbes, 2013; Noble and Heale, 2019) represented in *figure 2.* This enabled qualitative and quantitative data collection to inform design recommendations. The methods (Think-aloud and idea writing, screen recording and the system usability scale), rationale for selection and results are discussed in the following sections in more detail (2 - 5)

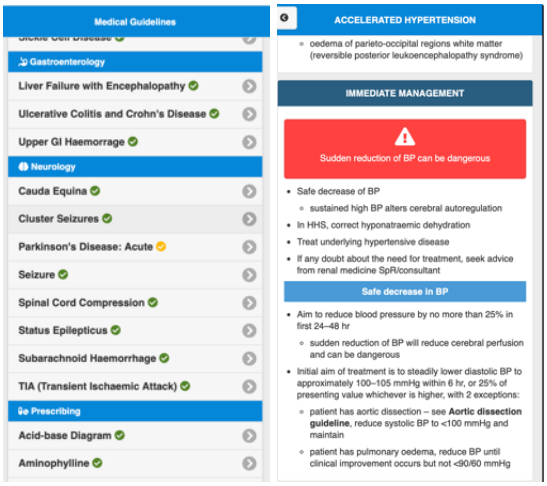


Figure 1: Example presentation of clinical information

1.1 Ethics

Ethical approval was granted by Keele University Research Governance in the Faculty of Natural Sciences (ERP2370) and from Research and Development at the University Hospitals of North Midlands NHS Trust. A letter of access was provided for the duration of the study.

2. Think-aloud

The think-aloud (Nielsen, 1992) technique was chosen to elicit feedback as it provided a method of understanding how users navigated the structure of the BCG app as well as their thoughts during the process of using it to complete basic clinical information retrieval scenarios. This method also allowed for the discovery of usability issues during information retrieval which may not have been identified during other methods of testing (i.e., focus groups).

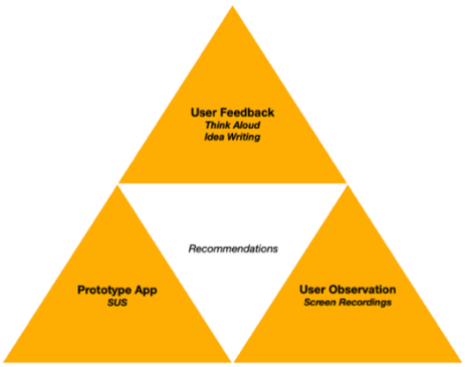


Figure 2: *Triangulation techniques used to evaluate the application*

2.1 Recruitment

Participants for this study were recruited via invitation emails which were sent via the Year Four Medical Lead to all fourth-year medical students at Keele University School of Medicine. In total, 38 students were recruited.

Participants

Participants were selected using a convenience sampling method and were required to have some medical knowledge. As access to clinicians was severely limited, it was decided that fourth year medical students would provide the adequate medical knowledge required for the basic information retrieval tasks and be accessible to the researcher. Demographics were not collected as this data would not provide relevant information in terms of design and implementation feedback. The purpose was to test the usability of the BCG app and therefore, selection based on demographics would not offer any further information required in terms of usability feedback.

2.2 Protocol

One to one sessions of fifteen minutes were arranged with all respondents who were offered certificates of participation. Participants were greeted and a brief overview of how think-aloud sessions are conducted was provided to allow users to understand the purpose and process of the session. Participants did not have access to the BCG app prior to the session and they were also not provided information on how the BCG app functions.

Participants were then asked to follow a process containing basic clinical scenarios. used to emulate clinical workflow. Research by Tu et al. (Tu et al., 2004) discusses the modelling of clinical guidelines for integration into clinical workflow and shows how a clinical workflow can be modelled using clinical scenarios. This is echoed in other studies such as Cossu et al. (Cossu et al., 2014), and UK based studies by Payne et al. (Payne et al., 2014) and Kwa (Kwa et al., 2015, 2014) where clinical scenarios were utilised during initial testing. The processes shown in these studies were used to inform the design of the clinical scenarios used for this study, described further in this section.

Screen and audio recordings were made of each think-aloud session using Apple QuickTime 10.4 and an iPhone X running iOS 13 tethered via USB.

2.3 Session overview

Clinical Scenarios

For this study, the clinical scenarios were developed with assistance of a Lead Respiratory Consultant at the Royal Stoke University Hospital. They were developed to ensure participants accessed specific guidelines and utilised guideline components such as text, warnings and decision algorithm tools.

For all sessions, the following process was followed:

Participants were provided with an overview of how to open the BCG app (the app was not opened at this stage).

Participants were provided with the three basic clinical scenarios. The clinical scenarios were based on three information retrieval tasks:

In the subsequent management of Unstable Angina, what is the recommended dose and method of administering Aspirin?

During fluid management in Acute Heart Failure, when should an echocardiogram be sought?

In the management flowchart of Hyperkalaemia, what is the recommended action where Plasma K+ 6.0-6.4 mmol/L and Acute ECG changes are present?

Scenario (a) was design to ask participants to retrieve basic text-based information. Scenario (b) was created to ask participants to retrieve text information contained in a warning, this enabled the analysis of how clinicians interact with the warnings contained in the BCG app. Scenario (c) was created to ask participants to retrieve information contained within a decision algorithm, again allowing analysis of how participants use the inline algorithm tools.

Audio and device screen recording was started (this is utilised for analysis discussed in section 6)

Participants were asked to access the BCG app and retrieve the required information (via scenarios) whilst discussing their actions and thoughts. Participants were asked to clarify comments during the session.

After completing the basic clinical scenarios, participants were given a brief demonstration of other features in the app e.g. Acronym support and Calculation tools.

Participants were asked to complete an SUS questionnaire (discussed in section 4).

During the think aloud, prompting questions were utilised when specific feedback was required but did not naturally occur during the session i.e. where a participant describes something as good they would be asked to elaborate and explain why it is “good”. These questions related to aspects such as design, layout, content and usability.

2.4 Think-aloud Analysis

Data Analysis

To identify themes from the think aloud session, audio recordings were transcribed verbatim and analysed by the primary researcher. User actions during screen recordings were analysed and coded.

Overall coded theme and category analysis

Table 1: *Main themes identified during the Thematic Analysis*

Theme	Description
MAIN MENU	*App content page*
GUIDELINE LAYOUT	*Design of the guidelines including Typeface/Colour, how the information is presented*
WARNINGS/ALERTS	*Presentation and content of warnings/alerts contained within the guidelines*
DECISION ALGORITHM	*Presentation and content of decision algorithms contained within the guidelines*
FILTER FUNCTION	*Presentation and content of filter functions contained within the guidelines and on the main menu*
FEATURES OR FUNCTIONS NOT PRESENT	*Suggestion or requirement of features and functions that are not currently available in the BCG app*

Table 2*: Number of comments related to each theme*

Main Menu	**9**
Guideline Layout	**94**
Warnings/Alerts (not specific to task)	**52**

Flowchart/Decision Algorithm tool	**74**
Text/font/colour	**7**
Filter Function	**16**

A total of 252 comments were coded over the 38 sessions analysed. In some cases, comments were considered neutral or irrelevant and therefore excluded from the final analysis. Examples include comments where participants would discuss unrelated information such as medical knowledge not relevant to the scenario or BCG app. On average participants made seven (n=7) comments that were coded/themed with a range of three to eighteen (3 – 18). Six themes were identified during the analysis shown in table 1. For each participant, an average of 3 themes were identified with a range of two to five (2 – 5).

Of the six identified in table 1, the themes most discussed (both positive and negative) were *GUIDELINE LAYOUT* (37.3% of comments) and *DECISION ALGORITHMS* (29.4% of comments). Details of the number of comments for each theme are provided in table 2. From overall comments, *GUIDELINE LAYOUT* and *DECISION ALGORITHM* represented a combined total of 66.6% (n=168/252). WARNINGS also represented a large portion of comments (20.6% of comments).

Comments for each theme (Table 1) were also categorised in terms of the categories which described comments overall. Similar studies have categorised in this way (Li et al., 2012) as it enables an understanding of the proportion of comments for each theme in terms of how they relate to aspects such as usability, clinical workflow etc. Therefore simplifying the identification of comments related to factors such as the content of the guideline versus the design on the guideline. Of the six themes identified (Table 1) four categories were created to code each comment (Table 3). Table 3 identifies these categories and provides a description of each. Of the four categories identified in Table 3, the categories most discussed were *USABILITY* (56% of comments) and *VISIBILTY* (23% of comments). From overall comments, usability and visibility represented a combined total of 79% of all comments (n=199/252).

***Table 3**: Categories of coding and description of each category*

Category	Description
Usability	*Comments which are considered to refer to how the app is used, how the information can be accessed and how the users 'feel' in terms of its use. (e.g. "I like how this looks")*
Visibility	*Comments which refer to the visibility, colour, salience, layout etc. (e.g. "I didn't notice it because it didn't stand out")*
Clinical Workflow	*Comments specifically refer to use of the app and its functions in wards/hospitals (e.g. "this would be really useful when treating patients as it can get busy on the wards")*
BCG Content	*Comments which specifically refer to the content itself – including text, knowledge and specific medical information/methods (e.g. "I would have expected this section to be above investigations")*

Participant comment analysis

Sessions were also analysed for consistent patterns in how participants utilised features of the BCG app. The following sections discuss the results of each particular theme presented in table 2 and provides examples of comments made by participants in relation to each.

2.5 Main Menu

All participants navigated the main menu without the need for prompting or further instruction. All participants were able to access the specific guidelines. In some cases, they utilised the filter function (n=30) – this is analysed further in this section. Some participants made specific positive comments in relation to the use of icons and headers for the sections provided. This can be summarised by the following participant quote:

"that's nice that you have this at the beginning so that you could flick through and see just an overview of all the things that you have on it"

Of the nine (n=9) comments made by participants in reference to the main menu, seven (n=7) were considered positive and two (n=2) were considered negative. An example of a positive comment referenced the use of categories:

“You've got headings which I like”

The majority of positive comments reference the layout and ease of use in terms of finding what they need. An example of a negative comment mentioned the following:

“maybe it'd be nicer if it was just the big blue header and then you can open and close”

The negative comments (n=2) in reference to the main menu all have similar themes in terms of presenting the content in an accordion type (open and close) view, as the above comment suggests.

2.6 Guideline Layout

The majority of respondents made general comments regarding the layout of the guidelines. A total of 94 comments were coded in reference to the guideline layout, a large proportion of all comments that were coded (37.3% of all comments). Of that total, 69 were considered positive and 25 were considered negative. As the following example highlights, most positive comments referenced the ease of finding information or the clarity of the layout:

“I think just how it's laid out signs and symptoms and then investigations and then differential diagnosis. I feel like it's laid out in a good order and there's not too much text as well. Cause I find that when I'm using NICE and stuff like that, there's so much text.”

In terms of negative comments, the majority of participants suggested a more collapsible layout may be beneficial. One user did specifically mention that in one of the guidelines, scrolling was undesirable. The participant stated:

“I think it's a bit long to like scroll down on set. I think just separating it a bit and bit might be a bit useful”.

Other comments suggested that there should be an overview of all the content (e.g. a content section or titles at the top of each guideline) to facilitate user understanding of the guideline layout:

“Maybe like at the top there could be like a mini, like contents where you could click on, for example, subsequent management and anything”

Feedback also suggested that the order of the content would be more beneficial if different from its current layout, for example:

“my only sort of thought with that is having the differentials above investigations. So as you read an investigations, you already know what really not helped.”

2.7 Warnings/Alerts

A large proportion of respondents specifically mentioned the layout of warnings or gave specific feedback regarding the information contained in the BCG app warnings (n=52/252, 20.63% of all comments). Of all the comments coded to particular themes, warnings/alerts received the majority of negative comments (45% of all negative comments). This was due to participants expecting the use of acronyms or shortened versions such as ‘ECG’ or ‘ECHO’. This was evident through comments such as:

“So you would expect acronyms to be in there too”

“It was more because I didn't see that it was anything to do with an echo”

Some participants suggested the information should be repeated in context within the guidelines. Summarised by some participant in the following comments:

“So I was expecting it to be in the standard text. Um, I normally would have looked at that… Perhaps a repeat of that. So, repeating the warning, in the information.”

“I was actually looking for a bullet that said echocardiogram. Okay. Um, so perhaps you could include it as both. It's like in the red and as a bullet point.”

Some participants also suggested warnings that contained too much text were harder to assimilate when scrolling through the BCG guidelines. In reference to the amount of text contained in a warning, one participant mentioned:

“I like things that are bullet pointed and then inset bullet point, and then the detailing.”

In some cases, negative comments were associated with users not finding the information contained within the warning. Whilst some suggested that the information should be repeated, other users specifically mentioned that

they felt the medical procedure would not necessarily be presented in a warning box, as the comment below suggests:

“I think I just assumed. That, that wouldn’t be. I didn’t read that. I don’t know why, although it looks like it’s designed to be more important. I guess I assumed that an echo wouldn’t be that important”

However, other participants suggested that the information in refence to an echocardiogram would not necessarily be expected to be in a section with fluid management.

“so maybe it’s just me missing it. And then if we hadn’t, since it’s about an echocardiogram, …put that in the fluid management”

These comments echo other participant comments referencing the repeat of information in the main text. It also highlights individual user behaviour and how participants assimilate the information contained in the guideline. One participant specifically mentioned their workflow may have contributed to them missing the information contained within the warning:

“I’m so used to just looking straight at the text rather than in boxes. Um, and usually I go back to boxes to see if things are important. Yeah. Um, but I’m usually, yeah, that’s hard to get straight to text, so that’s why I missed it”

The majority of positive comments referred to the salience of the warning, in particular the use of colour. Participant specifically mentioned the warning salience during the sessions:

“I definitely saw like the red warning thing, so I guess that is quite, it shows that it’s important. I guess if it’s immediate, that means that you probably want to put at the top, which you guys did and. This pops out because you don’t see this kind of thing on the other, on the other one that I saw”

“That’s quite nice to have like a big warning to make sure that you do what you need to do”

“Cause it’s an, a red box with a warning and like, I think anyone would automatically look and make sure like, what’s that warning about”

2.8 Decision Algorithm

All sessions were analysed in terms of how the participants interacted with the decision algorithm tool. The users had two options in terms of how to access the flowchart information they required to complete the scenario, a programmatic version and an image of the original flowchart/algorithm. However, they were not made aware of this in order to assess which method they would instinctively access. t is worth note that the decision algorithm tool is more salient in terms of design than the button to access the original version (figure 3). However, participants had access to both programmatic and original version of the decision algorithm within the same area of the guideline (Figure 3). Table 4 shows the number of participants utilising each version.

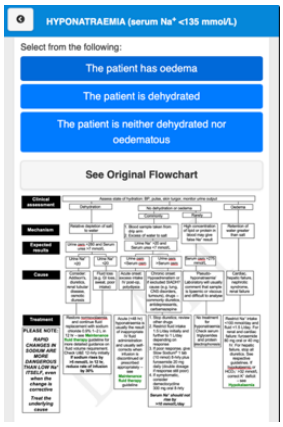


Figure 3: *Programmatic and original flowchart (inline)*

Table 4*: Number of participants utilising each version of the decision algorithm*

Utilised programmatic version	Utilised original version
37	*1*

Of all participants (n=38), all but one (n=37) accessed the programmatic version of the decision algorithm. Comments made by participants on the design and use of the inline decision algorithms were overwhelmingly positive. Of the seventy-four (n=74) comments made by participants in reference to the tool, sixty-seven (n=67) were positive and seven (n=7) were classified as negative. Specifically, one participant mentioned when comparing the two decision algorithms:

“so this is just a different way of presenting that digital flow chart. I think I liked the other (ref to new method) because this is too complicated (ref to original). And I think when needed quickly on the ward and you want to see something that probably not the best way”.

Another participant also reflected on the design, specifically stating:

“it helps you follow in your head. I find that flowcharts can be a bit much sometimes following it. Whereas this specifically just gives you the answer you need rather than everything on stuff. So, it makes it a bit easier to follow and easy to get the information you need”.

One participant also discussed the decision algorithm. Directly referencing the amount of information presented and reflecting on the need for specific information. This was also reflected in their comment, were they stated:

“sometimes when it's like branching and you having to look everywhere to find exactly what you need, it’s to the point”

Interestingly, there appeared to be a separate viewpoint on the use of information for learning as opposed to clinical use. This was highlighted specifically by one participant in reference to the presentation of the original decision algorithm (flowchart), stating

“I guess the original flow chart be good for learning”.

2.9 Filter Function

Participant screen recordings were analysed to determine if any utilised the filter function (Figure 4), both on the main menu and within the guideline (Table 5).

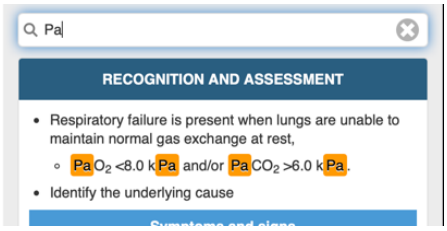


Figure 4: *Filter function available in each guideline and via the main menu*

Table 5: *Number of times participants accessed the filter function*

Utilised filter in the Main Menu	Utilised filter in the guideline
18	*12*

As the table highlights, participants accessed the filter function during the session with no prompting or instruction. The main menu filter was accessed by eighteen of the participants (n=18/38), and the guideline filter function was accessed by twelve participants (n=12/38). Participants also specifically mentioned using the filter function during use, describing it as a “quicker” or “faster” method of retrieving information. Overall, 16 comments were coded in reference to the filter functionality of the BCG app. Of these comments, 14 were considered positive, with general positive comments as mentioned. In terms of negative comments, 2 were identified by separate users. One user specifically mentions in terms of clinical workflow the following:

“If you didn’t know, you could type in potentially the symptoms or to go into cardiac”

Another participant also suggested that the filter function may be more useful if it allows the user to:

“move to the next part”

This suggests that the user is navigated to each highlight of the filter in a similar method that some PDF/Browser word filters function.

Features or functions not present

Participants mentioned aspects of clinical information that may be useful within the bedside clinical guidelines. In particular, drug calculation tools or information on specific treatments. As the scope of this study is to investigate the delivery of existing guidelines, it is beyond the scope of this study to investigate information on guidelines that do not currently exist. However, it is interesting to highlight that the information needs of participants does differ especially in terms of clinical expertise and interest.

Positive/Negative analysis

As well as identifying themes and categories, sessions were also analysed in terms of whether comments were positive or negative. This allows for an overall analysis of participants attitude towards the BCG app and enables the identification of specific features/themes that participants described in negative or positive terms. The following describes how comments were coded as positive or negative:

A ***positive*** reaction or general comment (e.g. “this is really great” or describing the use of a feature in a positive manner (e.g. “This would be really useful when…”)

A ***negative*** reaction or general comment (e.g. “I don’t like this..”) or any criticism, suggestion of alternative methods or ways in which the user prefers (e.g. “this is good but I would like it if it did…”

Each coded comment considered negative or positive was analysed by theme. Overall, of the 252 comments coded, a total of 182 were coded positive and 70 were coded negative. The majority of coded comments considered positive (n=182/252 or ~82%) focussed on *GUIDELINE LAYOUT* and the *DECISION ALGORITHMS*, both of which, as mentioned, received the most comments overall. Interestingly, the majority of negative coded comments also focussed on *GUIDELINE LAYOUT* (36% of all negative comments). However, this was most likely due to the high number of comments received overall. *WARNINGS/ALERTS* (46% of all negative comments) received a greater proportion of negative comments relative to overall comments. Of the 52 comments referencing warnings/alerts, 32 were coded negative and 20 coded positive.

Errors/Issues

The think-aloud sessions were also analysed for any occasions were participants encountered issues or errors. Table 6 describes the three areas created to describe the issues found.

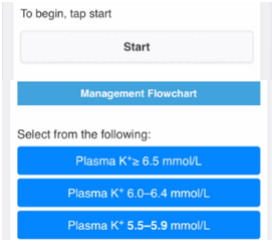
Table 6: *Issue types and descriptions*

Issue/error Type	Description
INFORMATION RETRIEVAL ISSUE	*Unable to retrieve the necessary information to complete the scenario or where the user selects the wrong information.*
USABILITY ISSUE	*Interacts with the app in a way they perceive negative due to its design or functionality*
OTHER	*Discovers a bug or app issue not related to information retrieval or usability*

Table 7: *Number of occurrences of each issue during think-aloud sessions*

Issue/error Type	Number of occurrences
Information retrieval	*18*
Usability issue	*9*
Other	*1*

A total of 26 issues/errors occurred over the 38 sessions, 68% of sessions. The 26 issues occurred in 21 sessions of the 38 with a range of 0 – 2 issues per session. Table 7 provides an overview of the types of issues and the number of occurrences for each type. Of the 18 occurrences of issues related to information retrieval, 9 occurrences were related to participants locating information incorrectly. Despite the scenario specifically asking users the 'dose of aspirin in subsequent management', 9 participants provided the initial dose contained in the management section. When prompted to locate the information in 'subsequent management' some users did state that an overview of the sections available in the guideline may be useful. This was highlighted in the comments contained in the Guideline layout section. The 9 other occurrences of information retrial errors all related to user not able to locate information contained in the warning box provided in the Acute Heart Failure guideline. This was due to the expectation of acronyms/short versions and the expectation of text contained in the warning would be repeated or available in the main guideline text, as mentioned in Warning/Alerts section. Of the 9 occurrences of usability issues, 2 occurrences related to locating the decision algorithm tool. During the first 3 sessions, the decision algorithm had to be activated by clicking the start button. After the initial usability issues this was changed to be inline without requiring activation, see figure 5. No further occurrences of this issue occurred in the remaining 36 sessions. This highlighted a clear usability issue that was resolved.



A(TOP): DECISION ALGORITHM TOOL THAT REQUIRES ACTIVATION *B(BOTTOM): INLINE DECISION ALGORITHM TOOL REQUIRES NO ACTIVATION*

Figure 5: *Display changes for decision algorithm tools*

The most prevalent usability issue was related to users mistaking a header for a button (Figure 5 shows the header and button). 5 participants (13.5% of participants) attempted to click the header for the tool before realising the tool was already present in the guideline. This represents 56% of the usability issues identified. Upon analysing the screen recording, all 5 participants failed to scroll down far enough to visibly see the tool, therefore assumed they could activate it using the header. This could also have been caused by the gap between the header and tool (see Example B), which does not conform with best practice (Wagemans et al., 2012). Most users acknowledged the error and, on some occasions, mentioned that this would not occur after they have become more familiar with how the BCG app works. Other usability issues included an occasion where one participant could not initially locate the 'Acute heart failure' guideline in the main menu, caused by the participant looking for heart failure and did not expect 'acute' to precede the title. Another issue identified was related to the filter function within the guideline. One participant attempted to move to the next guideline by searching for it in the filter tool, this was corrected by the participant without any interjection. A further issue was identified during the 18th think-aloud session. A bug was identified where the warnings did not display when using the filter function, this was categorised as an 'other' issue as it was not specifically related to usability or information retrieval. This was fixed before further sessions were conducted. Analysis of previous 17 sessions did not identify any other occurrences of this issue and the issue did not contribute to any negative comments or other issues identified during the previous sessions.

3. SUS

The System Usability Scale (SUS) (Brookes, 1996) was used to establish the usability level of the application from the clinicians' viewpoint. It also provided a baseline to measure future changes in the design and how they impact the usability. All participants (n=38) were asked to complete the SUS questionnaire post think-aloud session. Results were then analysed and compared to results of a previous study which had investigated the usability of a previous version of the app (Mitchell et al., 2020). This allowed for analysis of changes made to the application based on recommendations derived from previous UCD studies. These are discussed in previously published work (Mitchell et al., 2020).

3.1 SUS results

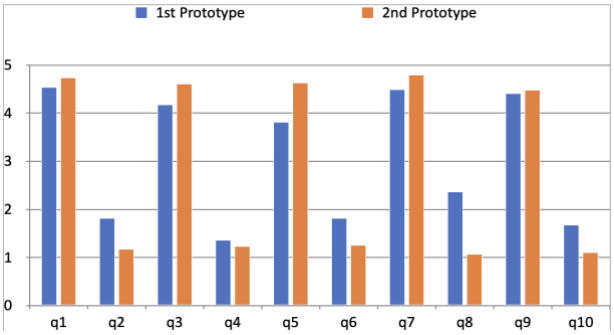


Figure 6: *A comparison of SUS results for the 1st and 2nd application iteration*

The app was shown to maintain a high usability score, with an overall score of 93.6 out of 100 (calculated utilising the methods described in (Brookes, 1996)). This result was higher than SUS scores discussed in previously published work (Mitchell et al., 2020), 81 out of 100. The consistent results in all sessions highlight a general consensus amongst participants that they highly rate the usability of the BCG app. This was also reflected in the positive comments/feedback discussed in the think aloud sessions. A comparison of the first and second application iteration is provided in figure 6.

4. Idea Writing

To further evaluate the app, an 'idea writing' session (Austin, 1994; VanGundy, 1984) was conducted. As discussed in a previous publication (Mitchell et al., 2020), the concept of using the idea writing methodology was due to the necessity to elicit information in a limited time. As access to clinicians is limited, idea writing allowed for a focus group to feedback based on a 'closed' method.

4.1 Idea Writing method

During this session, clinicians interacted with the application and were asked to feedback on each aspect of the design, which was presented as a 'concept'. Although this limited open discussion (by design), it allowed for more specific feedback regarding the design of the BCG app.

Participants

This session was conducted at the Wythenshawe Hospital, part of the Manchester University NHS foundation Trust. The session was conducted with four (n=4) participants. Participants were selected using the convenience sampling method.

4.2 Feedback

Feedback provided during the idea writing session was largely positive. Specifically, participants used words such as "very useful" and "good" to positively describe the app, an example includes:

"simplify the content as too wordy to be used in emergency although info all good - my suggestion is to use flowcharts as much as possible as first thing you see then have the fuller content below or linked to separate page"

Although this is related the authoring of the guidelines, this does specifically mention the need for succinct information delivery in an emergency, specifically delivery utilising the decision algorithm. Another comment refers to guideline titles:

"simplify and lose acute from the section titles as it makes it harder to search for subjects"

Although this has only mentioned by a single participant during the focus groups and think aloud sessions, an interesting point was raised regarding succinct information and how it is displayed in the content pages. It also has similarities to a usability issue which occurred during the think aloud sessions where a user was unable to locate the Acute Heart Failure guidelines because it was superseded by the word acute. Another comment also referenced the layout of guidelines, specific to warnings:

"warnings at top of pages"

This was in contrast to feedback received during other focus groups and think aloud sessions. However, it does highlight that individual preference may be a key factor in delivering clinical information and this requires further investigation.

5. Overview of results

Table 8 provides an overview of the main findings presented in sections 2-5. Each finding is presented with the method utilised and how it has affected the recommendations presented in the introduction of this paper.

Table 8: *Overview of method findings and outcomes*

#	Finding	Method	Outcome
1	*Inline decision tools caused less errors*	*Think aloud and video analysis*	***Adapting*** *existing recommendations to include inline activation*
2	*Warning design should be more explicit and salient*	*Think aloud and Idea Writing*	***Adapting*** *existing recommendations to include explicit, salient warnings*
3	*Warning text be repeated to avoid missing critical information*	*Think aloud*	*Adding a new recommendations based on repeat warning text.*
4	*Easier to find guidelines if unnecessary wording is removed*	*Think aloud and idea writing*	*Adding a new recommendations based on removing wording from titles in content pages*

6. DISCUSSION

This study utilised a mixed-method triangulation approach to inform the improvement of a mobile application for delivering bedside clinical guidelines. The use of the think-aloud technique with clinical scenarios and the 'idea writing' focus group, as well as the SUS methodology produced data which has informed on the impact of implementing recommendations and identified clear usability issues (i.e. decision algorithm activation). Despite the overlap in the findings of these methods, unique insights were elicited from participants. These methods also enabled evaluation of a clinical application where access to relevant users (clinicians) is extremely limited and restricted in terms of time. They also offer a unique insight into the use of these techniques as no studies that have combined these techniques to inform the delivery of bedside clinical guidelines could be found. The evaluation has provided a number of specific and general findings relevant to the development of the BCG app. In terms of layout, some participants referred to the order of content and specified alternative ordering. This is indicative of how preferences differ between individuals. Similar findings were also discussed in a previous publication, where personal preference has contributed to a large amount of variation in the apps clinicians utilise. This is further impacted by the requirements of the delivered information in terms of educational use as opposed to clinical use. Participants conveyed the need for a more in-depth delivery of information when learning. This is highlighted in Karen Davies's review on the information-seeking behaviour of doctors, which states two main behaviours when clinicians are seeking information, one seeking facts and another seeking literature (Davies, 2007). This also reflects the findings of the observational study (Mitchell et al., 2020), which found that Junior clinicians appear to use technology to establish knowledge which requires more information. Senior clinicians utilise technology for knowledge affirmation. The use of acronyms also suggests there are differences in the needs of individual clinicians from a knowledge perspective. Interestingly, the topic of warnings generated much discussion in terms of the information they contain. Specifically, the use of acronyms was expected by participants which is in direct contrast to feedback received during previous sessions (Mitchell et al., 2020). This may be due to the subject matter utilised within the warning. The scenarios utilised echocardiograms, a subject the participants were familiar with. It remains to be seen if other more complex subjects and less used acronyms would highlight knowledge gaps. However, previous findings highlighted the need to provide both acronyms and explications.

7. Conclusions

This study aimed to answer if the selected methods of user centred design were suitable when working with limited access to clinicians. Based on the feedback received and the adaption of recommendations, these methods have worked efficiently on providing feedback and evaluation for an app. The study also aimed to evaluate what design recommendations can be elicited/changed by utilising user centred design methodologies. The evaluation of the thirteen recommendations during this paper suggests that at least two of the original recommendations discussed in the introduction of this paper need to be adapted (as presented in Table 8):

Decision algorithms and Calculation tools should be displayed in-line with the guideline information, clearly outlined to distinguish from the main content, and ready to be used (i.e., does not require activation); Warnings should be succinct, explicit and adopt a salient design to ensure visibility. The findings also suggest the addition of two new recommendations , they are as follows: Text contained in alerts or warnings should also be available within the text it refers to; Remove unnecessary wording in titles e.g. Instead of 'Acute Heart Failure' use 'Heart Failure'. The adaption of previous recommendations and the addition of new recommendations has culminated in the creation of 15 recommendations for developing clinical information delivery applications for mobile devices. The following provides an overview of the final set of recommendations.

- Be cross platform
- Provide multiple methods of accessing content in list views (i.e., A to Z and Categories)
- Minimise unnecessary wording in titles (i.e., 'Acute heart failure' should be presented as 'heart failure'
- Have a menu that can be easily accessed, preferably using a tabbed menu design
- Utilise icons/images as well as headers
- Provide a basic filter function to filter content in both menu and information sections
- Minimise manual tasks (i.e., Drug dose calculations)
- Provide as many tools and resources as possible to minimise the requirement to use other systems
- Provide clear decision algorithms and calculation tools in line with content, and ready to use (i.e., does not require activation)
- Provide original content for any tools or decision algorithms (i.e., An original flow chart)
- Utilise acronyms, but also provide a method of understanding acronyms where possible
- Minimise the number of warnings/alerts to avoid 'alert fatigue'
- Display warnings/alerts in line with content, ensuring they are salient in design and succinct and explicit in content
- Repeat warning content within the main information
- Reduce the use of long sentences and provide information as succinctly as possible

Aside from the recommendations elicited from feedback and evaluation, it is also clear that further investigation into personalised delivery is required. Although a limited number of participants specifically mentioned layout, the feedback during the evaluation of both BCG apps highlights the eclectic nature of information delivery that satisfies user preference.

7.1 Limitations

It is worth note that the higher SUS score could be attributed to the fact that clinical students were utilised, a group that are familiar with mobile devices and clinical application use (Mitchell et al., 2020; Payne et al., 2012; Prescott et al., 2017). However, students participate in clinical practice through their university course, a requirement for all student clinicians in their final years of study. It is suspected that although this may have some effect on the results, it would not have a considerable impact as student and junior clinicians were utilised in the earlier SUS sessions and focus groups.

8. References

Abras, C., Maloney-Krichmar, D., Preece, J., 2004. User-Centered Design.

Austin, M., 1994. Needs Assessment By Focus Groups. American Society for Training and Development.

Brookes, J., 1996. SUS: a "quick and dirty' usability Scale, Usability Evaluation In Industry. https://doi.org/10.1201/9781498710411-35

Burton, Z., Edwards, H., 2019. A little less conversation, a little more high impact action. Futur Healthc J 6, 201–201. https://doi.org/10.7861/futurehosp.6-1-s201

Cossu, F., Marrella, A., Mecella, M., Russo, A., Kimani, S., Bertazzoni, G., Colabianchi, A., Corona, A., Luise, A.D., Grasso, F., Suppa, M., 2014. Supporting Doctors through Mobile Multimodal Interaction and Process-Aware Execution of Clinical Guidelines, in: 2014 IEEE 7th International Conference on Service-Oriented Computing and Applications (SOCA). p. 183 190. https://doi.org/10.1109/soca.2014.44

Davies, K., 2007. The information-seeking behaviour of doctors: a review of the evidence. Health Information & Libraries Journal 24, 78–94. https://doi.org/10.1111/j.1471-1842.2007.00713.x

Free, C., Phillips, G., Watson, L., Galli, L., Felix, L., Edwards, P., Patel, V., Haines, A., 2013. The Effectiveness of Mobile-Health Technologies to Improve Health Care Service Delivery Processes: A Systematic Review and Meta-Analysis. Plos Med 10, e1001363. https://doi.org/10.1371/journal.pmed.1001363

Heale, R., Forbes, D., 2013. Understanding triangulation in research. Évid Based Nurs 16, 98. https://doi.org/10.1136/eb-2013-101494

Kwa, A., Carter, M., Page, D., Wilson, T., Brown, M., Baxendale, B., 2015. NOTTINGHAM UNIVERSITY HOSPITAL GUIDELINES APP–IMPROVING ACCESSIBLITY TO 650 HOSPITAL CLINICAL GUIDELINES. Proceedings of the International Conference on Ergonomics & Human Factors 2015 220.

Kwa, A., Wilson, T., Carter, M., Page, D., Brown, M., Bennett, O., Miles, G., Baxendale, B., 2014. 0162 Developing A Mobile App To Improve Access And Application Of Key Hospital Guidelines. Bmj Simul Technology Enhanc Learn 1, A25.2-A25. https://doi.org/10.1136/bmjstel-2014-000002.60

Li, A.C., Kannry, J.L., Kushniruk, A., Chrimes, D., McGinn, T.G., Edonyabo, D., Mann, D.M., 2012. Integrating usability testing and think-aloud protocol analysis with “near-live” clinical simulations in evaluating clinical decision support. Int J Med Inform 81, 761–772. https://doi.org/10.1016/j.ijmedinf.2012.02.009

Littlejohns, P., Wyatt, J.C., Garvican, L., 2003. Evaluating computerised health information systems: hard lessons still to be learnt. Bmj 326, 860 863. https://doi.org/10.1136/bmj.326.7394.860

Mitchell, J., Quincey, E. de, Pantin, C., Mustfa, N., 2020. The Development of a Point of Care Clinical Guidelines Mobile Application Following a User-Centred Design Approach. Lecture Notes in Computer Science 294–313. https://doi.org/10.1007/978-3-030-49757-6_21

NICE, 2020. NICE [WWW Document]. NICE Clinical Guidelines. URL : https://www.nice.org.uk/guidance (accessed 9.2.20).

Nielsen, 1992. Evaluating the thinking aloud technique for use by computer scientists. Advances in Human-Computer Interaction Vol. 3 69–82.

Noble, H., Heale, R., 2019. Triangulation in research, with examples. Évid Based Nurs 22, 67. https://doi.org/10.1136/ebnurs-2019-103145

Norman, D.A., 1986. User Centered System Design. https://doi.org/10.1201/b15703

Pantin, C., Mucklow, J., Rogers, D., Cross, M., Wall, J., Partnership, T.B.C.G., 2006. Bedside clinical guidelines: the missing link. Clin Med 6, 98–104. https://doi.org/10.7861/clinmedicine.6-1-98

Payne, K.F., Weeks, L., Dunning, P., 2014. A mixed methods pilot study to investigate the impact of a hospital-specific iPhone application (iTreat) within a British junior doctor cohort. Health Inform J 20, 59 73. https://doi.org/10.1177/1460458213478812

Payne, K.F.B., Wharrad, H., Watts, K., 2012. Smartphone and medical related App use among medical students and junior doctors in the United Kingdom (UK): a regional survey. Bmc Med Inform Decis 12, 121. https://doi.org/10.1186/1472-6947-12-121

Prescott, O., Millar, E., Nimmo, G., Wales, A., Edgar, S., 2017. 21st century medical education: critical decision-making guidance through smartphone/tablet applications—the Lothian pilot. Bmj Simul Technology Enhanc Learn 3, 60–64. https://doi.org/10.1136/bmjstel-2016-000157

Takeshita, H., Davis, D., Straus, S.E., 2002. Clinical Evidence at the Point of Care in Acute Medicine: A Handheld Usability Case Study. Proc Hum Factors Ergonomics Soc Annu Meet 46, 1409–1413. https://doi.org/10.1177/154193120204601601

usability.gov, 2019. User Centred Design, US Government [WWW Document]. URL https://www.usability.gov/what-and-why/user-centered-design.html

VanGundy, A.B., 1984. BRAIN WRITING FOR NEW PRODUCT IDEAS: AN ALTERNATIVE TO BRAINSTORMING. J Consum Mark 1, 67–74. https://doi.org/10.1108/eb008097

Wagemans, J., Elder, J.H., Kubovy, M., Palmer, S.E., Peterson, M.A., Singh, M., Heydt, R. von der, 2012. A Century of Gestalt Psychology in Visual Perception: I. Perceptual Grouping and Figure–Ground Organization. Psychol Bull 138, 1172–1217. https://doi.org/10.1037/a0029333

How much Sample Rate is actually needed? Arm Tracking in Virtual Reality

Hildegardo Noronha and Pedro Campos

ITI/LARSyS - University of Madeira

Polo Científico e Tecnológico da Madeira

Caminho da Penteada, piso -2 9020-105 Funchal

Portugal

hildnoronha@gmail.com, pedro.campos.pt@gmail.com

There are plenty of studies dealing with the delays and other relations between head movements and visual response on Virtual Reality setups using head mounted displays. Most of those studies also present some consequences of deviating from those values. Yet, the rest of the human body remains relatively unmapped. In this paper, we present the data found during our research about vision-arm coordination. This data can be used to help build better and more efficient human-computer interfaces, especially those that rely on a virtual avatar with a body and have resource restriction like battery or bandwidth. We tested body tracking Sample Rates ranging from 15 Hz up to 120 Hz (corresponding to total latencies ranging from 37 ms to 95.4 ms) and found out no significant user performance differences. We did, however, find that a small percentage of users are, indeed, capable of noticing the changes in Sample Rate. Based on the found results, we advise that, if one is trying to save battery, bandwidth or processor cycles, a low body tracking Sample Rate could be used with no negative effects on user performance.

Virtual Reality. Arm-Tracking. Sample-Rate.

1. Introduction

Human body-tracking or motion capture is a field that has boomed with the advent of 3D movies and that is now expanding into gaming. The techniques used for capturing the human body movements and position have been evolving and new techniques keep emerging. Presently, the most commonly used techniques are optical/camera based and inertial based. There are other techniques such as mechanical, magnetic, acoustic and radio reflection tracking.

Even though there are tracking systems and techniques that do not, inherently, have those characteristics (for instance, optical based tracking), those systems can still benefit from this study. The benefits come from the fact that processing power, bandwidth and costs are limited resources and knowing the lower tracking limits of each body part on different situations allows the developer to fine-tune the tracking characteristics as well as the priorities allocated to each body part. This allows the building of cheaper products while maintaining the full quality of the tracking.

2. State-of-the-Art

The following paragraphs enable us to illustrate the importance of body tracking in the medical field. It also illustrates the importance of a better understanding of the intrinsic values of body tracking for each body part.

(Cloete et al., 2008) compared the kinematic reliability of both inertial and optical motion capture applied to clinical gait analysis. Both systems that were compared were professional, commercially available solutions and were probed at 100 Hz. They found out that the inertial motion capture had more errors than expected but found out that the problem was due to a lycra suit used and that those errors would be solved, based on a paper by (Dejnabadi et al., 2005), if the sensors were secured in place. On the optical side, they encountered issues with markers outside the camera view, shadows, and bad marker reflections. They conclude that the reliability is comparable for lower walking speeds. They also argue that the inertial system is a lot faster to set up than the optical one. This happens because the inertial system is a lycra suit while the optical system is an 8-camera system. The same would not be true if they would compare a strap based inertial system with a 1 or 2 camera system. (Cloete et al., 2010) studied, a couple of years later, the same systems from a repeatability point-of-view and concluded that inertial systems give enough repeatability to be used on clinical gait analysis. They noted, though, that those systems may perform less optimal on real patients due to body characteristics affecting the sensor placement.

On the studies of the previous paragraph, they used optical tracking systems with 8 cameras. One can argue that it was to achieve a higher degree of accuracy or it may have been due to a lack of better and cheaper solutions at the time (2008 and 2010). In 2012 (Wei et al., 2012) proposed a motion capture method using a

single depth camera and compared it with Microsoft Kinect (2012 version; the original version came out in 2010), which is also a single depth camera, and concluded that their method was more accurate.

(Lorincz et al., 2009) ran into a problem that could have been mitigated by the results on this paper. They run a group of sensors on patients, some of them were capturing movement through inertial sensors. Those sensors were fed by a battery and must run up to 18 hours per day. They also ran into issues with data storage and network bandwidth. The high volume of data (reported as 1200 byte/sec/node) as well as a big battery drain might come from the fact that the sensors are set at 100 Hz all the time when they could have been fine tuned to lower values while achieving a similar quality of results. The sample frequency could have been further lowered considering that the movements are not to be interpreted by the user in real time. The authors did throttle down the sensors when battery life was low rising the expected time of battery up to 32h, adding to the importance of more efficient sensor tuning.

(Witchel et al., 2012) made a comparison of four technologies applied to micro-movements. From the technologies they used, the most relevant for this paper, are the 8-camera optical tracking (a Vicon) and an accelerometer mounted on the head. They found out a good correlation between both systems, except for the yaw on the accelerometer. This happens because accelerometers cannot, directly and accurately, measure yaw movements. An important find is that, even without a gyroscope, they were able to match the rotation on the head to an expensive 8 camera tracking system, proving the quality and accuracy of a (striped down, accelerometer only) inertial tracking system.

(Aylward et al., 2007) gives us an example of implementation of inertial sensors on other fields. In this case, the paper focuses on dancing, but it is also tested on baseball illustrating the potential for the tracking of high speed, high acceleration movements while maintaining accuracy.

3. Materials and Methods

3.1. Subjects

The experiment was conducted using 41 volunteers (19 to 37 years old). All of them had good knowledge and contact with, at least, one of the following: computers, entertainment systems and gaming systems.

3.2. Materials

3.2.1. Hardware

We used the MPU-9150, a 9-degrees of freedom IMU. The accelerometer has a selectable range from 2 to 16 g. The gyroscope has a selectable range from 250 to 2000 º/. The magnetometer has a fixed range of 1200 µT.

3.2.2. Questionnaire

The users were requested to fill up a questionnaire with following questions graded from -5 to +5:

I felt that this iteration was -5 – Slower; 0 – Equal; 5 – Faster.

I felt that this iteration was -5 – Less Responsive; 0 – Equal; 5 – More Responsive.

I felt that this iteration was -5 – Harder; 0 – Equal; 5 – Easier.

3.2.3. Inertial Tracking

The system uses a series of inertial sensors positioned on several bones (arm, forearm, and hand – see Image 1). After some research and trial-and-error, we ended up using (Madgwick et al., 2011)'s algorithm to fuse the sensors' data.

3.3. Experimental Setup

3.3.1. Arm-Tracking

The sensors were placed in the arms so we can know the current orientation of each part of the arm. We then make use of forward kinematics to calculate the exact position of each bone.

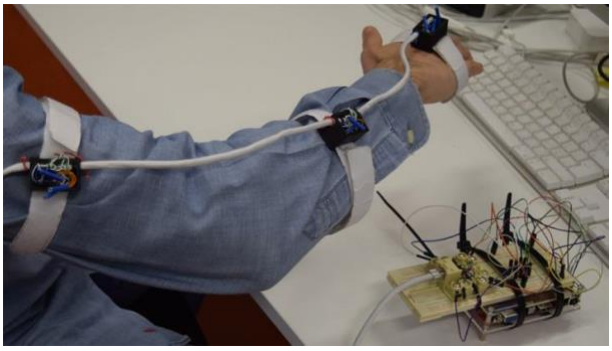


Figure 6: The Hardware Prototype

3.3.2. Independent Variable

The variable that is studied is the Sample Rate of the tracking of the arm movements. The magnetometer is set, permanently, at its maximum (8 Hz), because it is a value much lower than the values we were studying. The accelerometer and gyroscope Sample Rate is then set at 15, 30, 60, 90 and 120 Hz and the experiment is run.

3.3.3. Experimental Task

The user is presented with a virtual avatar, viewed from a first-person perspective (Fig. 2 and Fig. 3 illustrates this). Now the user can play around with the avatar for 30 seconds. After this period, the user's first task is started. The order of the experiments was alternated between users to avoid biasing the results based on learning effect. The Sample Rate order was always chosen by the random list generator from www.random.org/lists.

The following two tasks were crafted in a way that allow us to probe both slow (balancing a ball) and fast movements (hitting the bears in succession) and study how this affects the sample rate that the users feel they need and their measured performance at each sample rate.

3.3.3.1. Task 1 – Balance Ball

The user was requested to move the arm, so it stays on a starting position. There, a ball was dropped after a short count down. The user was then required to balance the ball for as long as he can. A perfectly vertical shadow indicated where the ball would fall. This was used as an aid to the lack of good depth perception to the user, where the user is expected to use depth on his movements. This task was chosen to evaluate conditions where most concentration lies on controlling an object on a slow and predictable environment (the ball responds only to the gravity and the forces the user applies). We measured the time the user is able to balance the ball. The test is repeated for 10 iterations and then the Sample Rate is, randomly switched. On each switch, the user answers the questionnaire.

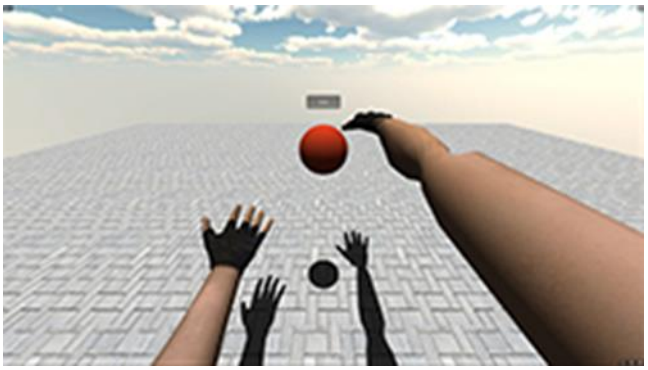


Figure 7: *The first task: Balance Ball*

3.3.3.2. Task 2 – Whack-a-Bear

The user is requested to hit a teddy bear that spawns on a random position on top of a table, as if playing a game of “Whack-a-Mole”. The spawn position was set such as it could never spawn too close to the previous position so a double hit could happen. This task was chosen to evaluate conditions where fast, far away, and precise movements are required due to the semi-randomness of the environment. The bear spawns 30 times for each Sample Rate, which is then randomly switched. On each switch, the user answers the questionnaire.

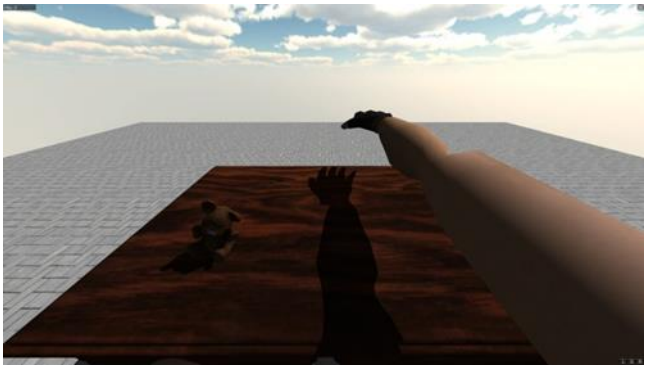


Figure 8: The second task: Whack-a-Bear

3.3.4. Data Extraction and Analysis

For the “Whack-a-Bear” experiment, we log the time it took for the user to touch the bear. For the “Balance Ball” experiment, we log for how long the user was able to balance the bear. The performance data was then analysed by calculating the average, standard deviation and also by comparing the averaged values of higher and lower Sample Rates. After those preliminary tests and analysis, we performed T-Tests of Student on each Sample Rate pair for the Balance Ball experiment time that the user was able to keep the ball in balance and for the Whack-a-Bear experiment time between bear hit. For each T-Test we also calculated the P-Value, for a two tail and the Effect Size. The used confidence interval was 95%.

The subjective data was averaged, and its standard deviation was calculated. We then proceeded to analyse discrepancies in the data (for instance, if the Sample Rate increases but the users think it was slower or harder, this indicates that the user may not be able to accurately notice (or, at least, express) what changed. We then averaged all the increases and decreases of the Sample Rates and performed the previous stated analysis.

4. Results

4.1. Balance Ball

The tests indicate a good effect when going from 15 to 120 Hz (t=0.07, P=0.95, effect size=0.01) but also a strong “non-effect” when going from 30 to 60 Hz (t=2.01, P=0.05, effect size=0.31). All the remaining data show no statistical relevance. When comparing the relation between the time it takes to complete the tasks on the different sample rates, we see no tendency in the data.

4.2. Whack-a-Bear

The tests show only a strong “non-effect” when going from 15 to 30 Hz (t=2.01, P=0.0). There seems to be no tendency in the data when comparing the times.

4.3. Questionnaire

For the Whack-a-Bear experiment, the user can always notice a good improvement (2.0) when going up from 15 Hz. But they can also notice a slight improvement (~0.5) when going down from 120 Hz, even to the other extreme of 15 Hz. For the remaining of the results, the users can guess about 53% of the time. As for the Balance Ball, there is a less pronounced effect of the user noticing an improvement by going either direction from one extreme to the other. The user is also more prone to correctly “guess” (65%) which direction the Sample Rate changed to, even if just slightly. There seems to be a bias towards positiveness on all the results, regardless of the tested direction.

5. DISCUSSION

The user performance values mostly indicate that no effect is present. When they indicate otherwise, those results are denied by other results. Since it makes no sense finding an improvement in both directions, those values may be disregarded as noise. This possibility is corroborated by the fact that the tests that may indicate an effect are those with the biggest standard deviation. This may be caused by an accumulated learning effect.

When looking at the relation between two Sample Rates for the Balance Ball, we can, easily, see that the ratio is split between being what it is expected (improving when increasing the Sample Rate and degrading when lowering the Sample Rate) and what is not (the opposite). In fact, it lays on the 50% mark. The Whack-a-Bear is more as expected marking at 20% contradictory results. Still, if we average all the relations, in both experiments we end up with a value of 1.0 for each which indicates that, on average, the performance is indifferent to the Sample Rate, for the tested interval. The indifference in performance was expected when the experiment was designed but only for the higher values. Having no performance difference between 15 Hz and 120 Hz came out as a surprise and one that is not easy to explain. We speculate that either the Madgwick's algorithm is good enough to compensate for the lower Sample Rate or that we did not design the experiment in a way that would reach a breaking point in speed and/or precision. Still, the reason could be that, in fact, the optimum value is, indeed, around or below 15 Hz.

As for what the user feels, a correct guessing of 53% and 65% for Whack-a-Bear and Balance Ball, respectively, seems to indicate that the user may not be sure what he is really feeling. It could also indicate that the user was not able to, correctly, communicate what he really felt or that the questionnaire was ill built or incomplete. There is also a bias towards positiveness that raises some red flags. This bias could be explained by either the user feeling a need to find an improvement, even if there is not one present or by the learning effect being strong enough for the user to confuse learning with technical improvement.

By adding personal remarks from the users' interactions during the experiments, we can state that we noticed that some users were really able to correctly and consistently guess the direction of the Sample Rate change. But those users represent a small portion of the whole sample. Unfortunately, we did not take note of the exact number of users. On the other hand, the majority of the users showed no clue of whether the Sample Rate had increased or decreased. This led us to conclude that there may be characteristic or subpopulation that has higher sensitivity to a Sample Rate change. As of this moment we were not able to identify what characteristic or subpopulation it may be.

6. CONCLUSION

Performance-wise we found no evidence that changing the Body Tracking Sample Rate would change the user performance when performing tasks, be it slow or fast or even precise. We found, however, that a small group of users may notice the change in the body tracking Sample Rate. We did not find where the exact threshold is, how strong that effect is and what makes those users being able to notice the body tracking Sample Rate change. This leaves a good margin for developers to save energy and bandwidth when tracking arms.

Acknowledgments

This research was partially funded by IDERAM through grants no. M1420-01-0247-FEDER-000019 and M1420-01-0247-FEDER-00003

REFERENCES

H. Dejnabadi, B. M. Jolles, and K. Aminian. (2005) A New Approach to Accurate Measurement of Uniaxial Joint Angles Based on a Combination of Accelerometers and Gyroscopes. IEEE Transactions on Biomedical Engineering, vol. 52, no. 8, pp. 1478–1484.

H. J. Witchel, C. Westling, A. Healy, N. Chockalingam, and R. Needham. (2012) Comparing four technologies for measuring postural micromovements during monitor engagement. p. 189.

K. Lorincz, B. Chen, G. W. Challen, A. R. Chowdhury, S. Patel, P. Bonato, and M. Welsh. (2009) Mercury: a wearable sensor network platform for high-fidelity motion analysis. p. 183.

R. Aylward and J. A. Paradiso. (2007) A compact, high-speed, wearable sensor network for biomotion capture and interactive media. p. 380.

S. O. H. Madgwick, A. J. L. Harrison, and R. Vaidyanathan. (2011) Estimation of IMU and MARG orientation using a gradient descent algorithm. pp. 1–7.

T. Cloete and C. Scheffer. (2008) Benchmarking of a full-body inertial motion capture system for clinical gait analysis. pp. 4579–4582.

T. Cloete and C. Scheffer. (2010) Repeatability of an off-the-shelf, full body inertial motion capture system during clinical gait analysis. pp. 5125–5128.

X. Wei, P. Zhang, and J. Chai. (2012) Accurate realtime full-body motion capture using a single depth camera. ACM Transactions on Graphics, vol. 31, no. 6, p. 1.

Omnichannel Heuristics for E-commerce

Sameer Kharel, Mikael Fernström and Bal Krishna Bal

Kathmandu University, Dhulikhel, Nepal and University of Limerick, Limerick, Ireland
sameer.kharel@ku.edu.np, mikael.fernstrom@ul.ie, bal@ku.edu.np

Many organizations are providing their services via web and apps, however, appropriate methods for measuring usability and user experience of a digital ecosystem seem to be largely lacking. E-commerce has become popular in developing countries like Nepal and its usefulness is found to be high for example during emergencies like the pandemic. The purpose of this research is to compare one of the mostly used inspection evaluation method i.e. Nielsen's heuristics with Omnichannel heuristics method which is developed considering the digital ecosystem in the context of e-commerce in omnichannel media.

The comparative study showed that Omnichannel heuristics detect more usability and UX issues than Nielsen's heuristics as much as two-third more than the latter. Omnichannel heuristics showed more high-priority issues, compared to Nelsen's heuristics. Omnichannel heuristics were found to be more effective for usability evaluation of e-commerce, compared to Nielsen's heuristics.

Heuristic Evaluation; Omnichannel; usability evaluation; e-commerce.

1. Introduction

Since the mid-1980s, there is increased interest in the field of Human Computer Interaction (HCI). Today's ubiquitous nature of computers has added more responsibility to people working in the HCI. Usability has been a core concept in HCI, which has slowly shifted toward user experience (UX). ISO 9241-11 defines usability as "the extent to which a product can be used by specified users to achieve specified goals with effectiveness, efficiency, and satisfaction in a specified context of use" (Bevan, 2001). Usability measurement has been an important part of user-centered design (UCD). The term emerged in the early 1980s. The main principle was to keep the user needs at the center of design (Marti and Bannon, 2009). In today's world, technology is changing more rapidly than ever before. People are trying to utilize technology to the fullest to accomplish things in their daily activities. Easy access to the Internet with high bandwidth has not limited its use to desktop but has expanded use via mobile devices. Users can now communicate or access the same service via multiple channels.

E-commerce and omnichannel have been hugely utilized by developed countries operating their business. Developing countries like Nepal were found to be trailing behind in mastery of e-commerce. The trend in the Nepal market scenario has been changing. E-commerce has become lucrative platform for entrepreneurs of Nepal as it is found to be profitable for their ventures (Ngudup et al., 2005) and the COVID-pandemic has been a game-changer for e-commerce businesses. Many people have started to use e-commerce due to situational demands in a country like Nepal. Each year new e-commerce enters the Nepal market, mostly with websites and app versions in parallel. It was found that the majority of Nepalese customers face issues when using e-commerce (Vaidya, 2019) which is believed to have much more potentials than what it is offering at the moment by using UCD during development of the e-commerce applications. It was found from interviews conducted with Nepalese UX designers that in most cases the design of websites was transformed into its app versions without doing proper research on it. From interviews, it was also found lack of communication between a website and app development teams leading to inconsistent design affecting seamless experience when switching from web version to app version and vice-versa. Usability and UX evaluation were rarely performed, citing lack of budget, time, and experts to conduct an evaluation.

Research found customers abandon e-commerce due to poor site organization (Sivaji et al., 2011). It is important to retain customers to get more revenue by improving the usability and UX of e-commerce. It was found that little emphasis was given on formal usability within the e-commerce sector compared with mainstream business applications (Sartzetaki et al., 2003; Jach and Kuliński, 2012). It was found that heuristic evaluation is a useful usability evaluation method for e-commerce (Sivaji et al., 2011).

2. Related works

Heuristic evaluation, which is the most used inspection method, is easy to perform, and it is inexpensive and effective to use (Inostroza et al., 2013). Usability problem detection depends upon the expertise of the evaluator and computer professional. The heuristic evaluation recommends a small group of experts rather than a single evaluator to evaluate an interface (Hertzum and Jacobsen, 2001). Jakob Nielsen's heuristics is one

of the most used heuristics (Nielsen, 1995; Molich and Nielsen, 1990; Liu, 2008). Nielsen and Molich developed this method focusing on telecom systems. It is often argued that it cannot be directly applicable to other software systems (Harrison and Duce, 2013). Another drawback is that it may not be appropriate to find software/device specific problems. Emerging smart devices usability problems cannot be addressed by traditional ways of thinking. New heuristic evaluation methods are needed to address new context i.e. omnichannel.

Zhang's heuristics were developed to address usability problems of specific devices. It was developed to inspect medical devices and particularly used to identify patient safety through the identification and assessment of usability problems (Zhang et al., 2003). Based on Nielsen's heuristic evaluation and Shneiderman's (Shneiderman's,1998) "eight golden rules," Zhang extended his heuristics, from Nielsen's 10 to 14. Zhang tried to incorporate Norman's "7-stages of actions." These heuristics, however, have the same disadvantage as previous. It is targeted to evaluate particular devices and could not address challenges posed by smart devices.

The above-mentioned heuristic evaluations lack the ability to address technology like mobile devices. Addressing this, Yáñez Gómez, R., et al. developed new heuristic guidelines (Yáñez et al., 2014). It consists of a list of 13 heuristics. It has incorporated every important aspect like pleasant and respectful interaction and privacy in the heuristic list. It has a sub-heuristic list which addresses the specific character of mobile, e.g. minimum input needed, general visual cues, fat-finger syndrome, among others. Gomez's heuristics argued to best fit for mobile device evaluation compared to the previously-mentioned heuristic evaluations. These heuristics are primarily focused on touch screen phones and tablets but has not considered smartphones that we now find in most people's pockets and the concept of multiple channel is missing.

The heuristics of Silva et al., which evolved from Nielsen, included 33 heuristics for evaluating smartphone apps targeted at older adults. Results of these heuristics show that all heuristics on the list were used in the evaluation of a smartphone app. Some of the unanswered questions of these heuristics are if this can be equally applied for desktop and mobile websites (Silva et al., 2015).

The above-mentioned heuristics did not consider a customer journey. Customer journey maps are used for understanding the customers' experience. It refers to a pictorial representation of interaction done by customers, possibly through different media at different times with the organization or service provider (M. S. Rosenbaum et al., 2017). Laia Bonastre and Toni Granollers considered customer journeys in their heuristic during purchase in a single channel, i.e. a website (Bonastre and Granollers, 2014). Nowadays, most e-commerce provides services through both website and app versions. The concept of multiple channels is lacking in the Laia and Toni heuristics i.e. website and an app of e-commerce in a single ecosystem. Multiple channel issues were addressed by Brad Aabel and Dilini Abeywarna for digital health (Aabel and Abeywarna, 2018). This heuristic cannot be directly applied to e-commerce. It also lacks concepts of sociability, privacy, and security, which are important for e-commerce. The purpose of Omnichannel heuristics is to measure the user experience of a website and an app of e-commerce considering it as a single ecosystem.

As Nielsen's heuristics is the mostly used heuristics evaluation, it is interesting to explore Nielsen's heuristics effectiveness for evaluating omnichannel context for e-commerce and compare results with Omnichannel heuristics. The Omnichannel heuristics were developed by reviewing previous work done to evaluate web and app, output of workshop conducted in Nepalese software companies, feedback from academics and Software companies working on UX, a case study of different e-commerce based on Nepal like "Daraz", "e-Sewa", "Hamrobazar", "IME Pay", "Khalti", and "NETTV", surveys with Google Docs users and The Buddha Air customers, Interviews with Daraz and e-Sewa customers and Interviews with UX designers, UI designers, developers, and Quality Assurance (QA) people.

3. Results

Three evaluations were conducted to compare the two heuristic evaluation methods. The evaluators were software developers, designers, UX engineers, and UX designers from Nepal based software companies. Different sets of five evaluators were used to evaluate "Buddha Air" using Nielsen's heuristics and Omnichannel heuristics. Three scenarios were used for evaluation. Scenario one was related to flight information and flight status, scenario two to flight routes, and scenario three to booking a flight. Similarly, three evaluators were used to evaluate "Foodmandu" and four scenarios were used for it. The first scenario was related to food order, the second with payment, the third with food delivery, and the fourth was related to listing new restaurants in Foodmandu. Five evaluators were used to evaluate "Daraz" using four scenarios. The first scenario was related to the registration process, the second scenario with a search for a product, the third scenario with payment, and the fourth scenario with listing purchasing experience in Daraz.

Min-Max normalization was used to normalize usability and UX issues found by using Nielsen's heuristics and Omnichannel heuristics. Its objective was to find average value of the issues and compare them.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

Table 1: Normalized Data

Heuristics	Buddha Air	Foodmandu	Daraz
Nielsen's heuristics	*0.40*	*0.37*	*0.23*
Omnichannel heuristics	*0.37*	*0.48*	*0.38*

A Comparison between the two heuristics was also done by the priority given on each issue by evaluators. Where P1 represents low priority issue and P4 represents high priority issue respectively.

Table 2: Issues by priorities of Buddha Air

Priority	Nielsen's heuristics	Omnichannel heuristics
1: Low	*42.86%*	*22.50%*
2: Medium	*37.14%*	*30%*
3: High	*20%*	*42.50%*
4: Catastrophic	*0%*	*5%*

Table 3: Issues by priorities of Foodmandu

Priority	Nielsen's heuristics	Omnichannel heuristics
1: Low	*22.22%*	*3.23%*
2: Medium	*59.26%*	*51.61%*
3: High	*14.82%*	*25.81%*
4: Catastrophic	*3.70%*	*19.35%*

Table 4: Issues by priorities of Daraz

Priority	Nielsen's heuristics	Omnichannel heuristics
1:Low	*10.53%*	*8.99%*
2:Medium	*39.47%*	*40.65%*
3: High	*42.11%*	*46.76%*
4: Catastrophic	*7.89%*	*3.60%*

Comparing evaluators' comments from three evaluations it was found Nielsen's heuristics were unable to address issues related to channel strength and seamless experience, which were addressed by Omnichannel heuristics. Omnichannel heuristics addressed both pragmatic and hedonic issues whereas Nielsen's heuristics only addressed pragmatic issues. Nielsen's heuristics missed the concept of collaborative work and the importance of brand value.

For full details evaluators comment refer https://sameerkharel.wixsite.com/nepal/evaluators-comments

4. Discussion

Table 5: Comparison between Nielsen and Omnichannel heuristics

Nielsen Heuristics	Omnichannel Heuristics
Visibility of system status	*Consistency*
Match between system and the real world	*Information design and content*
User control and freedom	*User control*
Consistency and standards	*Consistency*
Error prevention	*Prevent Errors*
Recognition rather than recall	*Consistency, partial(Branding, Seamless experience)*
Flexibility and efficiency of use	*Know your audience*
Aesthetic and minimalist design	
Help users recognize, diagnose, and recover from errors	*Sub-heuristics of "Prevent Errors"*
Help and documentation	*Help and Documentation*
	Ease of use
	Identify channel's strengths
	Sociability
	Privacy
	Security

Note: Heuristics in table 5 in the same row reflects similarities between two heuristics.

Table 5 shows that Nielsen's heuristics do not say anything related to channel strength whereas as Omnichannel do not say much about "Aesthetic and minimalist design". The concept of collaboration and emotional brand attachment has not been addressed by Nielsen. Sociability is an important concept for e-commerce which is not addressed by Nielsen's heuristics. Also, "privacy" and "security" have not been addressed in Nielsen's heuristics but is covered in Omnichannel heuristics, which is important for e-commerce. For detail heuristics see https://cfe7ec4e-6c4f-4b08-bedb-580ef0a5cba2.filesusr.com/ugd/722a37_62cf4a7b6033415780da4c722c3a1fc4.pdf.

Min-Max normalized values from three evaluations using Nielsen's heuristics and Omnichannel heuristics show that 66.67% of Omnichannel heuristics detected more usability and UX issues than Nielsen's heuristics. Three evaluations indicate that both heuristics found common usability and UX issues. Different sets of usability and UX issues were found by Nielsen's and Omnichannel heuristics. Nielsen's heuristics evaluators had less agreement between them compared with Omnichannel evaluators.

Few evaluators felt that Omnichannel is subjective and evaluators might get biased as they might want to prove their skill that may not be the case with Nielsen's heuristics. Some evaluators believed Omnichannel heuristics evaluation takes longer time for evaluation compared to Nielsen's heuristics. Few suggested that further classification was needed in Omnichannel heuristics. Sometimes lack of freedom is provided by Omnichannel heuristics as evaluators only stick to its sub-heuristics and are seen to provide less in-depth analysis on a few issues over Nielsen's heuristics.

5. Conclusion

The results from Omnichannel heuristic evaluation show that in most cases more usability and UX issues were detected than with Nielsen's heuristics. The result shows high priority usability and UX issues were found by Omnichannel heuristics compared to Nielsen's heuristics. It was found that in most cases comments written by Nielsen heuristics evaluators but not in Omnichannel heuristics were addressed by Omnichannel sub-heuristics. The result showed Omnichannel was able to detect both hedonic and pragmatic issues, but Nielsen's lack hedonic issues in most cases. Omnichannel heuristics look promising for addressing issues of channel strength and seamless experience, compared with Nielsen's heuristics.

6. References

Aabel, B. and Abeywarna, D., (2018) . Digital Cross-Channel Usability Heuristics: Improving the Digital Health Experience. Journal of Usability Studies, 13(2).

Bevan, N. (2001). International standards for HCI and usability. International Journal of Human Computer Studies, 55(4), 533–552. https://doi.org/10.1006/ijhc.2001.0483

Bonastre, L. and Granollers, T., (2014) A set of heuristics for user experience evaluation in e-commerce websites. In 7th International Conference on Advances in Computer-Human Interactions, pp. 27-34, IARIA.

Harrison R, Flood D and Duce D., (2013) Usability of mobile applications: literature review and rationale for a new usability model, Journal of Interaction Science,1(1):1-6.

Hertzum, M. and Jacobsen, N.E., (2001) The evaluator effect: A chilling fact about usability evaluation methods. International journal of human-computer interaction, 13(4), pp.421-443.

Inostroza, R., Rusu, C., Roncagliolo, S. and Rusu, V (2013). Usability heuristics for touchscreen-based mobile devices: update. Proceedings of the 2013 Chilean Conference on Human-Computer Interaction,, pp 24-29, ACM.

Jach K and Kuliński M., (2012) Heuristic Evaluation for e-Commerce Web Pages Usability Assessment, Advances in Usability Evaluation 460.

Liu, F., (2008) Usability evaluation on websites. In Computer-Aided Industrial Design and Conceptual Design, 2008. CAID/CD 2008. 9th International Conference, (pp. 141-144, IEEE.

Marti, P. and L. J. Bannon (2009) Exploring user-centred design in practice: Some caveats. Knowledge, technology & policy 22(1): 7-15.

Molich, R. and Nielsen, J. (1990) Improving a human-computer dialogue. Communications of the ACM 33(3): 338-348.

Ngudup, P., Chen, J. C., and Lin, B. (2005) E-commerce in Nepal: a case study of an underdeveloped country. International Journal of Management and Enterprise Development, 2(3-4), pp. 306-324.

Nielsen, J., (1995) 10 Usability Heuristics for User Interface Design, https://www.nngroup.com/articles/ten-usability-heuristics/ (Access 06-03-2018)

Rosenbaum, M. S., Otalora, M. L., & Ramírez, G. C. (2017). How to create a realistic customer journey map. Business Horizons, 60(1), 143–150. https://doi.org/10.1016/j.bushor.2016.09.010

Sartzetaki, M., Psaromiligkos, Y., Retalis, S. and Avgeriou, P., (2003) An approach for usability evaluation of e-commerce sites based on design patterns and heuristics criteria. University of Groningen, Johann Bernoulli Institute for Mathematics and Computer Science.

Shneiderman, B., (1998) Designing the User Interface: Strategies for Effective Human-Computer Interaction, Addison-Wesley Publ. Co., Reading, MA.

Silva, P.A., Holden, K. and Jordan, P., (2015) Towards a list of heuristics to evaluate smartphone apps targeted at older adults: a study with apps that aim at promoting health and well-being, 48th Hawaii International Conference on System Sciences, pp. 3237-3246, IEEE.

Sivaji, A., Downe, A. G., Muhammad Fahmi Mazlan, Soo, S.-T., & Abdullah, A. (2011) Importance of incorporating fundamental usability with social & trust elements for E-Commerce website. 2011 International Conference on Business, Engineering and Industrial Applications, pp 221-226, IEEE.

Vaidya, R. (2019). Online Shopping in Nepal: Preferences and Problems. Journal of Nepalese Business Studies, 12(1), pp 71–86.

Yáñez Gómez, R., Cascado Caballero, D. and Sevillano, J.L., (2014) Heuristic evaluation on mobile interfaces: A new checklist. The Scientific World Journal.

Zhang J, Johnson TR, Patel VL, Paige DL, Kubose T., (2003) Using usability heuristics to evaluate patient safety of medical devices, Journal of biomedical informatics, 1;36(1-2):23-30.

MailTrout: A Machine Learning Browser Extension for Detecting Phishing Emails

Paul Boyle and Lynsay A. Shepherd

School of Design and Informatics, Division of Cyber Secuirty, Abertay University, Bell Street, Dundee, UK

1600301@abertay, lynsay.sherpherd@abertay.ac.uk

The onset of the COVID-19 pandemic has given rise to an increase in cyberattacks and cybercrime, particularly with respect to phishing attempts. Cybercrime associated with phishing emails can significantly impact victims, who may be subjected to monetary loss and identity theft. Existing anti-phishing tools do not always catch all phishing emails, leaving the user to decide the legitimacy of an email. The ability of machine learning technology to identify reoccurring patterns yet cope with overall changes complements the nature of anti-phishing techniques, as phishing attacks may vary in wording but often follow similar patterns. This paper presents a browser extension called MailTrout, which incorporates machine learning within a usable security tool to assist users in detecting phishing emails. MailTrout demonstrated high levels of accuracy when detecting phishing emails and high levels of usability for end-users.

Phishing. Usable Security. Machine Learning. Browser Extension. Socio-Technical Security.

1. Introduction

Phishing emails generally attempt to persuade the recipient to reveal private or confidential information such as passwords or bank details and may deliver malware to infect the victim's machine. Information gained via phishing emails is used for fraudulent purposes by the sender, placing users at risk of identity theft, fraud, and significant financial loss.

Since the beginning of the COVID-19 pandemic, incidences of cyberattacks and cybercrime have increased considerably, including a sharp rise in phishing attempts (Lallie et al., 2021; Horgan et al., 2021). The pandemic has caused a fundamental shift in working practices and social interactions, creating an enhanced dependence on technology. Thus, users require additional support to identify potentially malicious emails.

Successful phishing scams can be costly for victims; in the UK, it is estimated that it takes 20 days and £960,000 to address the consequences of a single phishing or social engineering attack (Graham, 2018). To combat phishing attempts, email clients make use of spam filters to quarantine suspicious emails. However, these filters are not always successful; consequently, users require additional assistance to help them detect phishing emails in the form of anti-phishing tools and security education.

Anti-phishing tools may take the form of browser extensions, which can augment the users' browsing experience. These tools can identify different forms of phishing attacks; 'GoldPhish' is an Internet Explorer extension used to identify phishing webpages (Dunlop et al., 2010), while 'PhishAri' is a Google Chrome extension designed to detect phishing attempts on Twitter (Aggarwal et al., 2012). Anti-phishing tools may also make use of machine learning (ML), allowing systems to learn from existing data to make decisions without the need for human interaction. Previous work by Fette et al. (2007) was able to detect phishing emails based on features such as the number of hyperlinks present and the use of JavaScript.

This paper presents a prototype browser extension to detect phishing emails, which harnesses the power of machine learning to assist users in identifying phishing attempts, protecting them from becoming a victim of cybercrime. TensorFlow was used to develop and train an ML model using a dataset of fraudulent and legitimate emails. The model was evaluated for accuracy and converted for use in a prototype Google Chrome browser extension. The extension parses email text and evaluates sentiment and language to determine legitimacy. The extension was also tested with participants to evaluate its usability.

The remainder of the paper is organised as follows; Section 2 explores related work in phishing detection and machine learning. Section 3 describes the methodology. Results are presented in Section 4 and are discussed in Section 5. Section 6 presents conclusions and considers future work.

2. Related Work

2.1 Existing anti-phishing tools

Phishing emails are not a new problem; however, attempts have increased in the wake of the COVID-19 pandemic attempts (Lallie et al., 2021). Usable security research has investigated anti-phishing tools to protect users and increase the awareness of risks associated with phishing attempts. A vital consideration when developing security tools – especially those aimed at non-technical individuals – is ensuring that they are accessible and user-friendly. Kumaraguru et al. (2010) developed two anti-phishing tools: the embedded email-

based ‘PhishGuru’ and the online game ‘Anti-Phishing Phil’. To ensure the tools provided effective education, the developers followed a series of design principles, including ‘learning-by-doing’, which states that people learn better when they practice their skills. In PhishGuru, instructional materials are embedded into the user’s everyday tasks, such as checking their emails. Implanting the materials increases the prevalence of ‘teachable moments’ – optimal opportunities to convey a point or idea – increasing the tool’s educational potential. ‘Anti-Phishing Phil’ and the concept of embedding phishing content into games has been explored by Dixon et al. (2019). The work highlighted that users prefer integrated tools to help them learn, and they would not seek out a game for the sole purpose of learning about phishing.

Embedded tools may take the form of browser extensions. GoldPhish, developed by Dunlop et al. (2010), was an extension for the now deprecated Internet Explorer browser, allowing it to access the sites viewed by the user to identify phishing.

Aggarwal, et al. (2012) developed ‘PhishAri’, a Chrome browser extension used to detect phishing attempts on Twitter. The researchers found that phishing attacks carried out through social media sites have risen, and a common technique used is the obfuscation of malicious web links through URL shortening. The extension uses ML techniques to classify phishing URLs and tweets through characteristics of the URL, the tweet, and the author. The tool applies a red indicator to phishing tweets and a green indicator to safe tweets.

2.2 Machine Learning (ML)

The language patterns commonly reused in phishing attacks have generated interest in how ML can identify and protect users from phishing attacks due to its ability to classify data by identifying trends.

ML models require input data stored in a numerical format for processing. Data can come from a variety of sources, including images and text converted to numerical vectors. In the field of natural language processing (NLP), structured collections of text referred to as ‘corpora’ are used as datasets for training. Converted data can make predictions on non-numerical information, using qualities such as its visual appearance or use of language.

Fu et al. (2006) proposed a method for detecting phishing webpages by assessing visual similarities between a potential phishing site and a set of protected sites known to be legitimate. The research interpreted the colour and location of each pixel on a webpage as data when making a prediction. However, this method only detects phishing pages that look similar to those in the protected set, with less success at detecting phishing web pages outside of this set. Fu et al. (2006) cited natural language analysis to enhance the project, improving detection accuracy.

The GoldPhish browser extension by Dunlop et al. (2010) uses optical character recognition (OCR) to detect a company logo on a webpage and converts it to text. Google PageRank is then used to compare the top domains with that name to the current webpage. However, one potential issue with this method is that webpage logos may be highly stylised, rendering them difficult for OCR to interpret.

The aforementioned research has explored phishing webpages, which contain more graphical content than phishing emails. Image-based phishing detection is less flexible than text-based detection. It can only detect images similar to those used during model training and is dependent on the accuracy of external technologies, such as OCR. Thus, it is essential to focus on the text content using sentiment analysis, which has been applied to other contexts.

Tao and Fang (2020) proposed a multi-label sentiment analysis method to determine the sentiment of online reviews for restaurants, wines and films. This method allows the sentiment towards specific aspects of a sample to be analysed, rather than producing a prediction for the overall sentiment. For example, a review for a restaurant may express a positive sentiment towards the food but a negative sentiment towards the atmosphere.

While emails may contain some common features, such as greetings and sign-offs, these are not present in all emails. Also, compared to descriptions of specific features of an object, such as a wine’s variety or country of origin, these email features are more abstract and may be more difficult for an ML model to identify. However, this method used a multi-class approach, allowing samples to be classified as positive, negative, neutral or conflicted (both positive and negative). Such an approach may be applicable when identifying phishing emails, as it may produce more accurate results, considering different types of phishing attempts.

Other important factors to consider relating to the dataset used in training are its quality, size, and format. Halgaš et al. (2019) proposed a phishing classifier that uses a recurrent neural network (RNN) to evaluate an email’s text and structure. Researchers highlighted the ability of phishing emails to avoid filters due to their

changing nature and suggest that ML may be able to identify trends in phishing emails. Two datasets comprised of legitimate and phishing emails sourced from existing email corpora were used to train the model. Of the two datasets used, the RNN classified emails more accurately when trained with the smaller and less balanced of the two datasets, demonstrating that both quantity and quality of a corpus impact a model's accuracy. This method classified emails as either 'ham' (legitimate) or phishing. However, this binary classification system may have impacted the model's accuracy, given the many differences in language used in the numerous types of phishing attacks, such as extortion compared to unexpected money fraud.

Prusa et al. (2015) investigated the correlation between the size of a training dataset and the accuracy of a sentiment analysis classifier, explicitly studying the number of instances required to train a tweet sentiment classifier. The researchers found that as the size of the dataset used for training increased, the accuracy of the machine learning model improved. However, there was no significant improvement in the accuracy of this classifier after the use of a dataset containing 81,000 instances. The sentiments of tweets were classified as either positive or negative, which are very general terms (Prusa et al., 2015).

ML techniques can be applied to the field of usable security. Given the increased need for usable, anti-phishing tools and the ability of ML to detect patterns in data, this highlights the potential for these research areas to be combined, thus protecting users and enhancing phishing detection. In the following section, the methodology behind the research is outlined, explaining how an ML model was integrated into an anti-phishing browser extension to support end-users.

3. Methodology

An ML model was trained to classify emails as phishing or legitimate and was designed to produce a classification prediction based on an email's text contents. The browser extension operated by reading and processing selected text to generate an output in a popup window.

The browser extension and the ML model were integrated into a single extension named MailTrout (Figure 1). The browser extension selected and read text from the browser window and converted the text into a numerical sequence for processing. The ML model then generated a prediction based on the sequence. Finally, the browser extension displayed an output based on the prediction of the ML model.

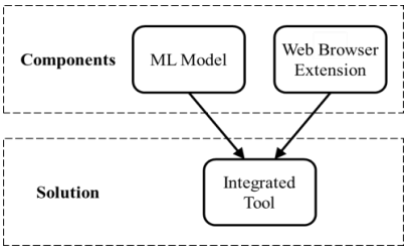


Figure 1: *Components within MailTrout*

3.1 Machine learning model

The ML model was developed using Python 3, the Python deep-learning library Keras (Chollet, 2015) and the open-source ML library, TensorFlow (Google, 2020a).

3.1.1. Algorithm selection

Artificial neural networks (ANNs) are computational algorithms based on the model of biological neurons in the human brain. ANNs can be used in ML to process input data and produce an output, such as a classification or prediction (Chen et al., 2019). Recurrent neural networks (RNNs) are variants of ANNs. The results of previous items in a sequence – such as words in a text – are stored to provide contextual information and produce results based on both the current and previous input. This method is ideal for NLP as it can evaluate the sentiment of text overall by evaluating words individually as well as in their context by considering the impact of the previous text (Lai et al., 2015).

Long Short-Term Memory networks (LSTMs) are an RNN architecture designed to cope with long-range dependences. As the distance between previous information and present input data grows, traditional RNNs become less effective at connecting this information to apply context. However, LSTMs are more capable of learning long-term dependencies, as they use multiple neural network layers to pass the neuron's output value and a memory cell state along the network, providing contextual information that can influence the output value

at each stage. Due to this technique, LSTMs are shown to outperform standard RNNs at learning context-free and context-sensitive language (Gers & Schmidhuber, 2001; Sak et al., 2014).

Bidirectional Long Short-Term Memory networks (BLSTMs) further improve the ability to learn long-term dependencies. BLSTMs operate on the input sequence from both directions, allowing the network to incorporate context from before and after the present item in a sequence. This method has proven to be powerful in tasks involving NLP, including sentiment analysis and classification (Wang, et al., 2015). For these reasons, the ML model was designed to use a BLSTM layer to process data.

3.1.2. Classification

Binary-class models allow data to be classified as one of two categories, typically 'positive' or 'negative'. While this approach could be applied in this research, the issue of the many differences in phishing email patterns and vectors had to be considered. Avanan's Global Phish Report (2019) classified the phishing emails reviewed into four vectors: spearphishing, extortion, credential harvesting and malware phishing.

Spearphishing attacks - phishing targeted at a specific individual, such as a high-level employee in an organisation - were commonly found to impersonate senior employees such as CEOs. Spearphishing uses social engineering to urge their victim to complete a task, such as granting the attacker access to company information or finances. This form of attack is known as business email compromise.

Extortion attacks use threats to pressure their victim, e.g. threatening to share compromising information, holding them to a cryptocurrency-based ransom. These emails often use email spoofing techniques and passwords uncovered from data leaks to add credibility to their claims.

Credential harvesting attacks aim to steal sensitive information from their victims, such as passwords or bank details. These attacks commonly impersonate trusted brands and lead the victims to phishing webpages, using social engineering to create a sense of urgency.

Malware phishing attempts seek to install harmful software on a victim's device. These exhibit characteristics similar to the aforementioned attacks.

Postolache and Postolache (2010) also identified numerous phishing vectors, including extortion and the impersonation of legitimate organisations and individuals. However, they also identified numerous vectors not covered by these terms, including advance-fee, lottery and investment fraud. These are examples of unexpected money and winnings scams, in which a scammer attempts to make a victim believe that they can receive a financial or material reward by following their instructions, such as by sharing their bank details or paying an upfront fee (Australian Competition & Consumer Commission, 2015). These investigations highlight the broad range of phishing email vectors in use and pose an issue for an ML model; as classification predictions are most accurate when items of a class have more similarities, a model's accuracy may be hindered by large differences in the data.

To reduce issues with accuracy, a multi-class approach was chosen for the ML model, in which text could either be classified as legitimate (HAM) or one of four classes of phishing: impersonation phishing (IMP), business email compromise (BEC), extortion (EXT) or unexpected money/winnings scams (UNX). This method ensured that data used for training could be sorted into classes of as little variance as possible. The approach helped ensure the ML model's accuracy, allowing the finished product to produce information specifically relevant to the type of phishing email that the user had likely received.

Finally, the ML model used the softmax function to output results as a probability distribution. Softmax normalises output by converting a vector of numbers to values between 0 and 1 that have a sum of 1, allowing each result to be interpreted as a probability (Goodfellow et al., 2016). This approach allows the model to output the certainty of its result, which may be helpful to a user when considering if they should follow the actions recommended by the browser extension in response to an email message they have received.

3.1.3. Datasets

The Fraud Email Dataset published by Verma (2018) was included in the final dataset used to train the ML model. Verma's dataset contains fraud emails described as 'Nigerian fraud' (advance-fee scam) taken from the CLAIR collection of fraud email (Radev, 2008), and legitimate emails taken from the dataset of Hillary Clinton's emails released by the US Department of State (Kaggle, 2019).

Verma's dataset was chosen as it required little formatting or review; the data did not contain any email header information, only the body content, which the ML model was designed to process. Also, all items were labelled as either fraud (1) or legitimate (0), allowing for easy relabelling to UNX and HAM respectively, for, compatibility

with the ML model's multi-class system. Additionally, the Python Reddit API Wrapper (PRAW) was used to collect extortion phishing emails posted on a series of Reddit threads titled "The Blackmail Email Scam" (EugeneBYMCMB, 2019). Suitable entries were labelled as EXT and added to the final dataset.

Online records of phishing emails are commonly presented in the format of screenshots rather than plaintext copies. In response to this, a script was developed to use OCR technology to read and store the text of saved images of emails. The Python script 'Google Images Download' was used to download results of online image searches for examples of phishing emails. This script allowed for multiple prefix and suffix terms to be added to a keyword for individual searches. The approach allowed for greater automation of image acquisition by appending names of well-known banks and commerce platforms to a search of "impersonation phishing email". The script also allowed for colour filters to be applied to searches, which was used to specify black-and-white images for forms of phishing that were unlikely to include colours or images (Vasa, 2019).

The image results required manual review as many were not suitable, including infographics and images on the subject of email phishing. The suitable images were then compiled into folders manually, separated by their classifications. The free OCR engine Tesseract was used to interpret the text from the images (Google 2020b). The Python wrapper tool PyTesseract was used to include Tesseract as part of a Python script (Lee, 2020).

All emails (Table 1) were compiled into one large CSV file.

Table 1: Different types of phishing email in the dataset.

Email Category	**Count**
Business Email Compromise (BEC)	*391*
Extortion (EXT)	*1427*
Legitimate (HAM)	*5287*
Impersonation (IMP)	*541*
Unexpected Money/Winnings (UNX)	*3581*
Grand Total	***11227***

3.1.4. Training

The email text from the dataset was split into portions for training and validation of 80% and 20%, respectively, following the commonly used Pareto Principle (McRay, 2015).

High-frequency words that consume processing time but do not contribute to sentiment were filtered from the dataset. These words are known as stop words. The Natural Language Toolkit (NLTK) is a Python library used for NLP and includes a corpus of stop words, including "the", "a", and "also" (Bird et al., 2009). This corpus was used as part of the ML training script to find and remove all stop words present in the training data.

Words in the dataset were converted to integers using a tokenizer, which converts text into meaningful data or 'tokens'. The tokenizer used was included in the Keras Python library and was created using the 10,000 most reoccurring words in the dataset vocabulary. The tokenizer was exported as a JSON file so that it could be used later. Both the training and validation sequences were tokenized, and a separate tokenizer was used to convert the data labels to integers (Google, 2020c).

The sequences used to train the model had to be equal in size, meaning that sequences had to be padded or truncated to fit a set length. Sequence padding involves adding zeros to a sequence until it is the desired length. This can be done from the beginning (pre-padding) or the end (post-padding. The sequences were pre-padded, as this method has been shown to produce the most accurate results when used with an LSTM model (Dwarampudi & Reddy, 2019).

The standard sequence length chosen was 500 words. The mean word count of the emails used in training was 201, and the standard deviation approximately 266. The sequence length was calculated as the mean plus one standard deviation rounded to the nearest hundred. On inspecting the distribution of word counts of the emails, it was confirmed that this length was suitable, as most emails were within this range. Emails with a word count greater than 500 were truncated. Truncation can be carried out from the beginning (pre-truncation) or the end (post-truncation) of the sequence. There is no widely accepted best practice for

sequence truncation for LSTMs; therefore, post-truncation was used to avoid removing keywords or phrases commonly located near the beginning of phishing emails, e.g. “Dear Customer”.

3.2 Web browser extension

The browser extension was designed to help the user classify an email as legitimate or phishing, using the ML model. The extension was developed for use in Google Chrome, which has the largest share (StatCounter, 2021).

Figure 2: *Screenshot of the browser extension.*

The browser extension uses a complementary colour palette of turquoise and gold, ensuring that text and buttons are high contrast and easy to distinguish visually. The extension also uses green and red icons to highlight what the user should and should not do in response to receiving a potential phishing email. Green is commonly associated with safety, while red is associated with danger. The extension uses this recognisable colour scheme to make the meaning of its messages clear. Screenshots of the prototype were tested using ‘Coblis’, an online tool that allows the user to view an uploaded image as a person would see it with a colour vision deficiency (CVD) or colour-blindness (Wickline, 2001).

3.2.1. Implementation

In order to select an email to be evaluated by the extension, the user highlights text using their cursor and then selects a button on the extension popup. This method was deemed the most transferable for use in different web email clients as it did not require email contents to be automatically detected, ensuring that the extension processed only the text a user wanted to evaluate. The extension popup displays on the right side of the page; this is common practice for web browser extensions and would be familiar to the user. This also prevents disruption to the user’s browsing experience; given that the user is likely to have left-aligned text on a page, the popup will not cover any important parts of the text.

3.2.2. Readability

To ensure that users could easily understand the instructions, certainty of classification, and information given by the extension, the Python package ‘TextStat’ was used to evaluate the readability and complexity of the text in the extension’s instructions and results (Bansal & Aggarwal, 2020).

TextStat can be used to produce a readability score using numerous established readability formulas. The Dale-Chall readability formula was used to calculate the US grade level of text, which was then used to determine the average age level. According to Begeny and Greene (2014), the Dale-Chall formula outperforms other commonly used readability formulas as a consistent and accurate indicator of text difficulty. Text within

the extension was written to be understandable by those aged 18 majority – the legally recognised threshold of adulthood – in most countries (UN General Assembly, 1989).

3.3 Integration of model into extension

The ML model was converted from a Hierarchical Data Format Files to the TensorFlow.js Layers format, allowing for use with JavaScript as part of the web browser extension. The Layers formatted model consisted of a JSON file of the model architecture and a binary weights file. The JSON file was loaded into JavaScript using TensorFlow.js, allowing the browser extension to make and output predictions using the ML model.

The browser extension was designed to use the ML model to make classifications on the client-side. This approach ensured the analysis of emails would be faster compared to loading the model from a server (Figure 3). The model was loaded from the JSON file stored by the extension and generated a prediction based on the sequence. The prediction consisted of an array of probabilities that the sequence was one of the potential email categories. The browser extension then displayed a result based on the classification with the highest probability.

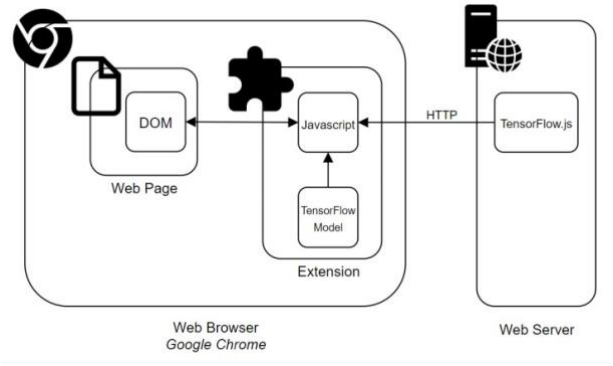


Figure 3: *System architecture.*

3.4 User testing

User acceptance testing was carried out to evaluate the browser extension's usability, with a total of 44 participants. Testing was conducted remotely owing to the COVID-19 pandemic and lockdown. Participants emailed the researchers indicating interest and were provided with a copy of the extension, along with an instructional YouTube video containing installation details.

An online survey used to record participants' feedback on the extension's usability was developed. Participants had to agree to an informed consent statement before proceeding with the experiment and were asked to provide demographic information. Participants were also asked about their familiarity with the terms used to describe the four categories of phishing emails. They were then given a fuller description of each category and asked to rate how likely they would be to identify an email of that category.

A scenario was given to participants to add a level of realism to the testing environment. This scenario stated that the participant was working for an organisation and had been asked to review their boss's email inbox using the MailTrout extension to identify phishing emails received. A webmail-style sandbox environment created using HTML, CSS, JavaScript, and PHP was developed and displayed to participants in the browser. The inbox randomly selected 10 test emails out of a possible 30 and displayed these to the participant one at a time, moving to the next email once the participant marked each as either legitimate or phishing. Once they completed the task, they were asked to consider how usable they found the extension, using an all-positive version of the System Usability Scale (SUS) as discussed by Sauro & Lewis (2011).

After using the extension, participants were again asked how likely they would be to identify phishing emails of each category. Additional questions explored how helpful they found the instructions provided for using the extension and how likely they would be to recommend the extension to someone looking to protect themselves against phishing emails. Participants were also asked to provide feedback on how the extension catered to any conditions they had which may impact their ability to use a browser extension, such as a specific learning difficulty (SpLD), CVD, or visual impairment. Participants were also given the opportunity to provide any other feedback they had about the extension overall.

4. RESULTS

Results showed that overall, the ML model classified emails accurately, and test participants were content with the usability of the extension. Additionally, they found it simple to use and felt it educated them on the techniques commonly used in phishing emails.

4.1 Model accuracy

The model was trained with a sequence size of 500 words, using pre-padding and post-truncation to reach this standard size. The size of 500 words was chosen because the majority of emails in the dataset fell within this range. Overall, the dataset contained 11227 records.

The model produced: 5930 true positives (TPs), 5287 true negative (TNs), 0 false positives (FPs), and 10 false negatives (FNs). The true categories of emails were recorded when counting FNs to understand the model's accuracy when classifying categories of phishing emails (Table 2).

Table 2: Total number of emails in the dataset and FNs.

Email Category	Total	No. of FNs
Unexpected Money/Winnings (UNX)	*3581*	*0*
Extortion (EXT)	*1427*	*0*
Impersonation (IMP)	*541*	*7*
Business Email Compromise (BEC)	*391*	*3*

4.2 User acceptance

To evaluate the usability of the extension, 44 participants (23 male, 21 female) over the age of 18 years were recruited for the pilot study. Participants ranged in age from 18-69 (with 59% falling into the 18-24 bracket), and varied in level of education, field of study, and country of residence.

Overall, participants' answers to the SUS questionnaire gave the MailTrout extension a score of 87.5 out of 100. Notably, younger participants gave the extension a higher usability rating than older participants. Only one participant reported being in the age range of 40-54. The two highest age ranges (40-54 and 55-69) were combined into one range of 40-69 to aid in presenting and interpreting the results. Participants aged 18-24 gave the extension the highest usability score on average, while those aged 40-69 gave the extension the lowest usability score. While the ratings received overall were positive, these findings demonstrate that older participants may have found the tool less usable (Figure 4).

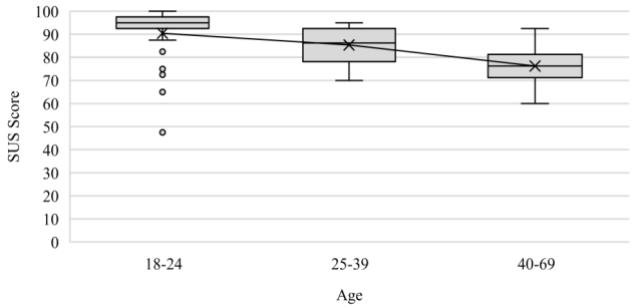


FIGURE 4: SUS SCORE BY AGE BRACKET.

Many participants remarked on how easy they found the extension to use and understand, describing it as refined and straightforward. Participants also found the speed and ability of the extension impressive. A common view among participants was that the extension was well designed, and the text was easy to read and understand.

Several participants expressed that the extension would be helpful to those who are less confident online and perhaps more vulnerable to phishing emails, such as the elderly. Another emerging theme was that participants said they would recommend MailTrout to people they knew who commonly receive phishing emails. Overall, participants felt they would be very likely to recommend the extension to someone looking to protect themselves against phishing emails, scoring their likelihood an average of 4.59 out of 5.

Before testing the extension, users were asked to rate their familiarity with terms describing the types of phishing emails. They were then given a description of each category and asked to rate the likelihood that they would be able to identify an email of each category. After using the extension, they were asked to rate the likelihood again, exploring if their level of knowledge increased. Results showed that participants knowledge of phishing emails and the associated categories improved post-test.

Some participants raised issues with the extension's functionality. Some reported that the result produced was more accurate if more text was selected for analysis. Therefore, users could receive less accurate results if they omitted some words when highlighting an email's text. Participants also expressed issues with interaction, notably the need to highlight text and click the extension icon. Others argued the extension often flagged emails as phishing where there were no typical characteristics of phishing attacks present in the text, such as requests for information or money or when the email appeared to have been sent by a trusted individual.

The results of the user testing were recorded to evaluate the accuracy of participants' classification of emails either as legitimate or phishing. These results are displayed as a confusion matrix – a table of the number of correct and incorrect predictions generated (Figure 5). Confusion matrices can be used to visualise accuracy and can include performance metrics. Using the FP and FN rates calculated, a confusion matrix was developed to present the participants' accuracy.

	Actual Positive 542	**Actual Negative** 125
Predicted Positive 525	**True Positive** *(TP)* 504	**False Positive** *(FP)* 21
Predicted Negative 142	**False Negative** *(FN)* 38	**True Negative** *(TN)* 104

FIGURE 5: CONFUSION MATRIX OF PARTICIPANT CLASSIFICATIONS.

The specific categories of phishing emails shown during experiments were also recorded to determine the number of phishing emails erroneously marked legitimate (false negatives), as shown in Table 3. These results showed that BEC emails had the largest number of FNs, suggesting they may have been the category detected with the least accuracy.

TABLE 3: *FNS PER CATEGORY DURING USER TESTING*

Email Category	**No. of Emails**	**No. of FNs**
Unexpected Money/Winnings (UNX)	*124*	*1*
Extortion (EXT)	*97*	*1*
Impersonation (IMP)	*143*	*5*
Business Email Compromise (BEC)	*178*	*31*

5. Discussion

This section discusses the accuracy and success of the ML model used by MailTrout and the usability of the integrated solution as a security tool.

5.1 Model

Using the FP and FN rates of PILFER and SpamAssassin as shown in Table 1 (Fette et al., 2007), categories were devised to rank the success of the ML model.

TABLE 4: *CATEGORIES OF ML MODEL FP AND FN RATES***.**

Category	False Positive (FP) Rates	False Negative (FN) Rates
'Excellent'	≤ 0.0012	≤ 0.036
'Good'	> 0.0012, ≤ 0.00135	> 0.036, ≤ 0.0715
'Average'	> 0.00135, ≤ 0.0022	> 0.0715, ≤ 0.13
'Poor'	> 0.0022	> 0.13

$$FP\ Rate = \frac{FP}{FP+TN} \qquad FN\ Rate = \frac{FN}{FN+TP}$$

FIGURE 6: FORMULAE FOR FALSE POSITIVE (FP) AND FALSE NEGATIVE (FN) RATES.

Using the formula shown in Figure 6, the FP of the MailTrout model was 0.0, with an FN of 0.00168, demonstrating the model's accuracy.

In comparison to existing research, the model outperformed other ML-based phishing detection methods. As shown in Table 4, the email classifier PILFER combined with a feature using the spam filter SpamAssassin developed by Fette et al. (2007) had an FP rate of 0.0013 and an FN rate of 0.036, while the trained SpamAssassin filter alone had an FP rate of 0.0012 and an FN rate of 0.130. The most accurate RNN phishing classifier developed by Halgaš et al. (2019) had FP and FN rates of 0.0126 and 0.0147, respectively.

However, these findings are somewhat limited by issues with the dataset. Firstly, due to the lack of data available, 80% of emails from the same dataset were used for training, with 20% for testing, following the Pareto Principle (McRay, 2015). Since the training and testing emails were from the same dataset, the model's familiarity may cause it to produce more seemingly accurate results than it would on unseen data. Models may learn the details of the training data with such specificity that they cannot make more general predictive rules that can be applied to new datasets, in an issue known as 'overfitting' (Dietterich, 1995). Due to the use of the same dataset for training and testing, the results of this study may suggest that the model is more accurate than it would be in a practical setting.

The dataset's quality may have been negatively impacted by the methods used to collect data or issues with the existing corpora. The Fraud Email Dataset (Verma, 2018) used as part of the training dataset contained some email metadata such as the date and time that emails were sent and encoded text for use with older email servers. This data was not valuable for training and may have caused the ML model to overfit or fail to identify words and phrases correctly. The dataset also had a lack of variety of legitimate emails; as the legitimate emails used all came from the released dataset of Hillary Clinton's emails (Kaggle, 2019), they may not have been reflective of the average email user's inbox.

When using the Python Reddit API Wrapper to extract comments from a Reddit thread of extortion emails (EugeneBYMCMB, 2019), some unrelated comments were extracted and added to the dataset. This was due to the thread containing general comments from users introducing or discussing the emails shared. Also, the OCR technology used to extract text from images of phishing emails may have produced inaccurate results due to an inability to understand stylised text or navigate unusual text layouts. The presence of text added to images to highlight common signs of phishing attacks may also have been picked up by OCR technology.

5.2 User acceptance

Considering the SUS adjective ratings proposed by Bangor et al. (2009), the SUS score of 87.5 given to the extension can be described as 'excellent', and highlights that the extension met its aim of being a usable security tool.

Participants reported that they found the extension easy to use and understand. One participant suggested that users would be more likely to keep using MailTrout due to the extension's accessibility and embedded nature.

"I like how easy it is to use, it's always in the corner so it isn't a complicated process that people will give up on easily"

Participants also remarked how impressed they were with the functionality of the extension.

"There are certain things in the tone of an email that I would not have flagged had it not been for the extension"

Commenting on the design and layout of the extension, one participant with strong colour vision deficiency (CVD) reported the colour scheme provided a high level of contrast and therefore had no issues using it. Other participants with specific learning difficulties (SpLDs) found the extension accessibly designed with a simple layout, colour-coding, and succinct information.

"I am extremely colour blind (strong deuteranopia) and had absolutely no issues using the web extension and found each colour to clearly stand out from its surroundings"

"I have dyslexia which makes using some text-based extensions difficult, this extension and the colour coded nature of the help box layout made it very accessible to use. Additionally the lists of what to look out for were to the point and easy to understand"

Participants also suggested that the extension could educate people on identifying phishing emails themselves, reporting that the information on what to look out for, what to do and what not to do was a particularly good feature.

"The Look Out/Do/Don't is a really good feature, as the user is learning as they use [the extension] rather than just relying on a traffic light system."

The responses to each statement in the SUS survey were very positive overall, generally averaging between 'Agree' (4) and 'Strongly Agree' (5). However, the average response to the first statement was found to be lower than that of all others. The first statement read *"I think that I would like to use this extension frequently*".

A possible explanation is that while participants provided positive feedback on the extension overall, they did not feel that they needed it themselves due to their ability to identify phishing emails unaided. This can be understood further using the theory of diffusion of innovations (DOI), which explores how new ideas and technologies are adopted. One of the characteristics of an innovation is its 'relative advantage', meaning the degree to which the innovation is perceived as better in comparison to existing measures. If a user perceives the relative advantage of an innovation as low, they will be less likely to adopt it (Rogers, 2003). Therefore, if participants believed they were able to identify phishing emails themselves with high levels of accuracy, they may have felt the relative advantage of using the extension was low. Therefore, they felt less likely to adopt the extension.

Participants' ratings of their ability to spot phishing emails (where 1 is poor and 5 is excellent) were compared to their answers for SUS statement 1 to understand this finding further. As hypothesised, participants who answered that they would not use the extension frequently reported they had a strong ability to spot phishing emails, suggesting that they would find using the extension unnecessary.

Another characteristic of innovation is 'observability', meaning the degree to which the effects of the innovation are visible. If a user perceives the visible results of innovation as low, they will be less likely to adopt it (Rogers, 2003). In DOI theory, 'preventative innovations' aim to lower the probability of an unwanted future event. Preventative innovations tend to take longer to be adopted by users due to the lack of observable impact of their use. However, if a user experiences a 'cue-to-action' – an event that causes them to undergo a behavioural change – then this can result in a more favourable attitude towards an innovation (Xiao, et al., 2014).

Security tools such as MailTrout may be considered preventative innovations as they aim to lower the probability of security failures, such as a user falling victim to phishing emails. Therefore, users may be more

likely to adopt the extension if they have experienced a cue-to-action, such as becoming the victim of a phishing attack.

Participants were shown to have an increased knowledge of types of phishing email after using the extension. Average familiarity ratings for each phishing email category increased as participants used the extension to learn what to look for in phishing emails. These findings demonstrate the extension's potential ability to educate users about identifying phishing emails in the long term.

The results also demonstrated a correlation between a participant's SUS rating and their demographic characteristics. Firstly, participants studying or working in a formal sciences subject, such as computing science, found the extension more usable than those in other subject areas.

A potential explanation for this result may be that those employed in formal science fields are more frequent users of computers and are therefore more comfortable learning how to use new tools. Participants in formal sciences may have had more experience, specifically with web browser extensions and would find learning to use the extension far less challenging than someone who has never used a browser extension before. They may also have had more exposure to phishing emails, especially if they are involved in cybersecurity, which may also have given them an advantage over users who are less familiar with the terms and techniques often associated with email phishing. Participants in formal sciences demonstrated an overall higher familiarity with categories of phishing emails than those in other fields throughout the experiment. However, it is important to note that the average SUS scores of each subject field were all in the 'excellent' category. Hence, differences between subject fields are a minor concern.

Younger participants found the extension more usable than older participants. The average SUS scores of the 18-24 and 25-39 age ranges were in the 'excellent' category, while that of the 40-69 range was in the 'good' category. While these ratings are positive, it demonstrates that older participants found the tool less usable. A potential reason for this may be that participants who were born after the 1980s – commonly referred to as 'digital natives' – are more likely to have grown up around digital technology and so have been familiar with computers from an early age. Conversely, users born before the 1980s – commonly referred to as 'digital immigrants' – grew up before the widespread use of digital technology and have not had the same experience, thus making it harder for them to learn how to use new technologies (Prensky, 2001). Younger participants may also have had more experience using browser extensions and dealing with phishing emails - this group demonstrated an overall higher familiarity with categories of phishing emails throughout the experiment than older participants. However, consideration should be given to the limited number of older participants who took part in the research, thus further work is required to explore this result.

6. Conclusion and future work

The research showed that due to the presence of common words and sentiment patterns across phishing emails, and the ability of ML algorithms to classify data by detecting recurring patterns, ML technology is well-suited to the task of identifying phishing emails. Additionally, the web browser extension format provided a suitable way to create an embedded learning tool, providing users with an opportunity to use the extension while completing everyday tasks, such as checking their emails. The extension demonstrated high levels of usability and accuracy when detecting phishing emails.

These findings also indicate that browser extensions can act as accessible security tools, requiring limited technical knowledge to use and can easily be incorporated within a person's routine online activities. Due to their simplicity and embedded nature, browser extensions may be beneficial for those with less experience of using the internet.

Furthermore, the COVID-19 pandemic has caused a fundamental shift in our lives, heightening the pace at which society adopts and utilises digital technologies. The increased reliance on digital communications means that more people may be likely to encounter phishing emails. Thus, more research is required in this field to protect potentially vulnerable citizens.

7. References

Australian Competition & Consumer Commission. (2015) Types of scams. https://www.scamwatch.gov.au/types-of-scams (Retrieved 13 March 2019).

Aggarwal, A., Rajadesingan, A., and Kumaraguru, P. (2012) PhishAri: automatic realtime phishing detection on twitter. 2012 eCrime Researchers Summit, Las Croabas, PR, USA, 23-24 October 2012, pp. 1-12. IEEE.

Avanan. (2019) Global Phish Report. https://www.avanan.com/global-phish-report (Retrieved 7 February 2020).

Bangor, A., Kortum, P. & Miller, J. (2009). Determining What Individual SUS Scores Mean: Adding an Adjective Rating Scale. *Journal of Usability Studies,* 4(3), pp. 114-123.

Bansal, S. & Aggarwal, C. (2020) *textstat 0.6.0.* https://github.com/shivam5992/textstat (Retrieved 3 March 2020).

Begeny, J. C. & Greene, D. J. (2014). Can Readability Formulas Be Used to Successfully Gauge Difficulty of Reading Materials?. *Psychology in the Schools,* 51(2), pp. 198-215.

Bird, S., Klein, E. & Loper, E. (2009). 2.4.1 Wordlist Corpora. In: J. Steele, 1st ed. *Natural Language Processing with Python.* Sebastopol: O'Reilly Media, Inc., pp. 60-62.

Chen, Y.Y., Lin, Y.H., Kung, C.C., Chung, M.H., Yen, I. (2019). Design and Implementation of Cloud Analytics-Assisted Smart Power Meters Considering Advanced Artificial Intelligence as Edge Analytics in DemandSide Management for Smart Homes. *Sensors (Basel),* 19(9).

Chollet, F. (2015) *Keras Documentation.* https://keras.io/ (Retrieved 14 March 2020).

Dietterich, T. (1995) Overfitting and undercomputing in machine learning. *ACM computing surveys (CSUR),* 27(3), pp. 326-327.

Dixon, M., Gamagedara Arachchilage, N.A. and Nicholson, J. (2019) Engaging users with educational games: The case of phishing. In Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems (pp. 1-6).

Dunlop, M., Groat, S. and Shelly, D. (2010) Goldphish: using images for content-based phishing analysis. 2010 Fifth international conference on internet monitoring and protection, Barcelona, Spain, 9-15 May 2010, pp. 123-128. IEEE.

Dwarampudi, M. & Reddy, N. V. S. (2019) Effects of padding on LSTMs and CNNs. https://arxiv.org/pdf/1903.07288.pdf (Retrieved 5 February 2020).

EugeneBYMCMB. (2019) The Blackmail Email Scam (part 3) : Scams. https://www.reddit.com/r/Scams/comments/biv65o/the_blackmail_email_scam_part_3/ (Retrieved 21 January 2020).

Fette, I., Sadeh, N. & Tomasic, A. (2007) Learning to detect phishing emails. 16th International World Wide Web Conference, Banff, Canada, May 2007, pp. 649–656. ACM.

Fu, A. Y., Wenyin, L. & Deng, X. (2006). Detecting Phishing Web Pages with Visual Similarity Assessment Based on Earth Mover's Distance (EMD). *IEEE Transactions on Dependable and Secure Computing,* 3(4), pp. 301-311.

Gers, F. A. & Schmidhuber, E. (2001). LSTM recurrent networks learn simple context-free and context-sensitive languages. *IEEE Transactions on Neural Networks,* 12(6), pp. 13331340.

Goodfellow, I., Bengio, Y. & Courville, A. (2016). 6.2.2.3 Softmax Units for Multinoulli Output Distributions. In: *Deep Learning.* Cambridge: MIT Press, pp. 180-184.

Google. (2020a) Tensorflow. https://www.tensorflow.org/ (Retrieved 15 January 2020).

Google. (2020b). tesseract-ocr / tesseract: Tesseract Open Source OCR Engine (main repository). https://github.com/tesseract-ocr/tesseract (Retrieved 6 February 2020).

Google. (2020c). *tf.keras.preprocessing.text .Tokenizer.* https://www.tensorflow.org/api_docs/python/tf/keras/preprocessing/text/Tokenizer (Retrieved 23 March 2020).

Graham, A. (2018) The cost of a cyber attack. IT Governance. https://www.itgovernance.co.uk/blog/the-cost-of-a-cyber-attack (Retrieved 7 February 2020).

Halgaš, L., Agrafiotis, I. & Nurse, J. R. C. (2019). *Catching the Phish: Detecting Phishing Attacks using Recurrent Neural Networks (RNNs).* Jeju Island: 20th World Conference on Information Security Applications, Springer.

Horgan, S., Collier, B., Jones, R., Shepherd, L. (2021) Re-territorialising the policing of cybercrime in the post-COVID-19 era: towards a new vision of local democratic cyber policing. Journal of Criminal Psychology, *Accepted/In Press*.

Kaggle. (2019) Hillary Clinton's Emails. https://www.kaggle.com/kaggle/hillary-clinton-emails/ (Retrieved 15 March 2020).

Kumaraguru, P., Sheng, S., Acquisti, A., Cranor, L.F., Hong, J. (2010). Teaching Johnny not to fall for phish. *ACM Transactions on Internet Technology (TOIT),* 10(2), pp. 1-31.

Lai, S., Xu, L., Liu, K. & Zhao, J., (2015). Recurrent convolutional neural networks for text classification. *Proceedings of the National Conference on Artificial Intelligence,* Volume 3, pp. 2267-2273.

Lallie, H. S., Shepherd, L. A., Nurse, J. R. C., Erola, A., Epiphaniou, G., Maple, C., & Bellekens, X. (2021) Cyber security in the age of COVID-19: a timeline and analysis of cyber-crime and cyber-attacks during the pandemic. Computers & Security, 105. 102248.

Lee, M. (2020) madmaze/pytesseract: A Python wrapper for Google Tesseract https://github.com/madmaze/pytesseract (Retrieved 6 February 2020).

McRay, J., 1st ed., (2015). Pareto principle. In: *Leadership glossary: Essential terms for the 21st century.* Santa Barbara: Mission Bell Media.

Postolache, F. & Postolache, M. (2010). Current and Ongoing Internet Crime Tendencies and Techniques. Preventive Legislation Measures in Romania. *EIRP Proceedings,* 5(1), pp. 35-43.

Prensky, M., (2001). Digital Natives, Digital Immigrants. *On the Horizon,* 9(5), pp. 1-6.

Prusa, J., Khoshgoftaar, T. M. & Seliya, N. (2015). *The Effect of Dataset Size on Training Tweet Sentiment Classifiers.* Miami: IEEE 14th International Conference on Machine Learning and Applications (ICML), IEEE.

Radev, D. (2008). CLAIR collection of fraud email, ACL Data and Code Repository, ADCR2008T001.https://aclweb.org/aclwiki/CLAIR_collection_of_fraud_email_(Repository) (Retrieved 15 February 2020).

Rogers, E. M. (2003). *Diffusion of Innovations.* 5th ed. New York City: Simon and Schuster.

Sak, H., Senior, A. & Beaufays, F. (2014). *Long Short-Term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling.* 15th Annual Conference of the International Speech Communication Association, Singapore, ISCA Archive.

Sauro, J. & Lewis, J. (2011). When designing usability questionnaires, does it hurt to be positive?. *Proceedings of the SIGCHI Conference on human factors in computing systems*, 7 May, pp. 2215-2224.

StatCounter. (2021) Browser market share worldwide. https://gs.statcounter.com/browser-market-share (Retrieved 6 May 2021).

Tao, J. & Fang, X. (2020). Toward multi-label sentiment analysis: a transfer learning based approach. *Journal of Big Data,* 7(1), pp. 1-26.

UN General Assembly. (1989). Convention on the Rights of the Child. *United Nations, Treaty Series,* Volume 1577, p. 3.

Vasa, H. (2019) Google Images Download. https://github.com/hardikvasa/google-images-download (Retrieved 15 March 2020).

Verma, A. (2018). Fraud Email Dataset | Kaggle. https://www.kaggle.com/llabhishekll/fraud-email-dataset (Retrieved 28 January 2020).

Wang, P., Qian, Y., Soong, F.K., He, L., Zhao, H. (2015). Part-of-Speech Tagging with Bidirectional Long Short-Term Memory Recurrent Neural Network. *ArXiv [Preprint].* (Retrieved 15 March 2020).

Wickline, M. (2001) Coblis - Color Blindness Simulator. https://www.color-blindness.com/coblis-color-blindness-simulator/ (Retrieved 10 March 2020).

Xiao, S., Witschey, J. & Murphy-Hill, E., (2014). Social Influences on Secure Development Tool Adoption: Why Security Tools Spread. *Proceedings of the 17th ACM conference on Computer supported cooperative work & social computing,* pp. 1095-1106.

Designing for affective warnings & cautions to protect against online misinformation threats

Fiona Carroll and Bastian Bonkel

Cardiff School of Technologies, Cardiff Met University, Llandaff Campus, Western Avenue, Cardiff CF5 2YB, Wales

fcarroll@cardiffmet.ac.uk, bbonkel2@cardiffmet. ac.uk

Social media's affordance for misinformation is compromising the glue that holds us and our society together. By influencing and manipulating our human behaviour particularly the decisions we make and opinions we form, it is polarising our existence in not only the virtual but also the physical world in which we live. Yet, despite being aware of the destructive nature of misinformation in general, many of us still don't seem to understand/ see the full danger on an individual basis. Hence, as we have witnessed during Covid 19, many people still continue to share this misinformation widely. The authors of this paper feel that there is an urgent need to support people in being more aware of false information whilst online. In this paper, we share thoughts around some of the mechanisms that people currently use to identify misinformation online. In particular, the focus is on a study that explores participant's experiences of ten different visualisation effects on a Facebook page. The findings highlight that some of these initial visualisation designs are more effective than the others in informing people that something is not quite what it should be. Like in the physical world, we propose the design of a set of affective online visual warnings and cautions that we hope can be further developed to fight online misinformation and counter it's current negative influence on society.

Misinformation, Warning, Caution, Affective, Visualisation effects, Awareness, Perception.

1. INTRODUCTION

Many countries around the world have spent years trying to build up a socially cohesive society. A society that 'works towards the well-being of all its members, fights exclusion and marginalisation, creates a sense of belonging, promotes trust and offers its members the opportunity of upward mobility' (Oecd 2012, p. 17). However, it seems that the Internet and particularly social media have very quickly started to erode this effort. Moreover, social media's affordance for online misinformation is compromising the glue that holds us and our society together. For example, the spread of misinformation during the coronavirus outbreak was rapid and caused huge uncertainty and tensions amongst people. So much so that the British Computing Society (BCS 2020) in their article '11 ways to fight Coronavirus misinformation' advised that bad spelling is a strong signal of misinformation. However, using grammar and spelling as an indicator of misinformation is becoming less and less useful. As research shows 'digital misinformation thrives on an assortment of cognitive, social, and algorithmic biases and current countermeasures based on journalistic corrections do not seem to scale up' (Ciampaglia 2018, p.147). In reality misinformation that has bad grammar and spelling is likely to increase people's vulnerability to the more sophisticated misinformation attempts. In this paper, we share thoughts around some of the mechanisms that people currently use to identify misinformation online. In particular, the focus is on participant's experiences of ten different visualisation effects on a Facebook page. The aim is to support people in taking more notice of potential misinformation threats. The following sections explore how we might enable people (through visual supports- warnings and cautions) to make the right decisions to counteract the spread of misinformation.

2. WHAT IS TRUE AND WHAT IS FALSE?

Our lives today are inextricably tied to the Internet and from this the acquisition of data. While this is empowering many of us, it is also proving to be very harmful especially as it is now more difficult than ever to decipher what is true and what is false in all this data. As Zhou and Zhang (2007, p.1) describe misinformation is the 'transmission of distortions or falsehoods to the audience'. It is distinct from disinformation where false information is spread with the intent to harm, misinformation is the unintentional spread of false information. Needless to say, both have become such a common part of our digital media environments that it is compromising the ability of our societies to form informed opinions (Fernandez and Alani 2018). Furthermore, it is people's emotions that has become the driving force for much of the widespread of misinformation. As a result it is becoming more and more difficult to centrally control. In detail, content that evokes high arousal positive (awe) or negative (anger or anxiety) emotions is more viral (Berger and Milkman 2012). The authors of this research are interested in the

emotional hook of misinformation. In particular, how we can engage the affective through designs to alert (i.e. warn and/ or caution) against misinformation.

3. MISINFORMATION, TRUST AND EMOTION

Trust and distrust have been considered as polar opposite constructs (Mal et al. 2018). Trust is the 'willingness to take a risk' and the level of trust is an indication of the amount of risk that one is willing to take (Mayer et al. 1995, p.1). Trusting is the inclination of a person 'A' to believe that other persons 'B' who are involved with a certain action will cooperate for A's benefit and will not take advantage of A if an opportunity to do so arises (Ben-Ner and Halldorsson 2010). In their paper, Schul et al. (2008) see the state of trust as being associated with a feeling of safety, they assume that a state of distrust is the mental system's signal that the environment is not normal, things may not be as they appear. Hence, individuals sense they should be on guard/ careful. If the environment is as it normally is and things really are as they appear to be, then the individuals see no reason to refrain from doing what they routinely do (Schul et al. 2008).

In terms of the affective, Martel et al. (2020) found both correlational and causal evidence that reliance on emotion increases belief in fake news. Furthermore, Greenstein and Franklin (2020, p.1) found the suggestibility for false details increased with anger. In attempt to counter this, the authors of this paper aim to use emotions to alert people to misinformation. As Kaiser et al. (2020) highlights, disinformation warnings can – when designed well – help users identify and avoid disinformation. Moreover, Bhuiyan et al. (2018) developed 'FeedReflect' which is a browser extension that nudges users to pay more attention. It uses reflective questions to engage people in news credibility assessment on Twitter. Other research (Lutzke et al. 2019) highlights the potential of simple interventions to prime critical thinking and slow the spread of fake news on social media platforms. As Fazio (2020, p.1) aptly states, it is about 'adding "friction" (i.e. pausing to think) before sharing can improve the quality of information shared on social media'. Supporting that, Pennycook et al. (2020) present results that show how simple and subtle reminders may be sufficient to improve people's sharing decisions regarding information about COVID-19. Therefore improving the accuracy of the information about COVID-19 on social media.

4. STUDY

This study took place at Cardiff Met University in July 2020. Its aim is to give some insight into individuals' perception of misinformation. In particular, to probe participant's experiences of ten different visualisation effects on a Facebook page in order to determine which afforded the most effective alert to the threat of misinformation.

4.1. Participants

Five hundred and thirty-two participants from the ages of 18 to 74 years completed the study. These included two hundred and seventy females and two hundred and sixty-two males. The majority of participants were from the age range 35-44 years old (one hundred and twenty-five participants). Also most participants (one hundred and sixty-one females and two hundred and four males) were 'employed for wages'. Others included homemakers, students, retired, self employed, out of work and looking for work, out of work but not looking for work, those unable to work, military and other. All participants (over eighteen years old and internet users) were globally recruited through the Dynata Insights Platform.

4.2. Methods & Procedure

The study consisted of four main parts. The first part was to probe participants around the concept of misinformation. To avoid priming, we asked participants if they thought it is easy to identify 'something' online that is not quite right (i.e. not quite as it should be)? The second part of the study was focused on gathering data on participant's thoughts and feelings on an image of an authentic Facebook page rendered ten times with a different visualisation effect (see fig.1). On each image, the visualisation effect was randomly applied to one of the three Facebook posts on the page. These ten effects (see fig.1) were based on designs from earlier studies (Carroll et al. 2018), (Carroll et al. 2020).

These included the different use of colour to **block**, **highlight** and **censor** the text on the Facebook post. They also included different explorations of the visual acuity of the text on the Facebook post: **blur**, **convolve**, **erode**, **fog**, **noise** and **wishy**. Finally, a more literal representation of a threat through broken **glass** over the text on the Facebook post was also investigated. The emphasis of the third part of the study was on which visualisation effect was the most effective in making participants more aware that something is not quite as it should be. Finally the last part of the study was interested in probing participant's opinions of what they think needs happen with regards to protecting themselves against misinformation threats online. The study took approximately 20-30 minutes in duration. It was conducted using the Qualtrics online survey software and open-ended questionnaire

questions were used to collect the data. The Ethics Board of Cardiff Met University approved the study methods and procedure and all participants provided online consent for study completion. The following presents a qualitative analysis of the online survey data.

4.3. Data Analysis & Results

For the first part of the study and in particular, the question: *In your opinion, do you think it is easy to identify 'something' online that is not quite right (i.e. not quite as it should be)? Please elaborate how you would best identify it*, we have applied six phases of thematic analysis (Braun and Clarke 2006). An initial read of the data generated codes such as 'yes; no; true; can; web; sure; hard; source; details; online; site; scams; questions; sense; research; grammar; easy; good; email; check; poor and new'. Building on these codes, themes such as *gut instinct, spelling and grammar, research, review, appearance, source (URL, website, email, padlock), experience of user, too good to be true, expectations, no/ not sure, yes, random* and *didn't understand question*, started to emerge and then time was taken to gather all data relevant to each potential theme. Finally, after a period of reviewing and refinement was undertaken, the following themes were determined to best demonstrate how participants decipher when something is not right online:

- **Intuition**: gut feeling usually makes me feel when something online isn't genuine Participant 36.
- **Appearance**: No, it's not that easy, some scams are very sophisticated. Bad spelling or grammar can sometimes be a giveaway, also asking for info a reputable company wouldn't request. Participant 98.
- **Reviews and Research**: I would look at reviews and research everything from different sites first then match the description up. Participant 27.
- **Source and Security**: In my opinion it is relatively easy to identify whether 'something' online is not quite right. There are ways to check the authenticity of certain websites and web pages such as anti-virus tracking software. Web browser address bars indicate whether websites or web pages could be trusted or not by symbols signifying whether they could be trusted such as the padlock. Participant 290.
- **User knowledge**: Yes. I have little trouble spotting these things, but I have been using the Internet for many years and am naturally sceptical. Participant 52.
- **Exceeded expectations**: If an offer seems too good to be true or if the advert does not seem professional. Participant 381.
- **Unrealistic demands**: Asking for personal information when it's literally not needed. Participant 97.

Interested to probe this further, it was clear from the data that the appearance of the content and the digital interface design plays an important role in helping one hundred and seventeen test participants to decipher that something was amiss. This theme of appearance included bad spelling and grammar which featured amongst seventy participants individual comments as a strong indicator of misinformation. Furthermore the parallels between these themes and cyber security awareness is important to highlight (especially, when cyber attacks can include various degrees of misinformation).

Part 2 of the study focused on the appearance of the Facebook posts and in particular, the ten visualisation effects (See fig.1). In detail, we asked participants to describe each visualisation effect/ alternation and its possible effect? It is interesting to see that words like danger, red, attention, warning, highlighted and grabbing are used to describe the **highlight** visualisation effect. Similarly, words like unsafe, warning, red, attention, alarming, danger are also used to describe the **block** visualisation effect. Whilst words like blurred, blurry, fuzzy, suspicious, confusing and ignore are used to described the **blur** visualisation effect and the **glass** effect is being described with words such as broken, cracked, smashed, confusing and annoying.

Moreover when we asked the question about which visualisation effect made them question the validity of what they were reading. The **blur** visualisation effect was more effective for women whilst the **block** visualisation effect

Figure 1: *Ten different visualisation effects*

for men. When probed about which visualisation effect made them feel uncomfortable, it is clear from the data that the **fog** visualisation (female 51% and male 31%) was the effect that most found uncomfortable. The majority of participant felt that they would ignore the **convolve** effect because it didn't really captivate or interest them. The **block** visualisation effect was of most interest to participants (29% and 22%); it was the visualisation that made them want to know more about why it was altered.

When asked if they felt nervous or calm looking at these visualisation effects, most participants (one hundred and thirty two participants) felt that the **block** and then **glass** (one hundred and thirty one participants) made them more nervous. When asked which effect made them more alert, the **block** visualisation (one hundred and thirty four participants) showed the highest number of participants that felt it made them more alert (see table 1).

Table 1: *Summary of semantic data captured from Block visualisation effect*

Semantics (Block)	1	2	3	4	5	Mean	Mode
Nervous (1) to Calm (5)	35.06%	23.17%	25.91%	9.45%	6.40%	2.289634	1
Relaxed (1) to Worried (5)	1.31%	4.10%	22.40%	32.81%	39.38%	3.716463	4
Attentive (1) to Inattentive (5)	14.52%	22.85%	36.69%	14.52%	11.42%	2.268293	1
Unaware (1) to Alert (5)	0.15%	1.03%	14.33%	35.27%	49.23%	4.14939	5
Confident (1) to Not Confident (5)	2.38%	5.11%	33.33%	20.81%	38.36%	3.457317	3

Part 3 of the study was primarily concerned with examining which of the ten visualisations effects was most successful in alerting/ making the participate more aware. In detail, we asked participants to rank each visualisation effect in order of which one makes them most aware that something is not quite as it should be? (1 [top] = Most aware and 10 [bottom] = Least aware). The **block** visualisation effect featured the most ranked at 1.

Finally, for part four of the study, participants were asked what they felt needed to happen online for them (and people in general) to care more about the validity and safety of the online experience. The findings strongly show that people need to be made more aware of what is happening. The frequency of words such as warnings (32 times), alerts (21 times), checks (9 times), messages (8 times) highlight that participants feel they need the support to become more aware.

5. CONCLUSION

In conclusion, we feel that there is currently a lack of support for people to identify a misinformation threat in the online environment. In the physical world we are provided with a range of techniques to enable us to determine whether something needs to be fully avoided or simply to take heed with. In the online environment, we don't have a set of standards or laws detailing what symbols /signs/ effects that determine what is dangerous or what might afford or connote careful and attentive behaviour.

Moreover, we feel that knowing the difference between the online warning and caution is essential for further online interactions. As an end user, we need to be able to perceive and understand that a caution online indicates a minor risk to ones person if proper safety practices aren't observed. Whilst also, to understand that a warning online is an alert to significant dangers. As this study has started to show, certain visualisation effects can trigger

certain feelings around online information. Also, in parallel, people seem to be naturally examining the presentation of their online environments as a means to detect if something is not quite as it should be. This research aims to support this behaviour further by providing end users with a more effective means to identify when something is lacking in integrity online. This particular study is the first in a series of studies to explore the development of effective online warnings and cautions. Similar to the physical world, the aim is to provide people with a system of warnings and cautions to protect them against online threats (including misinformation).

ACKNOWLEDGEMENTS

This research was funded by the Welsh Crucible, a consortium of Welsh higher education institutions and the Higher Education Funding Council for Wales (HEFCW). We are very grateful to Dr James Kolasinski, Cubric, Cardiff University who was a collaborator on this research project.

REFERENCES

BCS (2020). 11 ways to fight Coronavirus misinformation — BCS.

Ben-Ner, A. and F. Halldorsson (2010). Trusting and trustworthiness: What are they, how to measure them, and what affects them. *Journal of Economic Psychology*.

Berger, J. and K. L. Milkman (2012). What makes online content viral? *Journal of Marketing Research*.

Bhuiyan, M. M., K. Vick, T. Mitra, K. Zhang, and M. A. Horning (2018). FeedReflect: A tool for nudging users to assess news credibility on twitter. In *Proceedings of the ACM Conference on Computer Supported Cooperative Work, CSCW*.

Braun, V. and V. Clarke (2006). Using thematic analysis in psychology. *Qualitative Research in Psychology*.

Carroll, F., M. Webb, and S. Cropper (2018). Losing our senses online: Investigating how aesthetics might be used to ground people in cyberspace. *IEEE Technology and Society Magazine*.

Carroll, F., M. Webb, and S. Cropper (2020, sep). Investigating aesthetics to afford more 'felt' knowledge and 'meaningful' navigation interface designs. In *2020 24th International Conference Information Visualisation (IV)*, pp. 214–219. IEEE.

Ciampaglia, G. L. (2018). Fighting fake news: a role for computational social science in the fight against digital misinformation. *Journal of Computational Social Science*.

Fazio, L. (2020). Pausing to consider why a headline is true or false can help reduce the sharing of false news. *Harvard Kennedy School Misinformation Review*.

Fernandez, M. and H. Alani (2018). Online Misinformation: Challenges and Future Directions. In *The Web Conference 2018 - Companion of the World Wide Web Conference, WWW 2018*.

Greenstein, M. and N. Franklin (2020). Anger Increases Susceptibility to Misinformation. *Experimental Psychology*.

Kaiser, B., J. Wei, J. N. Matias, E. Lucherini, J. Mayer, and K. Lee (2020). Adapting Security Warnings to Counter Online Disinformation.

Lutzke, L., C. Drummond, P. Slovic, and J. Arvai´ (2019). Priming critical thinking: Simple interventions limit the influence of fake news about climate change on Facebook. *Global Environmental Change*.

Mal, C. I., G. Davies, and A. Diers-Lawson (2018). Through the looking glass: The factors that influence consumer trust and distrust in brands. *Psychology and Marketing*.

Martel, C., G. Pennycook, and D. G. Rand (2020). Reliance on emotion promotes belief in fake news. *Cognitive Research: Principles and Implications*.

Mayer, R. C., J. H. Davis, and F. D. Schoorman (1995). AN INTEGRATIVE MODEL OF ORGANIZATIONAL TRUST. *Academy of Management Review*.

Oecd (2012). Perspectives on Global Development 2012 Social Cohesion in a Shifting World Executive summary. *Development*.

Pennycook, G., J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand (2020). Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention. *Psychological Science*.

Schul, Y., R. Mayo, and E. Burnstein (2008). The value of distrust. *Journal of Experimental Social Psychology*.

Zhou, L. and D. Zhang (2007). An ontologysupported misinformation model: Toward a digital misinformation library. *IEEE Transactions on Systems, Man, and Cybernetics Part A Systems and Humans*.

Development of Usable Security Heuristics for Fintech

Stephen Ambore, Huseyin Dogan and Edward Apeh

Bournemouth University
Poole, Dorset, UK
sambore@bournemouth.ac.uk, hdogan@bournemouth.ac.uk, eapeh@bournemouth.ac.uk

Investments in cybersecurity over the years have led to the availability of strong technical countermeasures and innovations that are being increasingly leveraged to strengthen the security posture of financial services systems. The effort to improve the security posture of the human element of financial services systems has not matched the effort in developing technical countermeasures, thereby undoing the gains of the later. One area where such problem exist is in Fintech where emphasis is placed on developing innovative and secured technical financial models aimed at making financial services more accessible through the mobile phone. These Fintech solutions however have shortcomings in securing the human element. This study seeks to address this problem through the development of heuristics that can be applied in the evaluation or design of Usable Security in Fintech. This study developed twelve (12) initial Usable Security heuristics which were validated through expert review. The heuristics were developed through an iterative approach that comprises a survey of Fintech users, semi-structured interviews of Fintech solution providers and thematic analysis of relevant literature. The findings of the study show that application of the developed heuristic provides for Usable Security.

Usable Security. Fintech. Heuristics. Cybersecurity. Usability

1. Introduction

The high rate of mobile penetration, capability to generate insight from user data and the need for a better and personalized user experience in the use of financial services is driving the uptake of innovative models for financial services otherwise known as Fintech. Fintech refers to innovative models that enable the delivery of financial services in an agile manner (Mani, 2019, Addilah, 2019, Saksonova and Kuzmina-Merlino, 2017). These models leverage technologies like Application Programming Interface (API) Blockchain Technology, Biometry Technology, Artificial Intelligence, Data Analytics and Cloud to provide financial services to existing and new customer segment (EFInA, 2020).

In the UK, Fintech through challenger banks and neobanks like Monzo and Revolult are disrupting the financial services landscape (High, 2021). In response to this disruption, most incumbent banks now offer Fintech solutions to their customers through mobile financial services.

While Fintech has facilitated access to financial services in a cost-effective way, it comes with a secondary risk of cybersecurity to the customers. Fintech and digital platform provided a window for consumers to access financial services remotely during the lockdown occasioned by the COVID-19 pandemic in most countries where physical access to banks and stores were restricted. However, cyber fraud targeting Fintech increased during the same period (Glenny, 2021, Borrett, 2021).

Strong technological countermeasures like strong cryptographic algorithm, biometric authentication and improve methods to elicit informed consent exist to curb the growth of cybercrime. These technical countermeasures and innovation notwithstanding, cybercrime incidences still occur (Shetty, 2018). Most of these have been attributed to the human element who has been described as the “weakest link” in the security value chain because of their propensity to make errors or poor security decisions in the use of a system (Sasse et al., 2001, Pfleeger et al., 2014). Irrespective of the security controls put in place, the action or inaction of end-users can make a system susceptible to cyber-attacks. Analysing the psychological perceptions on why users make unsafe security decisions, West et al. (2009) posited that errors by end-users in the use of a system, and not sufficiently addressing human factor considerations during design are major contributors to cybersecurity risks. While investment in technical controls would help mitigate the risk of cybercrime, mitigating the vulnerability associated with the “weakest link” is imperative to build security controls that do not discourage good use practice and further jeopardize security objectives. For instance, Hof (2015) argued that though technology controls exist to secure systems, they might not be designed with usability as a primary objective.

Security systems are not foolproof but strengthening the human element will further improve the security posture of Fintech. Fintech is an important innovation as it stands to provide access to financial services to over 1.7 billion people globally who currently do not have access to financial services (Asli et al., 2018). More so, as most banks continue to leverage the mobile phone to provide financial services, more customers will depend on Fintech to access service putting more customers and their transaction at risk of cybercrime.

This study adopts a sociotechnical approach to improve the cybersecurity posture of Fintech, by examining elements that can improve the human factor from the perspective of users and solution providers in the

ecosystem. The study examined previous approaches adopted to improve cybersecurity from human perspectives and identified the need to develop heuristics that can be applied to improve Usable Security in Fintech, using a case study of mobile financial services. Heuristics are rules of thumb, for making inferences in an environment with limited time, knowledge or computational power (Hafenbrädl, et al., 2016).

While previous studies have developed heuristics, which were tested in other domains, to the best of our knowledge, no previous work exist on developing heuristics that will help evaluate and design Usable Security in Fintech (Feth and Polst, 2019).

Furthermore, early usability studies focused on improving user experience through usability inspection with a view to identifying usability problems (Nielsen, 1992). This study examined the Usable Security of Fintech from the perspective of the users and solution providers in the ecosystem.

The study presents an initial set of heuristics for the evaluation of Usable Security in Fintech. The recommendations of this study would serve as a guide for Fintech Developers, Systems Auditors, HCI and Cybersecurity experts looking to improve the security posture of Fintech solution.

The next section of this paper examines related work while section three (3) provides an overview of the methodology adopted in this study. Results of the studies conducted are provided in section four (4). A discussion on the findings and recommendation of the studies are contained in section five (5). The paper ends with a recommendation for future studies in section six (6).

2. background and related work

In this section, we reviewed prior work on improving the cybersecurity posture of systems with a focus on the human element. We then reviewed how Usable Security is evaluated and how it is incorporated into systems design. Furthermore, we reviewed various approaches adopted by previous studies in developing heuristics with a view to adapting the most appropriate approach into the study.

2.1 Improving Cybersecurity Posture Usable Security

Research efforts to strengthen the human element in the cybersecurity and HCI domain have focused on improving system usability as a means of improving the cybersecurity posture of systems. While early usability research focused on improving usability for users, some studies on improving the usability of security mechanism for Developers and Systems Administrators have been published (Nielsen, 92, Zurko and Simon,1996, Adams and Sasse, 1999, Wijayarathna, and Arachchilage, 2019).

The Mobile phone interface provides customers access to Fintech solution. Mobile Phone Operating System (OS) developers such as Microsoft, Android and Apple have published user interface design guides to facilitate the usability of applications that run on mobile phones (Android, 2018, Apple, 2018). Rule of thumb; otherwise known as the 10 heuristics for usability have also been proposed on how to ensure the usability of a system by users amongst others; preventing errors from occurring right from system design, providing a mechanism for timely feedback and provide necessary help and documentation on systems (Nielsen, 1995). Various usability models have also been developed. For instance, Harrison et al. (2013) proposed a usability model that considered the unique characteristics of mobile devices. Moreover, how Usability is designed in relation to Security is also important. While both Usability and Security are important, the way they are built into a system determines whether the implemented controls would meet the intended objective. The buttress to this argument, is the analogy of user authentication, Ferreira et al. (2009) posited that without a password, a system is more usable, and conversely, an authentication mechanism that frequently requests revalidation while highly secure might be less usable.

Various approaches have been proposed on how to design systems that are both highly secure and usable. A study by Bai et al. (2017) on balancing Usability and Security in the use of encrypted emails explained that encryption was difficult to use because of poor interface design and difficulty in key management. Furthermore, the paper reported the finding of a study that gauged participants understanding and how they valued Usability and Security trade-off in email encryption. Factors like privacy, ease of use and trust were observed to influence Usability and Security trade-off decisions. Also, Cranor and Buchler (2014) advocated considering Usability and Security together during the design. The opinion was that the end-user decision-making process does affect the balance between Usability and Security. They placed the onus on system designers to actively consider which decision requirements are assigned to end-users.

In a bid to improve Usability while minimizing threat scenarios, a study to analyse factors affecting both Security and Usability together was conducted (Kainda et al., 2010). The study proposed a Usability-Security threat model that identified factors to focus on when evaluating Usability and Security attributes. The study identified *Effectiveness, Satisfaction*, *Accuracy* and *Efficiency* as attributable factors that affect Usability only. It also

identified *Attention*, *Vigilance*, *Conditioning*, *Motivation* and *Social Context* as factors affecting Security only. However, *Memorability* and *Knowledge* affect both Usability and Security (Kainda et al., 2010).

In addition to the Usability and Security approaches discussed, Faily and Iacob (2017) proposed the use of a tool to ensure Usable Security. Their paper explains that the proposed tool; CAIRIS (Computer Aided Integration of Requirements and Information Security), facilitates the Usability Security engineering activity by providing the capability for persona development and threat modelling.

2.2 Usable Security Evaluation

To answer the question of how Usable Security can be evaluated in Fintech and how it could be incorporated in the design phase of Fintech solutions, we examined peer-reviewed Usable Security literature from 2010-2020. While some notable studies on system usability have been conducted in earlier years (Nielsen, 92, Zurko and Simon,1996, Adams and Sasse, 1999), the choice of papers was made to coincide with Fintech evolution and Usable Security research conducted in that period.

In a study to improve the usability of security measures Feth and Polst (2019) developed a heuristics-based usability evaluation model together with a model of how to apply the heuristics. The paper opined that the choice for heuristics was due to the reason that hard metrics for security are quite rare and difficult to apply in practice. To ensure the heuristics are human-centred, the heuristics incorporated HCD design principles. The intended audience of the heuristics are Developers and Systems Administrators (Feth and Polst, 2019). Similarly, in a study to address issues of consent data privacy concerns in health information system in the context of the social network paradigm, heuristics were developed to evaluate Usable Security on the system (Yeratziotis et al., 2012). In the same vein, Alarif et al. (2017) proposed a heuristics-based framework for evaluating E-Banking Security and Usability made up of 13 categories and 160 metrics (Alarif et al., 2017).

While the studies we referenced in this section, examined Usable Security evaluation in domains like health, and financial services, others were more component specific. For instance, Realpe et al. (2016) examined the Usable Security of user authentication, Eskandari et al. (2018) examined Usable Security of bitcoin key management, Green and Smith (2016) examined the usability of security APIs for developers and Schryen et al. (2016) examined the usability of CAPTCHAs.

Usable Security evaluation in the reviewed literature was carried out in three ways; experts review, user review, or systems analysis. A combination of user and expert review was also proposed (Nurse et al., 2011).

The studies reveal that heuristics are the most used usability inspection method and help identify errors that could be costly to address. While assessment of heuristics is at times considered unreliable. It often reveals problems that might otherwise affect system security (Yeratziotis et al., 2012).

2.3 Heuristics Development

While no single approach exists for developing heuristics, Table 1.0 provides a guide to steps taken to derive heuristics from.

TABLE 1.0: USABLE SECURITY ELEMENTS

#	STEPS	REFERENCES
1	DERIVE HEURISTICS FROM LITERATURE	YERATZIOTIS ET AL. (2012) FETH AND POLST, (2019) NURSE EL AL. (2011) JIMÉNEZ ET AL. (2012) QUIÑONES AND RUSU (2017)
2	REFINE HEURISTICS	
3	CATEGORIZE HEURISTICS	
4	REVISE FOR COMPLETENESS AND ADD MORE HEURISTICS	
5	PRIORITIZE HEURISTICS	

Usable Security evaluation has been conducted in several domains; however, none exist for the Fintech domain. This study seeks to develop heuristics for Usable Security evaluation in Fintech by adapting research effort in other domains to improve the security posture of Fintech from the human element perspective, using mobile financial services as a case study. In addition to evaluating the usability of user interface design by heuristics principles, usability metrics also exist for that purpose. For instance, the System Usability Scale (SUS) and the Quality in Use Integrated Map (QUIM) have been used to measure the usability of user interface design in specific application domains (Brooke,1996, Seffah et al., 2001, Sivaji et al., 2011).

The study also takes into cognisance existing frameworks and models for Usability and Security evaluation that can be leveraged to address risk identified by Open Web Application Security Project (OWASP) in a Fintech context (OWASP, 2016).

3. METHODOLOGY

The Usable Security heuristics for cybersecurity for Fintech was developed in three (3) iterations and validated by expert interviews. The first iteration was based on a survey of 698 Fintech users. The second iteration was based on a semi-structured interview of thirty-seven (37) participants, comprising Fintech solution providers and Bank Chief Information Officers (CIOs). The third Iteration was based on a thematic analysis of Usable Security evaluation papers published between 2010 to 2020 and an analysis of cybersecurity and Usable Security related framework and procedure. The heuristics developed as an outcome of these iterations were then validated through an interview of fourteen (14) cybersecurity and Usable Security experts.

3.1 Study Design

As described in section two (2), no single approach exists for developing heuristics. However, to ensure we address the major objective of this study which is leveraging human factor approaches to improve Usable Security in Fintech, we designed a study that considered the perspective of key stakeholders in the ecosystem, while taking cognisance of related efforts from literature and industry, this approach in addition to providing heuristics that would improve Usable Security, facilitates traceability from developed heuristics to practical problem it seeks to address. Table 2 provides an overview of the approach adopted in this study.

Table 2: Study Approach

Steps	**Study Method**	**Analysis Approach**	**Output**
Iteration 1	Survey of 698 fintech users	Principal Component Analysis	5 Usable Security Heuristics
Iteration 2	Semi-Structured interview of 37 fintech providers	Thematic Analysis Card sorting	5 Usable Security Heuristics
Iteration 3	Systematic Literature Review Document Analysis	Thematic Analysis	12 Usable Security Heuristics
Consolidated heuristics	Synthesized heuristics	Synthesis	Consolidated heuristics
Validation	Experts interview	ANOVA	Experts feedback on heuristics

3.2 Iteration 1: User Survey

The objective of the user survey was to gain understanding of observable and latent constructs that affect Usable Security for users of Fintech. To conduct this study, a survey instrument consisted of forty-three (43) questions. The questions consisted of thirty (30) Likert-type statements anchored by a five-point scale, ranging from 1 (‘‘strongly disagree” or “Never”) to 5 (‘‘strongly agree” or “always”). The remaining instrument constitutes twelve (12) multiple choice questions and one open-ended question.

The instrument was segmented into nine (9) sections for ease of administration. The questionnaire was then distributed both electronically and paper based. The electronic question was created using Bristol Online Survey (BOS), a survey tool made available by the university library of the authors, and circulated via email, and social media via WhatsApp and Facebook. Hard copies were distributed by hand to market placing targeting audience

without social media presence. The study was aimed at Fintech users who use Mobile Financial Services solutions. The questionnaires were distributed to 1000 respondents in Nigeria. However, only 698 completed questionnaires were returned. Table 3 provides a summary of profile of survey participants.

Table 3: User survey participants' profile

Age	%
18-24	20
25-34	35.6
35-44	36.7
45-60	6.7
= or > 61	1.0
Educational Qualification	%
Primary School Certificate	0.5
Secondary School Certificate	8.4
Diploma	12.3
Undergraduate Degree	42.7
Postgraduate Degree	35.2
Others	0.8
Monthly income	%
< = N 20,000	18.2
N 21,000 – N 50,000	15.6
N 51,000 - N 100,000	20.3
N 101,000 - N 250,000	23.4
N 251,000 - N 500,000	14.9
>= N 501,000	7.6

Princip…hen conducted on the data collated from the survey to identify elemer…search, PCA helped to expose latent variables not visible by using simple…ation (Abdi and Williams, 2010).

3.3 Ite…roviders and Bank CIOs

The ob…able Security elements that impact the practices of Fintech solution provide…ants are developers of Fintech and Bank CIOs. The recruitment proces…dsourcing from various online forums for Fintech solution providers and recommendations from financial services solution providers. Some participants were recruited from www.upwork.com, which provides the ability to filter and contact participants who met the set criteria. The website also provided verifiable evidence of past experiences of participants and their real identities. Sixty (60) participants were recruited but interviews were eventually conducted for twenty-two (22) participants. Four (4) of the participants were from the USA, Eight (8) from Asia, Seven (7) from Africa, two (2) from Europe and one (1) from the Middle East. The average years of experience for participants was eight (8) years. The most years of experience by any participant was fifteen (15) years, while the least number of years of experience by any participant was four (4) years. Irrespective of years of experience, participants have all worked on several successful Fintech projects. Ten (10) participants were Mobile Application Developers, six (6) were either Testers or Quality Assurance experts and three (3) had Governance related qualifications, like Project Management and Solution Architects. One (1) of the participants was a User Interface Design expert while two (2) were Business Relationship and Business Analysis experts. It should be noted that the skills mentioned above were primary expertise, as a number of the participants have played multiple roles in past projects.

The second group consisted of fifteen (15) Banking CIOs who have participated in the deployment of Mobile Financial Services making it a total of thirty-seven (37) participants for the study. The interviews were conducted over three (3) months.

Card sorting technique helped in arriving at the key factors that affect Usable Security from the perspective of the stakeholders (Nurmuliani, et al., 2004). Three (3) Information security experts conducted the card sorting exercise which culminated in the identification of Usable Security heuristics from the second iteration. An online

tool UsabiliTest, (Usabilitest, 2018) was used to conduct the card sorting exercise, the tool provided a user-friendly graphic user interface for card sorting and allowed participants to choose between open, closed or hybrid card sorting options.

3.4 Iteration 3: Literature review

Iteration one (1) and two (2) revealed elements central to Usability and Security and threw up a question on how Usable Security is evaluated and designed. The 3rd iteration of the study was designed to answer the question.

The process included the development of a search strategy and six search strings. The literature search was conducted in the following sources: Sources: ACM Digital Library, USENIX, Science Direct, IEEE Explorer Digital Library, Scopus, Google Scholar, Springer, ResearchGate. Only peer-reviewed papers published in English language between 2010 to 2020 were in scope for the studies. Eighty-eight (88) peer-reviewed papers were identified from the search and analysed using Thematic Analysis. Analysis of Usable Security framework was also conducted as part of the process.

3.5 Consolidation and Validation of Heuristics

This paper adapted the approach presented by Yeratziotis et al. (2012) and Feth and Polst (2019) and integrated the findings from all three iterations, giving rise to a set of heuristics principles and their descriptions.

Twelve (12) heuristics principles together with descriptions and derived heuristics were subjected to expert validation. The validation was conducted in the form of a semi-structured interview. Thirty (30) experts were contacted however, at the end of the validation period fourteen (14) participants took part in the validation, four (4) of the participants are experts based on the USA, four (4) in Nigeria, four (4) in UK, one (1) in Italy and the last one in Lithuania. While six of the experts are cybersecurity experts, seven work in the Human Computer Interaction (HCI) domain and one works in both. Of the fourteen participants, four work in the Financial Services sector, two in the Health sector, one in the Payment industry space, one from the Defense, others from Academia and freelance.

To validate the heuristics, a semi-structured interview with four (4) sections and twenty-nine (29) questions were deployed. All the interviews were conducted virtually as it was conducted during the COVID-19 pandemic where physical contact was restricted.

4. RESULTS

This study culminated in the development of twelve (12) Usable Security heuristics validated by experts. In addition to the heuristics, this section present findings from the studies leading to the development of the heuristics.

4.1 Iteration 1 Result

Principal Component Analysis (PCA) conducted on the data from the survey of 698 Fintech users indicated that out of the total number of respondents been analysed certain commonalities exist in 64% of them. The PCA also identified some observable components that when analysed in a correlation matrix exhibit certain correlations. Based on a comparison of the initial eigenvalues of the six (6) observable component, and extraction sums of square loadings, four (4) components explain 82.76% of the variation. An analysis of the PCA correlation matrix showed the relationship between the six (6) observable matrices. The analysis shows that Usability and Security have the highest positive correlation factor of 0.552, complexity variable has a negative correlation with both Usability (-0.302) and Security (-0.302). The coefficient of end-user privacy variable to Usability is 0.249 while the coefficient of end-user patching variable to security is 0.264. Furthermore, the relationship between the observable and latent factors was analysed using the model generated through the pattern matrix. The first latent component of the matrix loads heavily on Usability (0.869) and Security (0.841) but loads negatively on complexity (-.388). The second component loads positively on Patching and Complexity, while the third component loads only on Environment, while the last component loads heavily on Privacy and inversely on Complexity.

Based on PCA conducted on the data, five (5) heuristics were derived from the study as follows:

4.1.1 Complexity of System

The element addresses the complexity of security controls. While this was identified as a Usability attribute, participants believe addressing this will both improve Usability and Security. Furthermore, the study revealed that though the response from participants indicated that the system was not complex when the aggregate tasks that determine complexity were measured, the result showed the contrary.

4.1.2 Awareness of Privacy

Most participants indicated that they had more than an above-average knowledge of privacy. However, this differed in practice as participant phone use behaviours show a poor understanding of privacy. These participants store and use their logon credential in such a way that jeopardizes the security of their Fintech applications.

4.1.2 End-User patching

Lack of ensuring timely critical update poses a risk for Fintech users. While participants intuitively demonstrated a good habit of ensuring timely critical update on their devices, most are not aware of how this affects security.

4.1.3 Environmental Impact

While other factors results from direct user behaviour, this element measures the impact of factor external to the user and its impact on Usable Security. External factors like the environment of use might constitute a distraction to participants and has an impact on both Usability and Security of the system.

4.1.4 Usability and Security

Usability and Security are factors that have also been identified by participants to impact cybersecurity in Fintech. Furthermore, in ensuring a balance between Usability and Security in Fintech, our result show that Security concerns have more impact on trust than Usability concerns.

4.2 Iteration 2 Result

The heuristics derived from the first iteration were from the perspective of Fintech users, to ensure the final heuristics take cognisance of key stakeholders in the ecosystem, we conducted a second iteration of the study intending to identify more specific elements from the perspective of Fintech solution providers, that could further improve Usable Security in Fintech. To that effect, we conducted a semi-structured interview of Fintech solution development team (22) and bank Chief Information Officers (15), the rest of the section details the findings of the study.

Most development team members tend to play multiple roles. In one instance, a Developer was responsible for User Experience (UX) design, Security and Testing, in another instance a Developer was responsible for all processes from requirements gathering to documentation. While this might shorten development time, it might eliminate checks and balances that might have an impact on Usable Security of the final product. Furthermore, the study revealed that the level of awareness of stakeholders on Usable Security has an impact on how requirement for developing a solution are gathered. End-users are often not aware of what is technically and functionally feasible in securing a system before the development of the solution, as such depend on the development team to address security requirements in the system. However, users can provide input on how to improve usability when a prototype is made available.

Participants identify Agile as the predominant methodology used during the development of Fintech applications for mobile phones. Participants believe the Agile development method helps to achieve both Usability and Security objectives as it tends to reveal security loopholes at the early stages of development before it becomes expensive to correct. According to another participant, Agile provides for continuous interaction between clients and development team, facilitating the chances of deploying an acceptable solution. Participants also agree that development methodology alone was not sufficient to guarantee Usable Security. To achieve Usable Security, both Usability and Security must be deliberately planned into the development process.

Though standard usability and threat scenarios were considered during design, there seemed to be no clear-cut documented usability needs or requirements from customers, Developers depend on business requirements specifications, which regards security as a non-functional requirement. Usability considerations mostly come to the fore during testing. Usability testing is consistently done by in-house teams representing user interest, typically with automated testing tools. In general, there seemed to be no defined approach or minimum expectation during testing. Participants noted that tests must not only be conducted on end-user facing Fintech applications but also on the back-end servers. As one participant puts it "*Mobile apps with financial nature depend heavily on the back-end processes to accomplish tasks, for instance, where a user requests for an Account Statement or transactions, the front-end mobile app must wait for results from the back-end processes to complete before displaying to the user, as such, testing the efficiency of the back-end processes is therefore paramount to the success of the mobile deployment*". However, participants believe testing back-end and ensuring its security is the responsibility of the financial services provider. While functionalities, layouts and user experiences were designed by the development team, they depend on whatever back-end security infrastructure exists.

In deploying Fintech, solution providers are expected to comply with standards and guidelines specified by regulators in addition to payment industry standards like the Payment Card Industry Data Security Standard (PCI DSS). Based on the interviews, it was observed that development teams are guided by various generic development standards, security standards and government regulations. Controls against non-compliance to existing standards include penalties like fines and being placed on the policy violation list. By far the most potent control for ensuring compliance to standards as identified by participants is the reputational risk to the solution provider due to lack of adherence to standards.

While most participants agree that based on experience, Usability and Security should be considered together at every phase of Fintech solution development and deployment, some participants thought otherwise. For instance, one participant believed that a trade-off between Usability and Security should not be the focus during the development of Fintech. The focus he said should be on minimizing the possibility of threat scenarios and maximizing the accessibility of usability scenarios, with more attention given to minimizing threat scenarios. Another participant suggested a risk-based approach whereby the tilt should depend on where the risk lies. The use of analytics to continuously refine Usability and Security was also suggested. Another participant believed that the development team should worry more about Security and allow the users to worry about Usability because no matter the effort developers put in ensuring the balance, users will always have the final say on what is truly usable.

4.2.1 Card Sorting Results

A thematic analysis of the semi-structured interview data revealed factors that affect Usable Security from the perspective of the participants. Using card sorting techniques, the factors were categorized by three (3) Information Security experts and presented herewith as Usable Security heuristics from supply-side stakeholders.

Security and Usability: Eighty-Two (82) of the cards sorted identified security and usability as a factor that should be addressed to improve the security posture of Fintech. Thirty (30) of the eighty-two (82) cards were related to security assurance, fifteen to security, and the rest to usability.

Design: Participants believe system design is a very important element for improving Usable Security in Fintech. Twelve (12) of the cards identified design as a factor affecting Usable Security.

Communication: Communication and feedback in Fintech transaction affect user confidence and trust in the use of the solution. Thirteen (13) cards identified communication as an important Usable Security element for Fintech.

Quality: Quality relates to the correct elicitation and coding of user requirements and the testing of the solution based on these requirements. Eleven (11) of the cards identified quality as an important Usable Security element for Fintech.

Operations and Infrastructure: Environmental factors outside the control of the user, but within the control of the solution providers have an impact on the security of the Fintech applications. Twenty-nine (29) cards identified this factor as an element.

4.3 Iteration 3 Results

The first two iterations identified Usable Security heuristics from the perspective of stakeholders in the system. This iteration examines existing work from other domains with a view for identifying elements that can be applied to improve Usable Security in Fintech. Based on thematic analysis of Usable Security evaluation literature from 2010 to 2020, and an analysis of Usability and Security frameworks, the following heuristics were identified in Table 4.

Table 4: Usable Security Elements

#	Heuristic	Reference
1	Integrity	Gaehtgens et al. (2017) Feth and Polst (2019) Yeratziotis et al. (2012)
2	Proportionality	Feth and Polst (2019) Yeratziotis et al. (2012)
3	Transparency	Realpe et al. (2016)

#	Heuristic	Reference
		Feth and Polst (2019) Gaehtgens et al. (2017) Yeratziotis et. Al. (2012)
4	Empowerment	Alarifi et al. (2017) Melicher, et al. (2016) Feth and Polst (2019) Yeratziotis et al. (2012)
5	Identity	Gaehtgens et al. (2017) Feth and Polst (2019)
6	Reliability	Uzun et al. (2011) Alarifi et al. (2017) Hof (2015)
7	User Support	Feth and Polst (2019) Yeratziotis et al. (2012) Hof (2015)
8	Accessibility	Feth and Polst (2019) Hof (2015)
9	Authenticity	Yeratziotis et al. (2012) Khan (2015) Kainda, R., et al. (2010)
10	Compliance	Alarifi et al (2017)
11	Alignment	Hof (2015) Khan (2015)
12	Freedom	Hof (2015) Khan (2015)

4.3.1 Consolidate Heuristics Principle

This section presents a mapping of heuristics from the three () iterations. Usable Security as a factor from iteration one and two was not included as the entire heuristics is meant to address Usable Security. Table 5 shows Usable Security elements derived from the three iterations.

Table 5: Usable Security Elements

#	Iteration		
	One	Two	Three
1		Quality	Integrity
2	Complexity	Quality	Proportionality
3		Design	Transparency
4	Awareness of privacy		Empowerment
5			Identity
6	Environmental	-Design -Communication -Operations and Infrastructure	Reliability
7	-Awareness of privacy -Patching		User Support
8	Complexity		Accessibility
9			Authenticity
10			Compliance
11			Alignment
12			Freedom

The detail of the twelve (12) identified heuristics and their description is as shown below:

Integrity:

This factor address controls against the unauthorized modification of transaction data. It consists of measures put in place for data protection.

Derived heuristics:

Protected area should be inaccessible to unauthorized users

System should automatically test and install the required software update without making the system more vulnerable or less usable

Proportionality: Ensure security controls are proportionate to users' knowledge, time, transaction type and cognitive ability.

Derived heuristic:

System supports both novice and expert users

Users should be able to customize security to meet their individual preferences

Transparency: Ensure security controls and practices are comprehensible, verifiable and accessible for the user.

Derived heuristics:

System security status should be obvious to use irrespective of knowledge of the security mechanism

Users should be able to understand what security mechanism is active.

Empowerment: Enable users to express their systems security needs in the most efficient way

Derived heuristics:

User should be able to customize security preferences

User should be able to reverse certain security choices.

Identity: Ensure that users can be uniquely identified and verified with a high level of assurance

Derived heuristics:

Authentication options designed in a way to keep the cognitive load of users low

Reliability: Ensure service consistency and functionality on facilitating effective communication and feedback for user transactions and security actions

Derived heuristics:

The system should communicate error and transaction status to users in an understandable manner.

User Support: Ensure measures are put in place to support and educate users on the use of the system and security controls without additional cognitive workload on users.

Derive heuristics:

Security operations should be easy to learn and apply irrespective of user cognitive ability.

Only relevant security information should be provided

Accessibility: Ensure the system and security control do not discriminate against any user

Derived heuristic:

The security mechanism should have consideration for accessibility,

A visually impaired user should be able to differential a genuine from a rogue Fintech application

Authenticity: Ensure the system has valid certificates and the information should be available on the interface of use.

Derived heuristics:

System should alert users when they are interacting with non-trustworthy sources

Compliance: Ensure system and security control complies with extant policies, guidelines

Derived heuristics:

Test conditions and scenarios should address compliance to extant policies and regulations

Alignment: Ensure security mechanisms aligns with the usual flow of user activities, mental model and cognitive ability

Derived heuristics:

Security controls should not add to the cognitive workloads of the user

Freedom: Ensure security mechanisms guarantee a certain degree of freedom to users

Derived heuristics:

Security control should not limit user option in the use of the application

4.4 Heuristics Validation

All fourteen (14) experts that participated in the validation of the heuristics agreed on the importance of all twelve (12) heuristics and provided feedback they believe would further strengthen the heuristics. This section provides results from the heuristics validation interview.

One expert suggested "Consistency" might be a better description of the heuristics currently labelled "*Integrity*" as also addresses consistency of transaction throughout its life cycle. An expert noted that *Proportionality* might be difficult to implement as a decision needs to be made as to whether it should be implemented as a dynamically aware system or coded into the system during the design phase. Affordance was suggested as a more suitable description for *User Support.* Experts noted that it was important to take care that end-users were not burdened with too much documentation as it would counteract the objective of Usability. Experts recommended that *Compliance* should be further decomposed to address contractual requirements, legal requirements and regulatory standards. Experts recommended the merging of *Integrity* and *Reliability* and *Authenticity* and *Identity*.

ANOVA was carried out to determine differences in the mean perception of respondents by country and sector. The perception was gauged for when the factors are used to evaluate Usable Security and when they are used as a guide to design Usable Security into the system. Response from four experts was deleted from the model because they did not complete this section of the questionnaire. Table 6 below shows the descriptive statistics of the model.

TABLE 6: DESCRIPTIVE STATISTICS

Descriptives

		N	Mean	Std. Deviation	Std. Error	95% Confidence Interval for Mean		Minimum	Maximum
						Lower Bound	Upper Bound		
Weighted_Evaluation	Nigeria	2	33.00	.00	.00	33.00	33.00	33.00	33.00
	UK	3	28.33	8.08	4.67	8.25	48.41	19.00	33.00
	US	4	33.25	.50	.25	32.45	34.05	33.00	34.00
	Lithuania	1	33.00	.	.	.	.	33.00	33.00
	Total	10	31.70	4.47	1.41	28.50	34.90	19.00	34.00
Weighted_Design	Nigeria	2	33.00	.00	.00	33.00	33.00	33.00	33.00
	UK	3	29.67	9.45	5.46	6.19	53.15	19.00	37.00
	US	4	37.00	2.94	1.47	32.32	41.68	33.00	40.00
	Lithuania	1	33.00	.	.	.	.	33.00	33.00
	Total	10	33.60	5.76	1.82	29.48	37.72	19.00	40.00

The model shows that the mean of Nigerian experts is thirty-three (33) with a standard deviation of zero (0) while UK experts have a standard deviation of 8, which shows a more divergent view, the value is smaller for US experts.

ANOVA test was conducted to test the statistical significance of the elements when used for evaluation and when applied to design.

Table 7 shows the detail of the ANOVA test conducted. The test shows that there was no statistically significant difference between groups was determined by one-way ANOVA ($F(3, 6) = 0.74$, $p = .565$). No statistically significant difference, in the perception of the respondent.

Table 7: ANOVA

		Sum of Squares	df	Mean Square	F	Sig.
Evaluation	Between Groups	48.68	3	16.23	.74	.565
	Within Groups	131.42	6	21.90		
	Total	180.10	9			
Design	Between Groups	93.73	3	31.24	.92	.488
	Within Groups	204.67	6	34.11		
	Total	298.40	9			

5. DISCUSION AND CONCLUSION

The Usable Security heuristic principles presented in this work seeks to improve the usability of security mechanisms in Fintech applications. The heuristics developed can be applied to evaluate the Usable Security of existing systems or as a guide to design Usable Security during Fintech application development. While heuristics are generally developed from existing literature, extensive work was conducted to develop heuristics from a

sociotechnical perspective. The approach adopted facilitates heuristics traceability and reduce cybersecurity risk associated with the human element in the use of Fintech applications.

This study argued that cybersecurity issues still affect Fintech despite the availability of strong technical countermeasures. The proposed heuristics do not intend to replace existing technical countermeasures but make them more usable to end-users irrespective of their knowledge of the systems, security controls and physical ability.

The fourteen (14) experts that validated the heuristics all agree that the heuristics are apt in achieving the study objective but suggested that some of the elements could be merged, while the derived heuristics under each are retained. The experts also opined that the heuristics can be used in a Fintech sandbox process as criteria to ensure the Usable Security of the final product.

The suitability of the heuristics for evaluation of Fintech and design of Fintech solution was ascertained by participants. However, the level of importance was different for some element when used for evaluation compared to when used for design. Also, the view of the importance of each element was dependent on the domain of the evaluator, while HCI professionals tend to rank HCI related elements higher, security experts tend to rate security inclined elements higher. Irrespective of the level of priority given to the element by each group, they all emphasised the importance of all elements in the evaluation of Usable Security.

The development heuristics would be of benefit to Fintech Developers, Systems Auditors and Systems Administrators and end-users of Fintech solutions.

6. FUTURE WORK

This study has answered the research question of how to evaluate Usable Security in Fintech using a case study of mobile financial services, by developing twelve (12) Usable Security heuristics using an iterative approach that took cognisance of key players in the sociotechnical system.

Future work will involve using heuristics to evaluate Fintech solutions and compare them side by side with other usability heuristics. To determine how the heuristics will serve as a design guide, a hackathon will be organised where the heuristics principles will be used to guide development and then compared to existing development practices.

7 References

Abdi, H. and Williams, L.J., (2010). Principal component analysis. Wiley interdisciplinary reviews: computational statistics, 2(4), pp.433-459.

Abdillah, L. (2019, December). An Overview of Indonesian Fintech Application. In The First International Conference on Communication, Information Technology and Youth Study (I-CITYS2019), Bayview Hotel Melaka, Melaka (Malacca), Malaysia.

Alarifi, A., Alsaleh, M., & Alomar, N. (2017). A model for evaluating the security and usability of e-banking platforms. *Computing*, *99*(5), 519-535.

Asli, D., Klapper, L., Singer D., Ansar, S. and Hess, J, (2018), The Global Findex Database 2017: Measuring Financial Inclusion and the Fintech Revolution. Washington, DC: World Bank. doi:10.1596/978-1-4648-1259-0. License: Creative Commons Attribution CC BY 3.0 IGO.

Android, (2018). Android user interface development beginners guide, 2018. http://index-of.es/Android/Android.User.Interface.Development.Beginner.Guide.pdf. (Retrieved 22nd March 2021)

Apple, (2018). Human Interface guidelines. https://developer.apple.com/ios/human-interface-guidelines/overview/themes/, (Retrieved 22nd March 2021)

Bai, W., Kim, D., Namara, M., Qian, Y., Kelley, P. G., & Mazurek, M. L. (2017). Balancing security and usability in encrypted email. *IEEE Internet Computing*, *21*(3), 30-38.

Borrett, A. (2021). Techmonitor, Covid-19 has increased cybersecurity risk to the fintech ecosystem, https://techmonitor.ai/technology/cybersecurity/cybersecurity-risk-fintech-ecosystem (Retrieved 26th April, 2021)

Brooke, J. (1996). others,"SUS-A quick and dirty usability scale," Usability Eval. *Ind*, *189*, 4-7.

Cranor, L. F., & Buchler, N. (2014). Better together: Usability and Security go hand in hand. *IEEE Security & Privacy*, *12*(6), 89-93.

Enhancing Financial Innovation and Access (EFInA), (2020) FinTech Landscape and Impact Assessment Study

Eskandari, S., Clark, J., Barrera, D., & Stobert, E. (2018). A first look at the usability of bitcoin key management. *arXiv preprint arXiv:1802.04351*.

Faily, S., & Iacob, C. (2017, September). Design as code: Facilitating collaboration between Usability and Security engineers using cairis. In *2017 IEEE 25th International Requirements Engineering Conference Workshops (REW)* (pp. 76-82). IEEE.

Ferreira, A., Rusu, C., & Roncagliolo, S. (2009). Usability and Security patterns. In *2009 Second International Conferences on Advances in Computer-Human Interactions* (pp. 301-305). IEEE.

Feth, D., & Polst, S. (2019). Heuristics and models for evaluating the usability of security measures. In *Proceedings of Mensch und Computer 2019* (pp. 275-285).

Gaehtgens, F., Allan ,A., Zlotogorski, M., Buytendijk,F. (2017). Definition: Digital Trust. Gartner research Published: 24 May 2017 **ID:** G00329409.

Glenny, M., (2021). Financial Times, Pandemic accelerates growth in cybercrime. https://www.ft.com/content/49b81b4e-367a-4be1-b7d6 166230abc398?desktop=true&segmentId=d8d3e364-5197-20eb-17cf-2437841d178a#myft:notification:instant-email:content (Retrieved 29th April, 2021)

Gorski, P. L., & Iacono, L. L. (2016). Towards the Usability Evaluation of Security APIs. In *HAISA* (pp. 252-265).

Green, M., & Smith, M. (2016). Developers are not the enemy!: The need for usable security apis. *IEEE Security & Privacy*, *14*(5), 40-46.

Hafenbrädl, S., Waeger, D., Marewski, J. N., & Gigerenzer, G. (2016). Applied decision making with fast-and-frugal heuristics. *Journal of Applied Research in Memory and Cognition*, *5*(2), 215-231.

Harrison, R., Flood, D., & Duce, D. (2013). Usability of mobile applications: literature review and rationale for a new usability model. *Journal of Interaction Science*, *1*(1), 1-16.

High M, (2021). Monzo, Revolut and more - the rise of UK fintechs, https://www.fintechmagazine.com/venture-capital/monzo-revolut-and-more-rise-uk-fintechs (retrieved 8th May, 2021)

Hof, H. J. (2015). User-centric IT security-how to design usable security mechanisms. *arXiv preprint arXiv:1506.07167*.

Hof, H. J. (2015). Towards enhanced usability of it security mechanisms-how to design usable it security mechanisms using the example of email encryption. *arXiv preprint arXiv:1506.06987*.

Jiménez, C., Rusu, C., Roncagliolo, S., Inostroza, R., & Rusu, V. (2012). Evaluating a methodology to establish usability heuristics. In *2012 31st International Conference of the Chilean Computer Science Society* (pp. 51-59). IEEE.

Kainda, R., Flechais, I., & Roscoe, A. W. (2010, February). Security and usability: Analysis and evaluation. In *2010 International Conference on Availability, Reliability and Security* (pp. 275-282). IEEE.

Khan, H., Hengartner, U., & Vogel, D. (2015). Usability and Security perceptions of implicit authentication: convenient, secure, sometimes annoying. In *Eleventh Symposium On Usable Privacy and Security ({SOUPS} 2015)* (pp. 225-239).

Mani V, (2019), Cybersecurity and Fintech at a Crossroads, ISACA Journal / Issues / 2019 / Volume 1

Melicher, W., Kurilova, D., Segreti, S. M., Kalvani, P., Shay, R., Ur, B.,and Mazurek, M. L. (2016). Usability and Security of text passwords on mobile devices. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 527-539).

Nielsen, J. (1992). Finding usability problems through heuristic evaluation. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 373-380).

Nurmuliani, N., Zowghi, D., & Williams, S. P. (2004). Using card sorting technique to classify requirements change. In *Proceedings. 12th IEEE International Requirements Engineering Conference, 2004.* (pp. 240-248). IEEE.

Nurse, J. R., Creese, S., Goldsmith, M., & Lamberts, K. (2011). Guidelines for usable cybersecurity: Past and present. In *2011 third international workshop on cyberspace safety and security (CSS)* (pp. 21-26). IEEE.

OWASP Mobile Top 10, (2016) https://owasp.org/www-project-mobile-top-10/ (Retrieved 7th May, 2021)

Pfleeger, S. L., Sasse, M. A., & Furnham, A. (2014). From weakest link to security hero: Transforming staff security behavior. *Journal of Homeland Security and Emergency Management*, *11*(4), 489-510.

Quiñones, D., & Rusu, C. (2017). How to develop usability heuristics: A systematic literature review. *Computer standards & interfaces*, *53*, 89-122.

Realpe, P. C., Collazos, C. A., Hurtado, J., & Granollers, A. (2016). A set of heuristics for usable security and user authentication. In *Proceedings of the XVII International Conference on Human Computer Interaction* (pp. 1-8).

Saksonova, S., and Kuzmina-Merlino, I. (2017). Fintech as financial innovation–The possibilities and problems of implementation.

Sasse, M. A., Brostoff, S., & Weirich, D. (2001). Transforming the 'weakest link'—a human/computer interaction approach to usable and effective security. *BT technology journal*, *19*(3), 122-131.

Schryen, G., Wagner, G., & Schlegel, A. (2016). Development of two novel face-recognition CAPTCHAs: a security and usability study. *Computers & Security*, *60*, 95-116.

Seffah, A., Kececi, N., & Donyaee, M. (2001). QUIM: a framework for quantifying usability metrics in software quality models. In *Proceedings Second Asia-Pacific Conference on Quality Software* (pp. 311-318). IEEE.

Shetty M., (2018), Banks warn of new mobile malware, 232 banking apps in danger. https://timesofindia.indiatimes.com/business/india-business/banks-warn-of-new-mobile-malware/articleshow/62436145.cms. (Retrieved May 12, 2018).

Sivaji, A., Abdullah, A., & Downe, A. G. (2011). Usability testing methodology: Effectiveness of heuristic evaluation in E-government website development. In 2011 Fifth Asia Modelling Symposium (pp. 68-72). IEEE.

Usabilitest, (2018). https://www.usabilitest.com/. (Retrieved 31st August 2018)

Uzun, E., Saxena, N., & Kumar, A. (2011). Pairing devices for social interactions: a comparative usability evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 2315-2324).

West, R., Mayhorn, C., Hardee, J., & Mendel, J. (2009). The weakest link: A psychological perspective on why users make poor security decisions. In *Social and Human elements of information security: Emerging Trends and countermeasures* (pp. 43-60). IGI Global.

Wijayarathna, C., & Arachchilage, N. A. G. (2019). Why Johnny can't develop a secure application? A usability analysis of Java Secure Socket Extension API. *Computers & Security*, *80*, 54-73.

Yeratziotis, A., Pottas, D., & Van Greunen, D. (2012). A usable security heuristic evaluation for the online health social networking paradigm. *International Journal of Human-Computer Interaction*, *28*(10), 678-694.

Zurko, M. E., & Simon, R. T. (1996). User-centered security. In *Proceedings of the 1996 workshop on New security paradigms* (pp. 27-33).

Support Rather Than Assault – Cooperative Agents in Minecraft

Danielle Potts, Kate MacFarlane and Lynne Hall

Accenture UK and University of Sunderland

Danielle-19-93@live.com, kate.macfarlane@sunderland.ac.uk, lynne.hall@sunderland.ac.uk

With the dominant trope of the computer as adversary rather than enabler, reinforcement learning for games has mainly focused on the ability of agents to compete and win. Although cooperation is a product of learning, of understanding the player's requirements and applying agents' competences to fulfil them, there has been little investigation of reinforcement learning for cooperation in games. Reinforcement learning results in the agent adapting and changing, however, there are concerns that such adaptivity could alienate users if their cooperative agent outperforms them. To explore this, the paper outlines the development and training of cooperative agents reporting users' positive response to adaptive cooperation in games.

Reinforcement Learning, Cooperative Interaction, Games Artificial Intelligence, Collaboration, Minecraft

1. Introduction

This research explores how games technology and adaptable Artificial Intelligence can be fused to explore the gaps left by the lack of recent consideration of cooperative agents for players in non-competitive games worlds. It aimed to allow real-world users to interact with an adaptable agent in a game world. Unlike in most games, the agent would not be a competitor nor a danger. Rather, it would provide some sort of practical collaborative aid to the user. However, key challenges lie not only in identifying how agents can quickly learn to be useful and adapt, but also in their acceptability to users as cooperative agents.

Machine learning, in particular deep learning, has seen a surge of interest in recent years (Botvinick et al., 2019). It has been successfully implemented in previously unachievable tasks such as language translation and object detection (Arulkumaran, Deisenroth, Brundage and Bharath, 2017). Becoming cooperative, learning to be useful in response to a player, to assist rather than attack, is one such challenging task.

Machine learning is heavily used in games development (Kaliappan and Sundararajan, 2020) with games providing a powerful test bed for machine learning research (Dann, Zambetta and Thangarajah, 2018). Reinforcement learning has had particular success in games software (Arulkumaran, Deisenroth, Brundage and Bharath, 2017; Nair et al., 2018)) – with the goal of outdoing expert users and has been successfully applied to play games such as Chess, Atari, Doom and Starcraft (Botvinick et al., 2019, Xu and Chen, 2019). Reinforcement learning research in games has mainly focused on competitive agents, which compete with the player, they increase their skill level based on the progress of the player (Barros et al, 2020).

This paper discusses an alternative to competitive Non-Player Characters (NPCs), the development and training of NPCs that are adaptive agents designed to support a user in their game tasks. Section 2 introduces relevant related research. Section 3 explores the application of machine learning techniques to create cooperative agents. Section 4 reports a study of user responses to these cooperatively adaptable agents. The final section considers the results and future directions.

2. Related Research

AI is an intricate part of modern games for both world building and to increase challenge – a core piece of the experience (Xu and Chen, 2019). Most games use traditional techniques such as finite state machines (de Almeida Rocha and Cesar Duarte, 2019). However, with finite state machines, agents will always be limited to their routines and play styles. Velardo, (2019) discusses this in the context of the popular sandbox game Red Dead Redemption. There are thousands of NPCs in this game with seemingly complex behaviour changing dependent on player actions, however, the behaviours are still limited to a finite set.

Machine learning driven adaptive AI provides unique features with new behaviours and routines developing overtime, increasing the realism of the environment and allowing agents to better adapt to player skill levels. Game difficulty increasing as a player's performance improves is common in games. Microsoft attempt to utilise this with their Forza series. Forza employs adaptive techniques to adapt agents to human skill (Walsh, 2018) as the users play. Machine learning is used to estimate players' skill levels and to personalise bot skill levels for the player to compete against (Orland, 2013).

In The Last of Us Part II, developers Naughty Dog pushed the AI of their NPCs to achieve a higher level of realism for its players (Hara, 2020: Online). These characters would call out to nearby NPCs for help and clutch the area on their body where they had been wounded by the human player. They also appeared to have relationships with other NPCs, which manifested in NPCs appearing to be getting more aggressive if they saw their 'friends' under attack. While this gives the player another level of immersion into the environment, it is an increase to the threat level and game difficulty for the player.

Kahn, (2017) noted that competitive adaptable agents can appear threatening when they out compete people. Often, their purpose is to be better than expert players. This itself poses an issue with adaptable agents, especially in competitive games. The experience of agents that can endlessly adapt may appear unfair and frustrate users. A proposed solution to this is to avoid using competitive style games to introduce these agents, but rather instead introduce them in games that promote cooperation.

There are some examples of cooperative NPCs providing help and support to players in competitive gaming experiences. A notable example is Star Wars: The Old Republic (2021: Online). In this game the player has a choice of companion characters to travel alongside them as they level up and complete missions in this massively multiplayer online game from BioWare. These companion characters can provide a range of support roles (act as a healer, help with combat, or take damage) to the player character. This proved to be very popular with the player base for the game, with the relationships with the various companion characters being integral to their overall gaming experience.

There remains a gap, a continuing lack of emphasis on experiences that are non-competitive. However, in many other application areas of machine and reinforcement learning, the emphasis has not been on competition but rather on improving work processes (Daugherty and Euchner, 2020).

3. Cooperative Agents In Minecraft

The project aimed to develop and assess an adaptable agent who would, eventually through a period of training provide some sort of practical collaborative use to the user within a game world. The goal was not to produce a competitive agent, as much of the reinforcement learning research does, rather an agent who could add value for the user.

Game World - Minecraft

The game world selected for the user experience was Minecraft, currently one of the most successful computer games, with over 140 million registered user accounts, (Clement, 2021: Online). Minecraft was selected because it is an open world sandbox game, which gives players total freedom over what they do and how they play. They can choose to build, explore, or fight enemies – there is no single goal. And further, Minecraft is not competitive, with players often working together to achieve goals.

Minecraft as a deep learning task is well researched but poses some challenges such as the amount of time required just to achieve a single goal (Scheller, 2020). Further, because of its hierarchical nature, how best to tackle this issue is contested. Reynard, Kamper, Engelbrecht and Rosman, (2020) explored tackling the issue as a steppingstone style issue, creating many small tasks for the agent to solve before introducing it to the fully complex game world. They compared this to throwing the agent straight into the world, a sink or swim approach, discovering that building up from smaller to larger tasks was a more promising approach. The types of small problems given to the agent, and the data and rewards impact the learning outcomes for the agent.

Jaderberg et al., (2016) explore how this level of general intelligence requires the agent to solve many smaller auxiliary goals along the way. The issue with this is that the reward can become skewed, in that, the agent can become distracted from the ultimate goal. Moreover, the task of deciding what should be a reward and what shouldn't be is considered the sparse reward problem common in other environments besides Minecraft (Nair et al., 2018). Perez et al., (2019) addresses achieving the many subtasks in Minecraft using multi-agents, whereby one agent may be very good at navigating, and another may be very good at mining wood. Despite the challenges, Minecraft was chosen because it is not competitive.

Goals of the Agent and User

Minecraft requires a player to mine materials before they build structures. Mining resources can take some time and can become repetitive. Thus, an agent to help the player with this was the goal. The Minecraft agent was to be trained to do two things:

To navigate around the world logically

To mine trees. Mining trees produces wood, a necessary resource for creating tools and weapons in the game.

A user would be required to interact with the game as normal - simply to play the game as they wish. The agent could then be used as helper to gather a particular resource, in this case, wood. This would help to speed up their progress and remove a lot of the repetitive nature of the game.

Cooperative Agent Development

Reinforcement learning (RL) was the method selected to train the cooperative agent due to its effectiveness in game environments. It has been applied to Minecraft before (Milani et al., 2020). The solution to training the agent would use a deep learning model due to the high complexity of Minecraft and the size of the dataset.

RL focuses on the reward, but how to get to that reward is not always clear. Most methods will give the agent no prior knowledge about the environment it is expected to work within. In many cases, this solution works because the environment is of low dimensionality or complexity. As the environment complexity increases, so too does the difficulty.

When humans approach a problem, they rarely have zero knowledge about the domain or how to solve it. They consider a solution with existing knowledge, even if the new problem is not one within known experiences, able to make assumptions based on an accumulation of experiences (Efros et al., 2018). Replicating this, before entering an environment the agent is given some domain information. For example, in (Hester et al., 2017), the agent first learns about the problem via a deep neural network before interacting with the environment.

Q-Learning extends reinforcement learning by implementing an estimation technique, which aims to satisfy the Bellman Equation used to estimate future rewards. Given a current state and available actions it will perform in subsequent states based on the decision at the current point. This method is used to assess the goodness of a particular policy given the following state action pairs observable at future states (Sutton et al., 2018).

Deep Q-Learning subsequently implements deep neural networks into the equation and uses two networks. One network is the network being trained; the other is the 'target' network. When the network evaluates its actions, it measures the loss against the target network. Every so many episodes during training, the target network is updated by copying the parameters of the trained network promoting continuous learning (Silver et al., 2016). The output is an action along with a prediction, which is a percentage that represents the likelihood that each action will lead to reward maximisation.

To train the agent to play Minecraft, Hester et al.'s, (2017) work was the primary inspiration for the final model type – Deep-Q Learning from Demonstrations (DQfD). This is an imitation learning style approach to Deep Q-Learning. It utilises a Deep Neural Network for policy optimisation and a standard Convolutional Neural Network for image recognition (Kunanusont et al., 2017). The Python coded cooperative agent used Tensorflow and Keras.

Using MineRL to train the agent.

To train the agent the MineRL dataset was used. This is composed of 500 hours of image-based demonstrations broken down into different tasks and goals such as navigation and tree chopping, the two selected tasks for the agent. It has over 60 million state action pairs of human demonstration from the game. Observations are made as arrays of low-resolution images making them easier to compute. These were then pre-processed as standard for convolutional neural networks. Google Collab was used to speed up data pre-processing and parsing.

Two MineRL environments and datasets were used for training: Navigate and Treechop. The goal of navigate was to learn how to logically navigate the environment. This includes learning to walk, jumping in certain places and placing dirt to reach high places. In Treechop, the agent had to learn two things: how to walk and how to cut down trees.

Cooperative Agent Outcomes

Whilst the model did not entirely solve the issues it faced; it did make some progress. In Navigate, the agent was able to navigate the environment. In Treechop, the agent often did attempt to mine things. It struggled to successfully locate and target trees. There was clear evidence to show the agent was learning, for example with Treechop the agent's most common action became attack with rewards increasing over time. However, to be able to find the trees the agent also needs improve navigation, something which proved a challenge for the agent.

User Study

The purpose of the user study was to explore reactions to the cooperative agent and to investigate whether this engendered trust in cooperative intelligent agents in other environments.

Method

The original goal for the demonstration application was to allow users to try it in person. However, due to the Covid-19 pandemic this was not possible. With Minecraft being a paid application – the logistics of expecting people to set the application up on their machines was limited. To get around this challenge, a video of the agent was recorded which was shared with a survey to gather some insight about people's feelings towards AI and agents.

The users were given the context around the study, that this was an experimental application of AI in Minecraft looking at cooperative agents. They were then requested to watch a video containing a Minecraft agent attempting to mine some trees. After this, the users would be requested to answer a survey to record their reception and opinions around the video and to explore opinions around this kind of implementation and use of AI.

31 participants engaged in the study, a mixture of people from those who worked within tech and were at least familiar with machine learning and AI, and to those who had no experience or knowledge of it. Participants were recruited through personal contacts and opportunistic sampling due to the pandemic context.

Results & Interpretation

61% of respondents said that they somewhat trusted artificial agents, 31% did not trust them at all and only 8% had a positive level of trust in them. This lack of trust is also seen other with AI-based technologies, for example some people distrust voice assistants believing that they are continually and cleverly listening to them. The survey results indicated that participants were wary of intelligent agents suggesting they might be predisposed to have a negative view of the helper agent.

However, just as the lack of trust in Alexa does not deter use, neither does the lack of trust in agents impact on the use or intention to use agents. 76% of participants indicated that they believed that they regularly interacted with artificial agents on a daily basis. Participants indicated that they were willing to work with intelligent agents, 39% stated that they would be willing to work with an agent with general intelligence, 44% might be willing to and 17% said that they would not be willing to work with them.

Participants were keener to interact with agents in their personal life with 48% happy to interact with an agent out of the workplace and a further 32% who might be. There was even more willingness to engage when agent usefulness was considered, like the helper agent in Minecraft. When asked if they would find it useful to have a helper agent in Minecraft to carry out basic tasks, 64% of the respondents said that they would, 22% thought that it might be a useful addition and14% said they would not find it useful. Only 1 participant would still feel threatened by agents like the one in the demonstration video,14% said that they might feel threatened, but the majority (83%) said that they would not feel threatened.

However, participants identified they would be less comfortable with a helper agent in the workplace. 23% of the sample believed that AI would impact the workplace in a negative way. However, mainly the view was positive with 57% of participants believing that AI would enhance the workplace and create new types of jobs. In comparison, many of the participants still felt that AI/agents were somehow threatening with 54% believing that AI is threatening to some degree.

Generally, people found the idea of having an agent in a game as a helper to be positive. Participants also suggested other types of tasks in Minecraft, as well as other games in which this type of AI could be useful to them, like defending against enemies and to help those with increased accessibility needs.

Conclusion and future work

Games provide a powerful test environment for reinforcement learning research. Such gamification provides a powerful tool for training algorithms before integrating them into real world systems (Riedmiller et al., 2018). Moving towards more cooperative gameplay has significant potential for some player experiences. Using adaptable helper agents could lead to greater game longevity and more immersive gaming in both competitive and non-competitive gaming environments.

This study has highlighted the potential of cooperative agents in game worlds. A key issue that this study identified is the lack of trust in AI-based technologies. Cooperative helper agents offer a clear route to engender higher levels of trust in AI systems, particularly, if this was the guise in which agents were first introduced to users.

However, there are significant limitations to the study at this point, primarily that user engagement involved watching a video rather than interacting with the agent in Minecraft.

Future work focuses on replicating the experiment with the user working on activities with a need for wood, with the agent cooperating by chopping that wood. This will benefit from a more refined agent architecture where the

agent is provided with the underpinning skills necessary to achieve the high-level task. The Minecraft helper agent could be further developed to provide other support services like gathering food and build tasks.

References

Arulkumaran, K., Deisenroth, M., Brundage, M. and Bharath, A, "Deep Reinforcement Learning: A Brief Survey". *IEEE Signal Processing Magazine*, 34(6), 2017, pp.26-38, Doi: 10.1109/MSP.2017.2743240

Barros, P., Tanevska, A. and Sciutti, A., 2020. *"Learning from Learners: Adapting Reinforcement Learning Agents to be Competitive in a Card Game",* viewed 10 May 2021 https://arxiv.org/abs/2004.04000

Clement, J., 2021, *Minecraft number of players worldwide.* Statista. Available at: https://www.statista.com/statistics/680139/minecraft-active-players-worldwide/ (accessed 5.10.21).

Dann, M., Zambetta, F. and Thangarajah, J., 2018. "Integrating Skills and Simulation to Solve Complex Navigation Tasks in Infinite Mario", *IEEE Transactions on Games,* 10(1), 2018 ,pp.101-106, doi: 10.1109/TCIAIG.2017.2696045

Daugherty, P. and Euchner, J., 2020. "Human + Machine: Collaboration in the Age of AI", *Research-Technology Management*, 63(2), 2020, pp.12-17, doi: https://doi.org/10.1080/08956308.2020.1707001

de Almeida Rocha, D. and Cesar Duarte, J., 2019. "Simulating Human Behaviour in Games using Machine Learning", *2019 18th Brazilian Symposium on Computer Games and Digital Entertainment (SBGames)*, 2019, doi: 10.1109/SBGames.2019.00030

Efros, A., Griffiths, T., Pathak, D., Agrawal, P. and Dubey, R., 2018. "Investigating Human Priors for Playing Video Games", *International Conference on Machine Learning*, viewed 19 December 2020, https://arxiv.org/abs/1802.10217

Gough, C., 2020. *Minecraft Unit Sales Worldwide* 2020 | Statista. Available at: https://www.statista.com/statistics/680124/minecraft-unit-sales-worldwide/#:~:text=Minecraft%20unit%20sales%20worldwide%202016%2D2020&text=Since%20its%20release%20in%202011,and%20Grand%20Theft%20Auto%20V (Accessed: 19 December 2020)

Hara, R., 2020. The Last of Us 2 Dev Reveals Major AI Improvements. *Game Rant*. Available at: https://gamerant.com/last-of-us-2-ai-improvements/ (Accessed: 10 May 2021).

Hester, T., Vecerik, M., Pietquin, O., Lanctot, M., Schaul, T., Piot, B., Horgan, D., Quan, J., Sendonaris, A., Dulac-Arnold, G., Osband, I., Agapiou, J., Leibo, J. and Gruslys, A., 2017. "Deep Q-learning from Demonstrations", *Association for the Advancement of Artificial Intelligence*, viewed 19 December 2020, <https://arxiv.org/abs/1704.03732>

Jaderberg, M., Mnih, V., Czarnecki, W., Schaul, T., Leibo, J., Silver, D. and Kavukcuoglu, K., 2016. "Reinforcement Learning with Unsupervised Auxiliary Tasks", *DeepMind,* viewed 19 December 2020, https://arxiv.org/abs/1611.05397

Home, 2021. *The Elder Scrolls Online*. Available at: https://www.elderscrollsonline.com/en-us/home (accessed 5.10.21).

Kahn, J., 2017. *Bloomberg - Are You A Robot*? Bloomberg.com Available at: https://www.bloomberg.com/news/articles/2017-03-31/video-games-without-the-players (Accessed 19 December 2020).

Kaliappan, J. and Sundararajan, K., 2020. *"Machine Learning in Video Games.* Handbook of Research on Emerging Trends and Applications of Machine Learning", 2020, pp.425-443, doi: 10.4018/978-1-5225-9643-1

Kunanusont K, Lucas SM, Perez-Liebana D (2017). General Video Game AI: Learning from screen capture. DOI: 10.1109/CEC.2017.7969556

Milani, S., Topin, N., Houghton, B., Guss, W., Mohanty, S., Nakata, K., Vinyals, O. and Kuno, N., 2020. "Retrospective Analysis of the 2019 MineRL *Competition on Sample Efficient Reinforcement Learning",* viewed 19 December 2020 <https://arxiv.org/abs/2003.05012>

Nair, A., McGrew, B., Andrychowicz, M., Zaremba, W. and Abbeel, P., 2018." Overcoming Exploration in Reinforcement Learning with Demonstrations", *2018 IEEE International Conference on Robotics and Automation (ICRA),* 2018, doi: 10.1109/ICRA.2018.8463162

Orland, K., 2013. How Forza 5 and the Xbox One Use the Cloud to Drive Machine-Learning AI. *Ars Technica.* Available at: https://arstechnica.com/gaming/2013/10/how- forza-5-and-the-xbox-one-use-the-cloud-to-drive-machine-learning-ai/ (Accessed 19 December 2020)

Perez D, Liu J, Abdel Samea Khalifa A et al. (2019). General Video Game AI: a Multi-Track Framework for Evaluating Agents, Games and Content Generation Algorithms. DOI: 10.1109/tg.2019.2901021

Reynard, M., Kamper, H., Engelbrecht, H. and Rosman, B., 2020. "Combining primitive DQNs for improved reinforcement learning in Minecraft". *2020 International SAUPEC/RobMech/PRASA Conference, 2020*, doi: 10.1109/SAUPEC/RobMech/PRASA48453.2020.9041025

Riedmiller, M., Hafner, R., Lampe, T., Neunert, M., Degrave, J., Van de Wiele, T., Mnih, V., Heess, N. and Springenberg, T., 2018. "Learning by Playing – Solving Sparse Reward Tasks from Scratch", *Google Deepmind*, viewed 21 May 2020, https://arxiv.org/abs/1802.10567

Scheller, C., Schraner, Y. and Vogel, M., 2020. "Sample Efficient Reinforcement Learning through Learning from Demonstrations in Minecraft". *NeurIPS2019 Competition & Demonstration Track,* viewed 19 December 2020 <https://arxiv.org/abs/2003.06066>

Silver, D., Guez, A. and Hasselt, H., 2016. "Deep Reinforcement Learning with Double Q- learning", *Association for the Advancement of Artificial Intelligence*, viewed 19 December 2020,<https://arxiv.org/abs/1509.06461>

Star Wars: *The Old Republic* Available at: https://www.swtor.com/ (accessed 5.10.21)

Sutton, R., Bach, F. and Barto, A., 2018. *Reinforcement Learning*. Massachusetts: MIT Press Ltd, ISBN: 978026203924

Velardo, V., 2019. *The Reality of Red Dead Redemption 2'S AI (Part 1).* Medium. Available at: https://medium.com/the-sound-of-ai/the-reality-of-red-dead-redemption-2s- ai-part-1-c276e9da2763 (Accessed 20 December 2020).

Walsh, A., 2018 *Forza Motorsport 7 Drivatars Being Improved, Race Adjudication Pushed To 2019* – Fullthrottle Media. Available at: Fullthrottlemedia.co.uk (Accessed 19 March 2020)

Xu, L. and Chen, Y., 2019. A Hierarchical Approach for MARLÖ Challenge. *2019 IEEE Conference on Games (CoG*), 2019, doi: 10.1109/CIG.2019.8847943

Virtual Training Environment for Gas Operatives: System Usability and Sense of Presence Evaluation

Asghar, Oche A Egaji, Luke Dando, Mark G Griffiths and Phil Jenkins

University of South Wales and GATC Ltd.

ikram.asghar@southwales.ac.uk, alexander.egaji@southwales.ac.uk, luke.dando@southwales.ac.uk

Training of gas operators in real-life settings often has associated risks to health and property. The use of a virtual environment to train gas operators has the potential to offer risk-free training. This study tests the usability of a virtual environment specifically designed to teach new gas operatives in near real-life scenarios. Thirty-two participants tested the virtual environment and performed different tasks required to complete gas safety checks. We used SUS (System Usability Scale) and sense of presence questionnaires to collect data from these participants. The SUS analysis indicated that most participants belonging to a different gender, age, and virtual reality experience groups were comfortable in the VR training environment. The sense of presence data analysis also confirmed similar results as all sense of presence factors scored high regardless of the demographics characteristics of the participants. However, there is still a need to add different scenarios to make the virtual environment into a comprehensive training course.

Virtual environment; Training; Gas operative; Sense of presence; System usability

1. Introduction

The recent statistics show that carbon monoxide (CO) poisoning has resulted in thousands of incidents. These incidents have caused more than 697 deaths inside the UK in the last two decades (Safety, 2018). The biggest reason for CO emissions is improper installation of gas ovens, boilers, and cookers. Lack of maintenance of such appliances also contributes to such incidents.

Usually, in gas leak incidents, the gas operatives are called in to fix such situations. Therefore, proper training of gas operatives is imperative to avoid any causalities. However, training new gas operatives in real-world settings has its associated risks. Developing a full-scale gas leak training scenario will not only come at a high cost but will have associated risks to life and property as well. Considering these factors, the researchers have previously advised incorporating virtual reality (VR) as a potential solution to reduce these risks to life and property (Asghar, Egaji, Dando, Griffiths, & Jenkins, 2019). Additional research has also shown that incorporating VR in traditional training can provide a more interactive environment for the trainees and positively impact their skill set and knowledge retention (Fredricks, Blumenfeld, & Paris, 2004).

In literature, we can find many examples of virtual environments to help train people in risk-free environments. Examples include VR training for firefighters in fire evacuation (Cha, Han, Lee, & Choi, 2012). A simulator for tunnel fire evacuation (Ronchi et al., 2015). VR-based emergency rescue training system for railway cranes operators (Xu et al., 2018). VR application for training children to cross railway crossings safely (Dando, Asghar, Egaji, Griffiths, & Gilchrist, 2018). Another VR application to train miners in mine incidents (Kizil & Joy, 2001). These examples motivated us to develop a virtual environment for safety training of the gas operatives explained our previous work (Asghar et al., 2019).

The current paper study aims at analysing the usability of a virtual training environment for the gas operatives. Two data collection instruments called system usability scale (SUS) and sense of presence questionnaires are used to test the usability and level of immersion of the gas operatives while trained through the virtual training environment. In total, 32 participated in the virtual environment testing across multiple sessions. Most of the participants supported the idea of the virtual environment and appreciated the skills and knowledge this system offered to them.

The rest of the paper is divided into five sections. A summary of the virtual training environment is presented in the second section. The research process used for this study is summarised in the third section. Research results and discussions are shown in the fourth section. Finally, conclusions from this study and future work recommendations are offered in the fifth section.

2. Virtual Training environment

The virtual training environment is based on two scenarios. The first scenario presents a typical residential home in virtual settings. The gas operatives are expected to explore any gas or CO emitting appliances in the virtual home and select these as potential hazardous appliance for gas leaks. The gas operatives can use teleporting

devices to enter the house. Teleporting helps them open or close the doors, select appliances, take their readings, and mark any appliance deemed dangerous. The gas operatives can select multiple appliances within one scenario. At the end of this scenario, the gas operatives can see their results in the form of a summary. The gas operatives can repeat this scenario many times. The first scenario is presented in figure 1.

Figure 1. First Scenario with Potential Actions

The second scenario requires the gas operatives to follow a storyboard of a gas leak situation, which can occur in real-home settings. In this scenario, the gas operative will have a time window of 40 seconds to follow the systematic approach from hazards' identification, inspection, and solution. The second scenario showing potential actions that a gas operative can take in a particular situation is presented in figure 2.

The gas operative will read a particular appliance, mark that appliance as hazardous or disconnect that appliance. If the gas operative cannot find the gas leak in 40 seconds or commits any mistakes in this process, there will be an explosion in the virtual training environment. This will add to the realness of dangers involved in such situations in real-life settings. Both scenarios are expected to aid gas operatives and be an excellent addition to their existing training courses.

Figure 2. Second Scenario with Potential Solutions

The current virtual training environment is kept limited to two scenarios only, as previous research has shown that excessive VR exposure can result in dizziness and sickness for some users (Viirre & Ellisman, 2003).

3. Research Process

This study followed a three-tier research strategy to accomplish the main research aim as presented in figure 3. Tier one focused on the design of the virtual training environment; tier two concentrate on testing and data collection from the gas operatives, and tier three concentrate on the analysis of the collected data. The ethical approval for the study was obtained from the Faculty of Computing, Engineering and Science, University of South Wales.

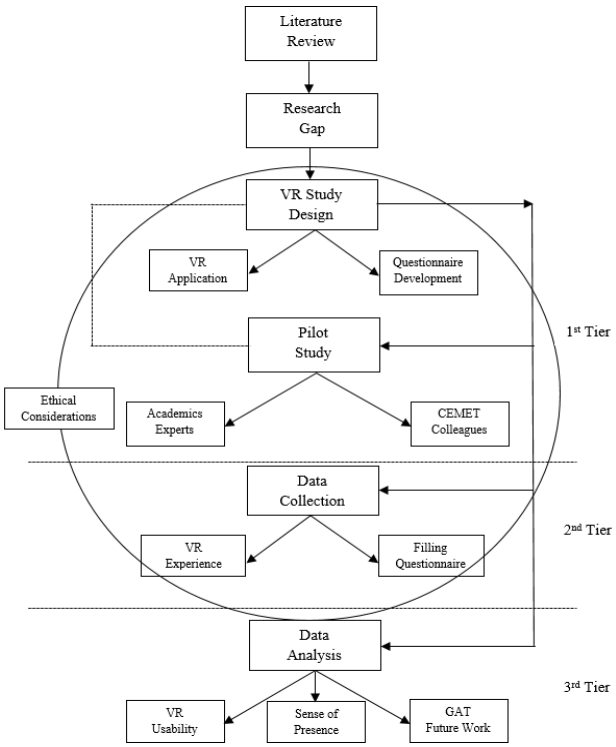


Figure 3. Research Process used for the Study

4. Results and Discussion

In total, 32 participants were trained in the virtual training environment. The training process involved each participant going through both scenarios in the virtual world and filling out both questionnaires at the end of their experience. Table 1 summarises the demographic information for the participants.

Table 9: Participants Demographic Information

Gender	*Female*	*12*	*38%*
	Male	*20*	*62%*
Age	*<= 40*	*14*	*44%*
	> 40	*18*	*56%*
Previous VR Experience	*Yes*	*20*	*62%*
	No	*12*	*38%*

We have followed the standard SUS data analysis process for data analysis and result interpretation (Harrati, Bouchrika, Tari, & Ladjailia, 2016). For the sense of presence questionnaire data, we used the Shapiro-Wilk and the Mann-Whitney U tests.

4.1 SUS Data Analysis

The average historical score for SUS is 68. A SUS score greater than 68 means that the system usability is good, and a lower than average score means there is some problem with the usability of that system (Bangor, Kortum, & Miller, 2008). For the current study, 29 participants have SUS scores of above 68, which means most participants were satisfied with the usability of the virtual training environment. Only three participants with scores of less than 68 were not fully satisfied with the usability of the virtual training environment.

Another dimension in SUS analysis is known as grading levels of SUS scores. In such research, a score equal to or greater than 80.3 means grade A, 68 and above is grade C, and equal to or less than 51 means grade F (Sauro, 2011). The average SUS score of all 32 participants who used the VR training environment is 85.31. This overall survey score corresponds to grade A, which means that the participants did enjoy using the virtual training environment, and they will recommend it to others.

Table 2 shows SUS scores based on the demographics of the participants. Interestingly, all groups scored grades A, which confirms the quality of the virtual training environment.

Table 2: SUS Scores for Demographics

Demographics	Group	SUS Final Score
Gender	*Female*	*86.67*
	Male	*83.38*
Age	*<= 40*	*87.68*
	> 40	*82.22*
Previous VR Experience	*Yes*	*85.38*
	No	*83.33*

4.2 Sense of Presence Data Analysis

For the sense of presence data collection and analysis, we have considered the work of Witmer & Singer, as they are believed to be the pioneers in this domain. Their work revolved around the theories of involvements and immersion. Based on their theoretical and empirical research work Witmer & Singer determined multiple factors, including control, sensory, distraction, and realism, contribute to the sense of presence and immersion (Witmer & Singer, 1998).

4.3 Factors Impact on Sense of Presence

The sense of presence was measured on a 7-point Likert scale. The basic descriptive statistics show mean values for control (6.16), sensory (5.97), distraction (4.06), and realism (5.44). The higher values of control, sensory, and realism indicate that sense of presence in the virtual training environment was high amongst the participants. The low value for distraction suggests a minor degree of interference by objects and actions in the virtual training environment, and participants could concentrate on the tasks. In summary, all four factors in this study contribute to a high sense of presence in the virtual environment.

4.4 Demographics Impact Sense of Presence

For the second part of the analysis, we wanted to test the impact of demographic characteristics on the presence factors. The Shapiro-Wilk test showed that all demographic characteristics have significant values ($p < 0.001$), which means the data set in this study was not normally distributed. Therefore, we have to use some non-parametric tests for testing differences among different data groups. As all demographic characteristics, including gender, age group, and previous VR experience, have only two possible values, we used the Mann-Whitney U test for further analysis (Nachar, 2008).

The results of the Mann-Whitney U test are summarised in table 3. These results show no significant values of ($p < 0.05$) for any demographic characteristics of the participants. This indicates that none of these characteristics has a significant impact on the system's usability under test. Therefore, we can conclude that all participants, regardless of their gender, age, and VR experience, enjoyed the virtual training environment as the factor mean scores indicated in the last subsection.

From the data and results for both questionnaires, we can conclude that system usability and sense of presence are highly appreciated by the gas operatives who got the chance to have hands-on experience in the virtual training environment.

Table 3: Mann-Whitney U test for Demographics

Factors		Control	Sensory	Distraction	Realism
Gender	*Mann-Whitney U*	*102.00*	*114.00*	*100.00*	*108.00*
	Z	*-1.114*	*-0.775*	*-1.856*	*-0.543*
	Asymp. Sig	*0.265*	*0.439*	*0.063*	*0.587*
Age	*Mann-Whitney U*	*123.00*	*119.00*	*112.00*	*96.00*
	Z	*-0.181*	*-0.882*	*-1.268*	*-1.326*
	Asymp. Sig	*0.856*	*0.378*	*0.205*	*0.185*
VR Experience	*Mann-Whitney U*	*118.00*	*114.00*	*100.00*	*116.00*
	Z	*-0.124*	*-0.775*	*-1.856*	*-0.181*
	Asymp. Sig	*0.902*	*0.439*	*0.063*	*0.856*

Furthermore, the usability results are slightly better than the previous study (Asghar et al., 2019). This indicates that involving end-users in the testing process and modifying the system based on their feedback results in improved system usability and sense of presence.

The study limitation includes that this system is still undergoing further development based on user feedback. Furthermore, as only 32 people participated in the testing, the results cannot be generalised and need further testing.

5. CONCLUSION AND FUTURE WORK

A virtual training environment for the gas operatives is evaluated in this paper. This virtual training environment can be integrated into existing modules of gas operatives training that can help in improving the skill set and decision-making capabilities among the gas operatives. The potential advantages of using a virtual training environment are its ability to construct close to real-life situations at less cost, risk-free training of new gas operatives, and execution of the same problem multiple times.

The virtual training environment is tested with 32 participants with the help of the SUS and sense of presence questionnaires. The average SUS score for all participants was 85.31, equivalent to a Grade A in the SUS analysis. There are no significant differences between different age groups, gender, and participants with or without previous VR experience. The sense of presence data analysis supports the results from the SUS scores reflecting high system usability and close to a real-life virtual training environment.

Future work will address some of the limitations of the current paper by testing the virtual training environment with a much larger sample size to generalise the results. In addition, more VR scenarios can add extensive details and interaction opportunities between the gas operatives and their trainers by having trainee and instructor views (multiplayer).

Acknowledgements: The authors would like to acknowledge the European Regional Development Fund (ERDF) and the Welsh Government for funding this study. We would also like to recognise the role of the Gas Assessment and Training Centre in providing expert knowledge for designing the virtual training environment. Finally, our gratitude goes to all members of the Centre of Excellence in Mobile and Emerging Technologies, the University of South Wales, for their contribution in various capacities in this study.

6. References

Asghar, I., Egaji, O. A., Dando, L., Griffiths, M., & Jenkins, P. (2019). *A Virtual Reality Based Gas Assessment Application for Training Gas Engineers.* Paper presented at the Proceedings of the 9th International Conference on Information Communication and Management.

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

Cha, M., Han, S., Lee, J., & Choi, B. (2012). A virtual reality based fire training simulator integrated with fire dynamics data. *Fire Safety Journal, 50*, 12-24.

Dando, L., Asghar, I., Egaji, O. A., Griffiths, M., & Gilchrist, E. (2018). *Motion rail: a virtual reality level crossing training application.* Paper presented at the Proceedings of the 32nd International BCS Human Computer Interaction Conference.

Fredricks, J. A., Blumenfeld, P. C., & Paris, A. H. (2004). School engagement: Potential of the concept, state of the evidence. *Review of educational research, 74*(1), 59-109.

Harrati, N., Bouchrika, I., Tari, A., & Ladjailia, A. (2016). Exploring user satisfaction for e-learning systems via usage-based metrics and system usability scale analysis. *Computers in Human Behavior, 61*, 463-471.

Kizil, M., & Joy, J. (2001). What can virtual reality do for safety. *University of Queensland, St. Lucia QLD*.

Nachar, N. (2008). The Mann-Whitney U: A test for assessing whether two independent samples come from the same distribution. *Tutorials in quantitative Methods for Psychology, 4*(1), 13-20.

Ronchi, E., Kinateder, M., Müller, M., Jost, M., Nehfischer, M., Pauli, P., & Mühlberger, A. (2015). Evacuation travel paths in virtual reality experiments for tunnel safety analysis. *Fire Safety Journal, 71*, 257-267.

Safety, C.-G. (2018). CO-Gas Safety's Statistics on Deaths and Injuries. Retrieved from http://www.co-gassafety.co.uk/data/

Sauro, J. (2011). Measuring usability with the system usability scale (SUS).

Viirre, E., & Ellisman, M. (2003). Vertigo in virtual reality with haptics: case report. *CyberPsychology & Behavior, 6*(4), 429-431.

Witmer, B. G., & Singer, M. J. (1998). Measuring presence in virtual environments: A presence questionnaire. *Presence, 7*(3), 225-240.

Xu, J., Tang, Z., Yuan, X., Nie, Y., Ma, Z., Wei, X., & Zhang, J. (2018). A VR-based the emergency rescue training system of railway accident. *Entertainment Computing, 27*, 23-31.

Paper session 4: Value-based HCI. Session chair: Matthias Laschke

Digital mobility services: A population perspective

Joy Goodman-Deane, Jakob Kluge, Elisabet Roca Bosch, Nina Nesterova, Mike Bradley and P. John Clarkson

University of Cambridge, Institut für Zukunftsstudien, Universitat Politènica de Catalunya and Breda University of Applied Sciences

jag76@cam.ac.uk, j.kluge@izt.de, elisabet.roca@upc.edu, Nesterova.N@buas.nl, mdb54@cam.ac.uk, pjc10@cam.ac.uk

Digital mobility services have great potential to increase passengers' transportation options, improve their experiences and reduce exclusion. For example, they can facilitate access to information and support, and join transport modes together more seamlessly. However, these advantages will only be available to those who can access and use these services effectively. To facilitate the development of usable and inclusive services, information on the range of potential users' digital interface capabilities, attitudes and current use of digital services is needed. A population-representative survey examining these issues was carried out with 1010 participants in Germany in 2020. As well as self-report questions, it examined basic digital interface competence using simplified paper prototyping. The results are examined in terms of the characteristics of groups that are particularly vulnerable to either digital or transport exclusion. Older people (aged 65+), people with disabilities and people with low levels of education were found to have particularly low levels of digital technology access, use, attitudes and competence. Caution is thus required when rolling out digital mobility services. Non-digital alternatives are needed to ensure an inclusive service. When digital interfaces are used, they need to be designed carefully to be usable by and reassuring to digital novices.

Inclusive Design. Digital transport. Digital exclusion. Vulnerable to exclusion groups. Older people. Disabilities.

1. Introduction

Digital mobility services have great potential to improve passengers' transportation options and experiences, offering a wide range of mobility innovations to meet changing lifestyles. For example, they can provide better access to information and support, and can help passengers to combine different types of transport modes together for a single journey. Furthermore, they can facilitate essential travel while reducing the need for direct human contact when, and for whom, this is important, for example during a pandemic or for people with communication difficulties. Examples of these services include map applications, route planners, vehicle sharing systems and ticketing and payment facilities.

However, these services will only be useful for people who can access and use them effectively. Despite steady increases in internet use in the EU, 9.5 per cent of the population have never gone online, with large differences between countries and sub-groups (European Commission, 2020). Larger numbers do not own a smartphone (Taylor and Silver, 2019). Furthermore, using the internet or a smartphone does not guarantee the ability to operate complex digital services.

This is a particular issue for digital mobility services because some of the groups that could benefit the most from improved access to transport are also at higher risk of digital exclusion. For example, there is low digital technology use among people with low education, older people and those who retired or inactive (European Commission, 2020). There is a danger that, rather than helping, digital mobility services may exacerbate the existing disadvantages for such groups.

As a result, care needs to be taken in the design of digital mobility services to ensure that they are appropriate for and can be used by these groups. To do this, it is important to understand the characteristics and needs of people in general and of vulnerable to exclusion groups in particular, considering aspects such as technology use, digital interface competence, transport needs and current use of digital mobility services.

1.1 Vulnerable to exclusion groups

An examination of the literature (e.g. Hoeke et al. 2020, Durand and Zijlstra, 2020) has identified seven groups that are more likely to be affected by digital mobility exclusion:

Older people: This group has lower levels of technology use and digital interface competence and may also experience mobility issues, capability loss and psychological constraints, such as anxiety, about falling or catching the wrong bus.

Women: Although many European countries report little gender gap in digital technology use, there are still noticeable gaps in some countries. In addition, women often have lower financial resources and different transport needs and patterns. Inherent biases and differences in attitudes towards technology also play a part.

People with low levels of education: Education attainment is correlated with a range of digital skills and hence ability to use digital mobility services.

People with low levels of income: Low income affects access to and ownership of technology devices, as well as car ownership and transport patterns.

Inhabitants of rural areas: Transport provision and needs, as well as demographic breakdown, differ between rural and urban areas. Rural areas may also lack communication infrastructure (e.g. wireless communications services).

Migrants: This group may experience barriers to technology and transport use due to language and culture. Some may also have different transportation needs.

People with disabilities: This group often experiences difficulties with transport use and may require additional information and assistance when travelling. They may also have difficulty with certain interfaces.

Previous research tends to focus on aspects that may cause and exacerbate difficulties for a particular vulnerable group. However, in reality, people belong to multiple groups. Digital division and mobility poverty should be considered as multi-layered phenomena (Kuttler and Moraglio 2020; Durand and Zijlstra, 2020).

The study described in this paper adds to this work by providing initial results from a population-representative survey of 1010 adults in Germany. It examines a range of variables of relevance to digital mobility services. This initial analysis in this paper examines the characteristics of each of these vulnerable groups defined above separately but this is merely preparatory to a more in-depth analysis of how the groups interact.

1.2 The wider project

The survey described in this paper is part of a larger research project, examining how to foster a sustainable, integrated and user-friendly digital travel eco-system that improves accessibility and social inclusion, along with the travel experience and daily life of all citizens (Dignity project, 2021).

As part of this project, a survey is being conducted in five different European countries (Belgium, Germany, Italy, the Netherlands and Spain). The surveys in some of these countries are still underway, having been delayed due to COVID-19 restrictions. This paper reports on initial results from the German survey, which was the first of these surveys to be completed.

2. Method

2.1 Overview of method

The German survey was conducted by forsa, a German independent market and opinion research institute. Participants completed the questionnaire face-to-face with an interviewer. Each interview took 20 to 30 minutes. Ethical approval was obtained from the University of Cambridge Engineering Department ethics committee.

2.2 German sample

The ADM face-to-face sampling system was used in the German survey to obtain a population-representative sample of 1010 adults. The ADM framework is a three-stage stratified random sampling design and is frequently employed in market, media and social research in Germany (Häder, 2016). After the selection of sample locations, private households and target persons within these households were selected at random using a random route procedure. At least four contact attempts were made for each target household or person. No incentives were offered to participants.

The distribution of the sample compared with that in the German population as a whole is shown in Table 1. A weighting variable was calculated to better represent the population, taking region, age and gender into account. The final column of Table 1 and all results presented in this paper use this weighting.

Table 1: Sample distribution. German population percentages come from the German census, the German Federal Statistical Office, the World Bank, UN DESA and Vuma Touchpoints, obtained through Statista (undated). Figures for education are from Statistisches Bundesamt (undated). Smartphone use in the survey refers to those who used a smartphone at least once a week. Sample percentages are given as a proportion of those who responded to the question.

Variable	Value	% in German population	% in unweighted sample	% in weighted sample
Gender	*Male*	*49.3%*	*48.4%*	*49.0%*
	Female	*50.7%*	*51.6%*	*51.0%*
Age	*16-39*	*33.3%*	*35.9%*	*33.4%*
	40-64	*41.2%*	*44.4%*	*41.3%*
	65-74	*12.0%*	*12.7%*	*15.1%*
	75+	*13.5%*	*7.0%*	*10.1%*

Location	*Urban*	*77.4%*	*71.0%*	*70.7%*
	Rural	*22.6%*	*29.0%*	*29.3%*
Technology use	*Use smartphone*	*81.7%*	*85.8%*	*81.9%*
	Do not use smartphone	*18.3%*	*14.2%*	*18.1%*
Education	*Currently attending school*	*3.6%*	*1.5%*	*2.6%*
	No school leaving certificate	*4.0%*	*1.7%*	*2.0%*
	School leaving certificate (secondary general or intermediate or equivalent)	*60.1%*	*71.1%*	*62.3%*
	University entrance qualification or higher	*31.9%*	*25.6%*	*33.0%*

2.3 Questionnaire

The survey questionnaire was adapted from a previous survey conducted in the UK in 2019 (Goodman-Deane et al, 2020). Some questions were omitted or modified based on the experiences in the UK survey and subsequent validation test. A module was added focusing on the use of technology for transport (see Section 2.3.2).

The questionnaire was developed in English and then translated into German and the other survey languages by professional translators. They were translated back into English and checked by the survey creators before adjustments were made and the translations finalised.

The questionnaire covered a range of topics as described below. Most questions were multiple-choice self-report, except for digital interface competence as described in Section 2.3.4.

2.3.1. Technology access and use

Participants were asked multiple-choice questions about their access to and frequency of use of the internet, computers, tablets and smartphones. They were then asked whether they had performed various technology activities recently. A first set of questions asked about activities in the last 3 months, and a second set examined activities that are commonly performed less frequently or relate to a deeper knowledge of technology devices, over the last 12 months. A list of activities is given in Section 3.2.

The questions about technology access and use were based on items in the Internet Access Survey 2017 (Office for National Statistics, 2017) to allow for comparison with national UK statistics. The questions were slightly abbreviated from those asked in the UK survey.

2.3.2. Use of technology for transport

Participants were asked to rate their confidence in their ability to plan an unfamiliar, local public transport journey using a computer and using a smartphone, on a scale from 1 (Not at all confident) to 10 (Totally confident). This provides an estimate of participants' self-efficacy with digital mobility services in different forms.

Additional self-report questions examined what sources participants used to obtain information about public transport, how often participants used particular digital mobility services, and whether and why participants felt limited in their regular travel within their region.

2.3.3. Attitudes towards technology

Overall attitudes towards technology were examined using the ATI (Affinity for Technology Interaction) scale. This examines "whether users tend to actively approach interaction with technical systems or, rather, tend to avoid intensive interaction with new systems" (Franke et al, 2018). The ATI scale comprises nine self-report items with a six-point response scale from "completely disagree" to "completely agree".

To explore attitudes further, some additional questions were added using the same response scale, examining aspects such as willingness to explore an unfamiliar interface and confidence in using new technology.

2.3.4. Basic digital interface competence

This module assessed participants' performance on eight basic digital interface tests using simplified paper prototyping. In each test, the participants were shown a picture of a smartphone interface on a paper showcard. An example is shown in Figure 1. The interfaces were created in English, based on those used in the UK survey, and then adapted for use in different countries with different languages and locations.

Participants were asked to indicate on the showcard what they would do to achieve a particular goal. For example, one of the goals for the interface in Figure 1 was to change the number of adults (Erwachsene) in the accommodation search (Unterkunftssuche). In some cases, achieving a goal might require several

actions. Participants were asked to indicate just the first action they would do, by indicating on the showcard. The interviewer coded each response as one of a set of predetermined options. This simplified paper prototyping method was used to keep the length and cost of the interviews down, enabling a larger sample size.

Figure 1: *Example of one of the interfaces used in the digital interface competence tests: a mock-up of a website to search for accommodation options.*

The interfaces and questions were chosen to cover a range of common, basic digital interface patterns on a smartphone: search, changing settings, creating a new event, opening a menu with more options, going back to a previous screen, activating a drop-down menu, activating an on-screen keyboard and setting favourites. As such, the tests examined a basic level of digital interface competence, rather than the capability to perform complex tasks on a digital device.

2.3.5. Other modules

Other modules examined demographics, as well as basic measures of sensory, cognitive and motor capabilities.

3. Results and Analysis

The analysis was conducted in SPSS v27 and the dataset was weighted by region, age and gender to better match the population as a whole. All the results reported in this paper use this weighting.

For brevity and clarity, this paper reports on selected summary results, calculated from the responses to individual questions. These were selected to cover key aspects covered in the survey.

Significance testing was conducted using Mann-Whitney U tests, comparing each vulnerable group against the rest of the survey sample on each variable of interest. Because the results were weighted, some of the frequency counts were non-integer and had to be rounded to the nearest integer for the analysis. The significance threshold was adjusted to $p<0.007$ using Bonferroni correction because multiple tests were performed on each variable. Note that, due to the size of the survey sample, differences may be statistically significant but small in magnitude. Due to space and the preliminary nature of the analysis, effect size is not analysed in this paper.

3.1 Definitions of vulnerable to exclusion groups

The results were examined for each of the groups identified as being particularly vulnerable to digital mobility exclusion (see Section 1.1). These groups, their proportions in the survey sample and their definitions are given below:

- Older people (25.3% of sample): those aged 65 and over.
- Women (51.0%): those giving their gender as female.
- Low education (32.2%): those listing their highest level of education as secondary general school-leaving certificate or below. This roughly corresponds to ISCED levels 0-2 (Eurostat, undated). Those currently attending general school are not included in this group.
- Low income (14.1%): those with a net monthly household income below a poverty line of 1040 euros for a single-person household (Statistisches Bundesamt, 2021). The poverty line for multi-person households was calculated from this based on the OECD-modified household size (OECD, undated).
- Rural inhabitants (28.3%): those living in a postal code in an area identified as “rural distinct with some densification” or “sparsely populated rural district” according to the official classification from the Federal Office for Building and Regional Planning (BBR). Note that this definition means that some people who are counted as rural may live in small towns.
- Migrants (9.7%): those who did not acquire German citizenship at birth. This includes both those who acquired it later and those who are not German citizens.

- People with disabilities (15.8%): those reporting being “very limited” in their daily activities due to issues with their eyesight, hearing, hands, mobility, reach, memory or concentration.

Note that these groups are not independent. In particular, the vast majority (85.5 per cent) of those reporting a disability were aged 55 and over, with 67.8 per cent of them aged 65 and over.

3.2 General technology access and use

The survey examined whether participants had access to various kinds of technology. The results for the different groups are shown in Figure 1.

Ownership of “any mobile phone” was generally high, with the lowest level being 89 per cent among people with disabilities. Access to tablet devices was the lowest, with only 42 per cent of the sample as a whole having access to a tablet. This was also very varied, with only 18 per cent of older people having access to one.

Access varies between groups. Older, low education, low income and disability groups had significantly lower rates of access than the rest of the sample on all these technologies (Mann-Whitney, $p<0.007$). The other groups did not differ significantly on any of these variables. These lower levels of access were particularly pronounced for digital technologies (i.e. excluding “any mobile phone”) among older people and those with disabilities.

Participants were also asked about their technology experience (see Section 2.3.1). A summary variable was created to represent the total number of activities performed recently out of the following 18: e-mail, voice/video internet calls, social media, online news, internet search, finding information about goods/services, buying goods/services, internet banking, booking travel, mapping applications, moving/copying files, moving files between devices, installing software on a computer, installing apps on a smartphone/tablet, changing settings, word-processing, editing photos, video or audio, and writing code.

The results are shown in Figure 2. For presentation purposes, the number of activities were categorised into High (13-18 activities), Medium (6-12) and Low (0-5). They are presented in this order so that a longer bar for the first category represents a higher amount of technology experience. All groups, except rural inhabitants and migrants, reported significantly lower numbers of technology activities than the rest of the sample (Mann-Whitney, $p<0.007$).

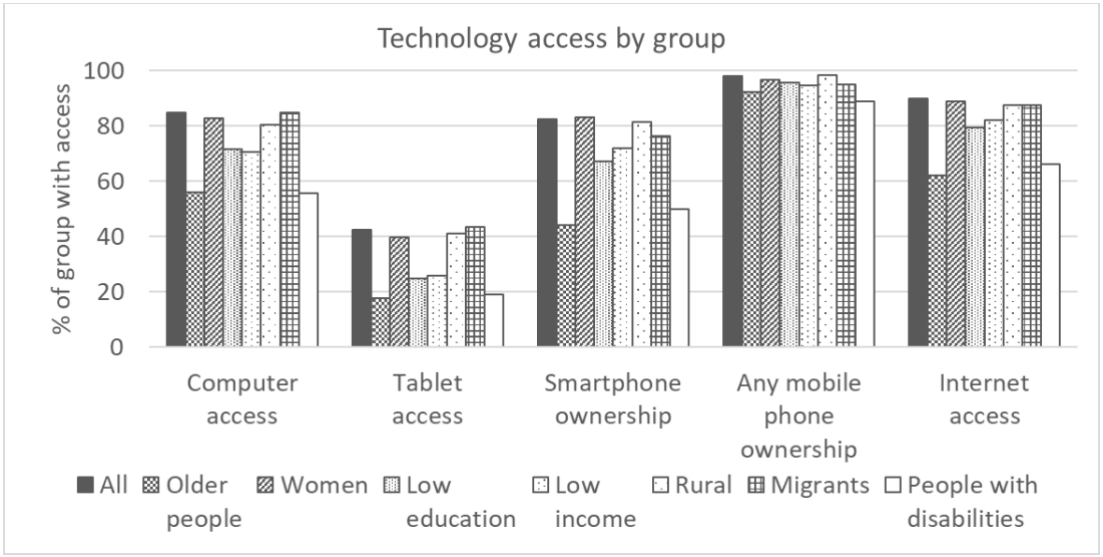


Figure 1: *Access to various digital technologies by group. Ownership of smartphones and mobile phones is used rather than general access because these are personal devices.*

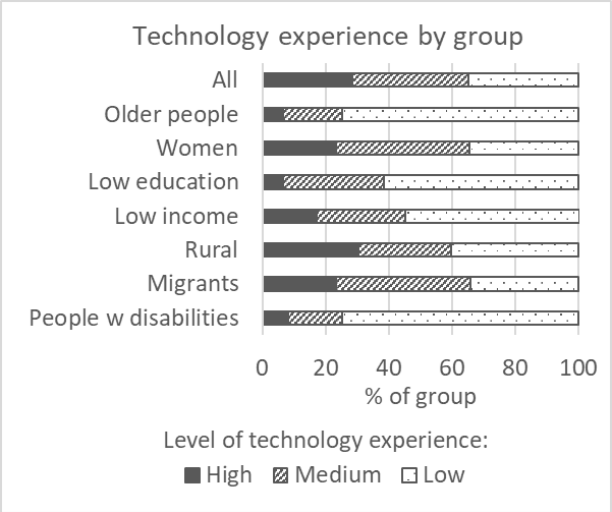


Figure 2: *Technology experience by group (based on the number of technology activities conducted recently)*

The level of technology experience varied widely between the vulnerable-to-exclusion groups, with particularly low levels amongst older people, people with low education and people with disabilities.

3.3 Use of technology for transport

The survey also examined how people obtain information about public transport, e.g. schedules, routes, cancellations and congestion. Participants chose up to three information sources. Figure 3 shows the percentage of each group mentioning any digital information source, such as websites, social media and navigation apps. A lower proportion of older, low education, low income and disabled groups used digital sources than the rest of the sample (Mann-Whitney, $p<0.007$).

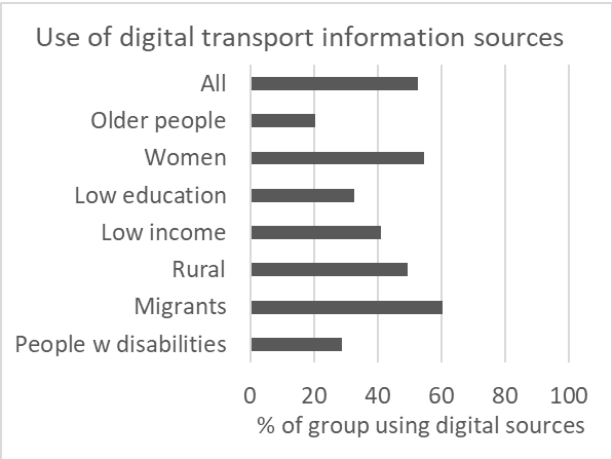


Figure 3: *Use of digital information sources about public transport by group*

The survey then asked about specific digital mobility services. The figures for the survey as a whole are shown in Figure 4. More detailed response options were used in the survey but are amalgamated into three frequency categories in the graph for visual clarity. 21 per cent of the sample had used any of these digital mobility services in the last 3 months, and 11 per cent at least once a month.

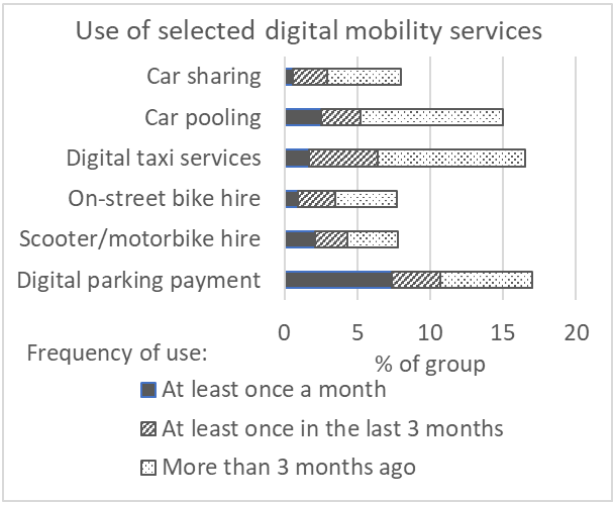


Figure 4: *Use of selected digital mobility services in the sample as a whole. There were also a small number of responses of "I don't know" (max 0.6%), with the remainder replying "Never".*

When broken down by group, the numbers for some of the services are extremely small. Thus Figure 5 examines the use of any of the digital mobility services itemised in Figure 4. There is a big variation between groups. Older people, people with low education and people with disabilities had particularly low usage of these services (Mann-Whitney, $p<0.007$).

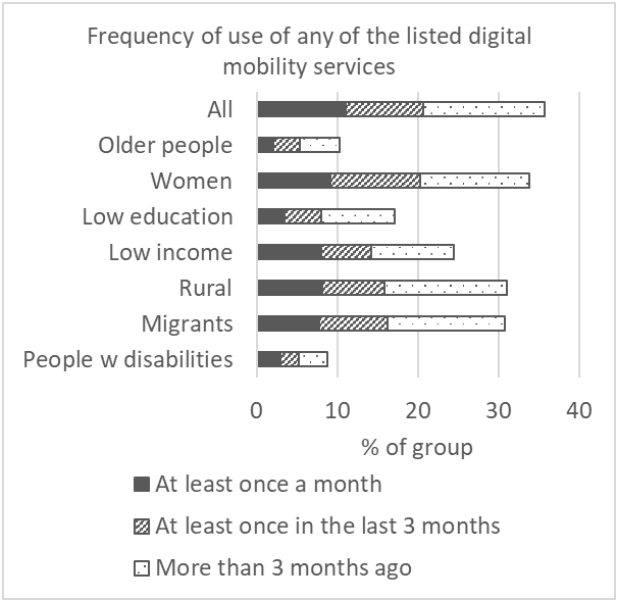


Figure 5: *Use of any of the digital mobility services listed in Figure 4 by group.*

Participants also rated their confidence in planning a local transport journey using a computer and a smartphone, as shown in Figure 6. For presentation purposes, responses were categorised into High (8-10), Medium (4-7) and Low (1-3).

All groups except rural inhabitants and migrants had significantly lower levels of confidence with both a computer and a smartphone than the rest of the sample (Mann-Whitney, p<0.007). Older people, people with low education and those with disabilities had particularly low levels of confidence.

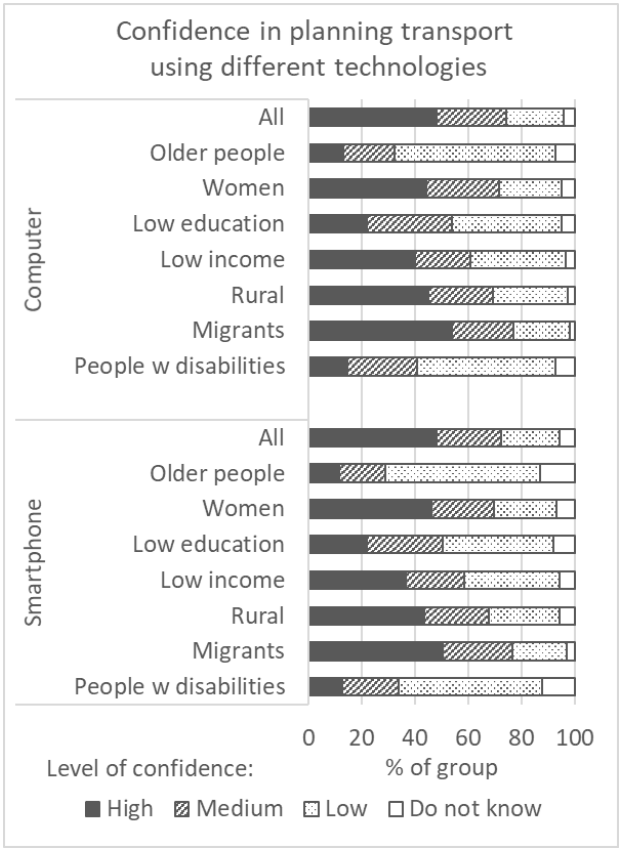


Figure 6: *Levels of confidence in planning a local transport journey using a computer and a smartphone*

This module also examined whether and why people felt very limited in their regular travel within the region. Figure 7 shows the responses overall and for reasons related to digital skills. The survey examined a range of other reasons for limitations, but this paper focuses on digital aspects.

Higher proportions of the older, female, low education and disabled groups reported feeling very limited because digital skills were needed to plan travel or use the transport (Mann-Whitney, p<0.007). The picture is different when examining limitations in travel for any reason: higher proportions of all groups except women and migrants reported these overall limitations (Mann-Whitney, p<0.007). The highest levels of limitations were experienced by people with disabilities (of whom 74 per cent felt very limited overall) and older people (65 per cent).

3.4 Attitudes towards technology

The ATI (Affinity for Technology Interaction) scale gives each person a score between 1 and 6. For presentation purposes, these are categorised into High (>4), Medium (3-4) and Low (<3). The results are shown in Figure 8.

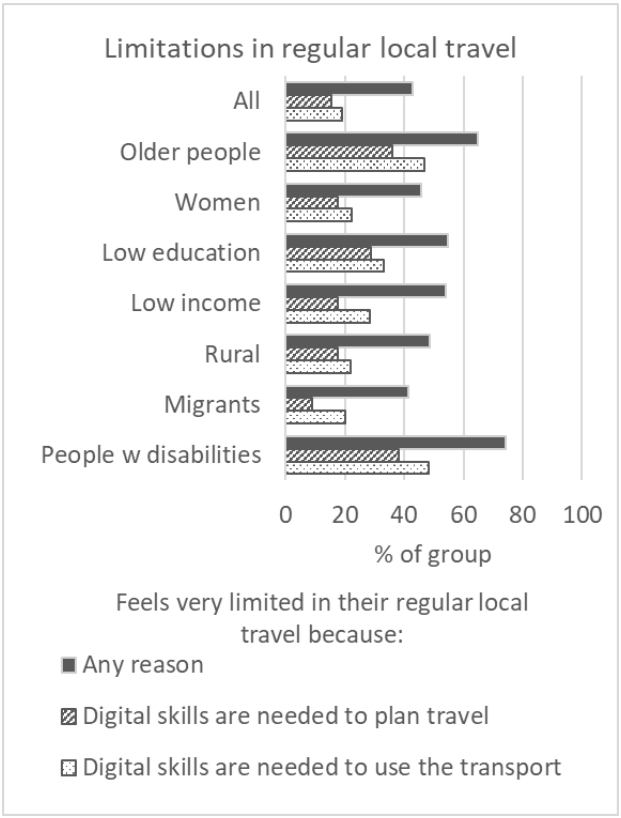


Figure 7: *Limitations in regular travel within the region, for any reason and for reasons related to digital skills.*

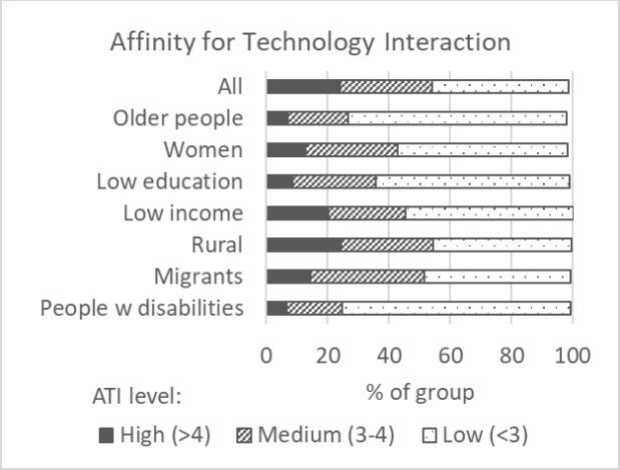


Figure 8: *Percentage of each group with Low, Medium and High ATI (Affinity for Technology Interaction) scores*

Older people, women and those with low education, and those with disabilities had significantly lower ATI levels than the rest of the sample (Mann-Whitney, $p<0.007$). Levels were particularly low among older people and those with disabilities.

3.5 Basic digital interface competence

Participants completed eight interface tests as described in Section 2.3.4. Their responses were coded into correct and incorrect, with "I don't know" coded as incorrect. The total number of tests done correctly was calculated. The total was recorded as Missing data if participants declined to do at least half of the tests. The test examined a basic level of digital competence, so the number of tests correct was categorised as described below:

- Low: 4 or fewer tests correct. We estimate that people with these scores are likely to struggle on many modern digital interfaces, particularly on smartphones and tablets.
- Medium: 5 or 6 tests correct. These people are still likely to have some difficulties
- High: 7 or 8 tests correct. a fairly high level of basic digital interface competence. This does not necessarily translate to competence with more complex interfaces and tasks.

The results are shown in Figure 9.

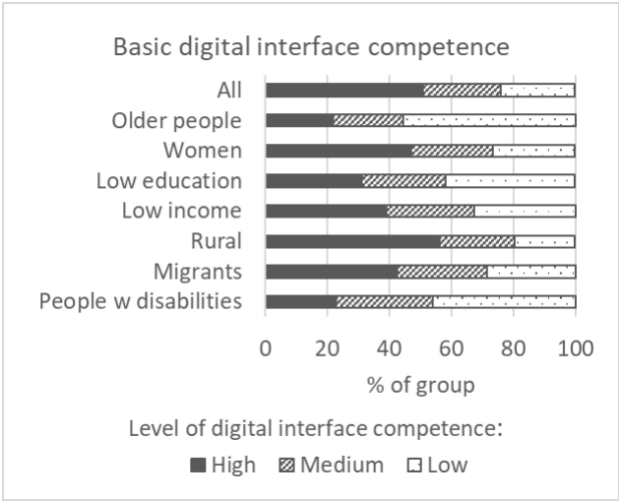


Figure 9: *Digital interface competence by group (based on the number of interface tests done correctly)*

All groups except rural inhabitants and migrants had significantly lower digital interface competence than the rest of the sample (Mann-Whitney, $p<0.007$), with particularly low levels amongst older people and people with disabilities. Rural inhabitants actually had higher competence levels than the rest of the sample (Mann-Whitney, $p<0.007$).

4. Discussion

4.1 Groups with the lowest digital technology use

Older people reported particularly low levels of digital technology use. This is consistent with previous work which identified negative correlations between age and the use of digital technology in the UK (Goodman-Deane et al, 2020). Similarly, Frid et al (2013) discussed the lower usage of technology among older people in various European countries, and Koch and Frees (2016)'s survey on internet use in Germany found that age was negatively correlated with smartphone usage. The current study extends this work, finding low levels of technology use among older people, both in general and in the context of transport. In addition, this group had low levels of access to a range of technologies, including computers, smartphones, tablets and the internet, as well as more negative attitudes towards technology and lower basic digital interface competence.

In the current analysis, the sample was divided into just two age groups in order to examine a range of vulnerable groups. However, the 65+ age group is large and very varied. Previous work has found that older groups within this range (e.g. 75+) have even lower levels of technology use and competence (e.g. Hargittai et al, 2019). This could be explored further in future work.

Another group with very low levels on all technology variables (including technology access, use,

attitudes and competence) were people with disabilities. This may be partly due to the overlap with the older age group, as 68 per cent of those with disabilities were aged 65 and over. However, the group also includes many younger people. Further analysis is needed to explore the intersectionality between these groups, and the differences between younger and older people with disabilities, and between those in each age group with and without disabilities.

This group is heterogeneous in other ways as well as age, including the range of disabilities (sensory, motor and cognitive) and the levels of severity. A range of design adaptations and accessibility features are required to meet these needs.

A third group with low levels on all technology variables is people with low levels of education. This group had slightly higher levels on some of the technology variables than older people or those with disabilities, but still much lower than the rest of the sample. For example, 42 per cent of this group had low levels of digital interface competence as measured by the interface tests, compared to 24 per cent of the sample as a whole.

4.2 Intermediate groups

People with low income and women were significantly lower than the rest of the population on some of the variables, but not on others. The size of the difference from the rest of the sample was also smaller than for the groups in Section 4.1.

People with low income had significantly lower levels of technology access, general technology use and digital interface competence than the rest of the sample. They did not differ in their general attitudes towards technology, but did have lower confidence in planning transport journeys digitally, both using a computer and using a smartphone. A higher proportion reported being very limited in travel due to difficulties during trips, but not prior to travel, because digital skills were needed.

Women reported lower levels of general technology use, attitudes towards technology and competence with technology, but some of the differences were small. They did not differ in their technology access, use of digital transport or overall limitations in travel.

Some previous studies have found gender differences in technology use (OECD, 2018), while others have not. For example, Goodman-Deane (2020b) found no significant gender differences in technology use and competence in the UK. This may be due to differences between countries. Alternatively, the larger sample size (n=1010) in the present study may have enabled detection of smaller differences between groups.

4.3 Groups with highest technology levels

At the opposite end of the scale, the survey found that migrants were similar to the sample on all the variables. Rural inhabitants differed only in two variables: a higher proportion of them reported being very limited in travel, but not for reasons related to digital skills. Furthermore, this group had a significantly *higher* level of digital interface competence than the rest of the sample.

The survey results thus indicate that migrants and rural inhabitants in Germany, considered as groups as a whole, are not at greater risk of digital mobility exclusion on the grounds of general technology access, experience, attitudes and competence. However, they may still have specific needs when it comes to other aspects. For example, the survey did not consider language issues, which are likely to be a particular concern for migrants. The survey did find that rural inhabitants reported greater limitations in transport. Transport needs are different in rural and urban areas, due to the increased distances, logistical issues with transport provision and differences in infrastructure. These should be taken into account when considering digital mobility services in rural areas.

4.4 Transport services

The survey found low numbers using the listed digital mobility services: car sharing, car pooling, digital taxi services, on-street bike hire, on-street scooter or motorbike hire and digital parking payment. The usage of these services may have been affected by the Covid-19 pandemic. Nevertheless, 64 per cent of participants had *never* used any of these services, with even lower usage in most of the vulnerable groups, especially older people and those with disabilities. This indicates that there is still a long way to go before these services become truly mainstream. Designers and developers should not assume that potential users will know how to access or operate these services. Clear and simple explanations may be required.

Larger numbers used digital sources of information about public transport. 53 per cent of the sample as a whole, and 66 per cent of those who reported using public transport, said they used these information sources. Nevertheless, many do not use them. Digital information sources need to be provided in conjunction with other non-digital means of obtaining important transport information.

The survey also found high levels of mobility poverty, i.e. people who reported feeling limited in their regular travel within their region. 44 per cent of the sample as a whole reported feeling “very limited”. Rates within the vulnerable groups varied from 42 per cent for migrants to 74 per cent for people with disabilities. Digitalization of transport products and services is not the only reason for this

mobility poverty, but it does play a part, especially for certain groups. 51 per cent of older people, 51 per cent of people with disabilities and 39 per cent of those with low education reported feeling very limited in their travel because digital skills were needed to either plan travel or use transport. This highlights the importance of ensuring inclusivity and usability when rolling out digital mobility services.

4.5 Design implications and challenges

It is important for designers and developers to consider carefully who their potential users could be, and what the characteristics of these people are. Particular care is needed if the potential users include older people, people with low education or people with disabilities.

Many older people and people with disabilities do not have internet access (38 and 34 per cent of these groups respectively). Even more (56 and 50 per cent) do not own smartphones. In fact, 18 per cent of population as a whole do not own a smartphone. Thus, while smartphones offer great potential for transport services due to their portability and mobile internet access, they cannot be deployed alone to provide an inclusive service. It is important to offer alternatives. This is highlighted by the numbers of those who are limited in their travel because of requirements to use digital technology, e.g. to access travel information or purchase or present tickets.

One possibility is to offer telephone information and booking lines. The numbers excluded by such services are much smaller, as 98 per cent of the sample and 92 per cent of older people own mobile phones. However, even these services do not cover absolutely everyone. In particular, note that visitors to a country may not have mobile signal coverage, or the cost of using a mobile phone may be prohibitive for them.

Another possibility is to offer fixed screens or kiosks displaying information or offering functionality at stops and stations. These overcome the technology access issues, but can still result in exclusion due to the digital interface competence and attitudes of potential users.

As a result, efforts are needed to make any digital interface easier to use, whether on a web browser, smartphone or kiosk. This is particularly important to ensure that people with low digital technology experience and competence are included.

For example, users with low digital interface experience are unlikely to understand the icons, language and conventions of digital interactions. If the target group is likely to include such users, it is important to include text explanations alongside icons. Similarly, these users may be unaware of hidden digital interface conventions and controls. Examples include gestural controls such as ‘pinch to zoom’. To prevent exclusion, it is important to provide hints or tips, or offer these interactions in an alternative, more visible format, such as through a zoom button or menu option.

Other issues arise due to attitudes towards technology. Some people are scared or hesitant about using unfamiliar technology, and will not try exploring an unfamiliar interface in case they break something or perhaps buy the wrong ticket. It is thus important to provide clear reassurance and confirmation for actions. Easy and obvious ways to ‘undo’ an erroneous action also help. As well as increasing the likelihood of successful use, they provide users with reassurance that they can recover from mistakes and give them more confidence to use the system.

5. Conclusions and Further Work

This paper has presented results from a survey of 1010 people in Germany in 2020, examining various characteristics related to the use of digital mobility services: technology access, general technology use, attitudes towards technology, basic digital interface competence and the use of technology for transport. It described the characteristics of seven groups that were identified as being particularly vulnerable to either digital or transport exclusion. Older people (aged 65+), people with disabilities and people with low levels of education were found to have particularly low levels of digital technology access, use, attitudes and competence.

The survey also found large numbers of people reporting being very limited in their regular travel because of the need for digital skills to plan travel or use transport. These numbers were particularly high among older people and those with disabilities.

Caution is thus required when rolling out digital mobility services. Non-digital alternatives are needed to ensure an inclusive service. When digital interfaces are used, they need to be designed carefully to be usable by and reassuring to digital novices.

The survey described in this paper is currently underway or completed in four other European countries. An earlier version of the survey was also conducted in the UK. Further work will compare the findings between the countries to get a cross-Europe picture.

Further analysis can examine intersectionality, overlaps between groups and differences within a group, e.g. between older and younger people with disabilities, or between smaller age groups within the older population. In addition, the analysis in this paper has only given summaries of key variables. Further analysis could examine more detailed variables, e.g. response to additional questions about attitudes, such as willingness to explore an unfamiliar interface, and performance on particular

interface tests. The results could go further to inform the design of inclusive interfaces.

Further work could also compare the results from this survey with qualitative findings from interview and observational studies.

6. Acknowledgements

This research was done as part of the Dignity project which received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement number 875542. We would like to thank all the research partners on this project for their input into designing and carrying out the survey design. We are also grateful to Camelia Chivaran from the University of Campania for helping with the descriptive analysis, and to Maribel Ortego from the Universitat Politècnica de Catalunya for advice on the statistical analysis.

7. References

Dignity project (2021). https://www.dignity-project.eu/ (retrieved May 2021).

Durand, A., Zijlstra, T. (2020) The impact of digitalisation on the access to transport services: a literature review. Netherlands Institute of Transport Policy Analysis.

European Commission (2020) Digital Economy and Society Index (DESI) 2020. https://digital-strategy.ec.europa.eu/en/library/digital-economy-and-society-index-desi-2020 (retrieved Apr 2021).

Eurostat (undated). International Standard Classification of Education (ISCED) https://ec.europa.eu/eurostat/statistics-explained/index.php/International_Standard_Classification_of_Education_(ISCED)#Implementation_of_ISCED_2011_.28levels_of_education.29 (retrieved Apr 2021).

Franke, T., Attig, C., Wessel, D. (2018) A Personal Resource for Technology Interaction: Development and Validation of the Affinity for Technology Interaction (ATI) Scale. International Journal of Human–Computer Interaction, 35(6), 456-467.

Frid, L., García, A., Laskibar, I., Etxaniz, A., Gonzalez, M.F. (2013). What Technology Can and Cannot Offer an Ageing Population: Current Situation and Future Approach. In Biswas et al (eds) A Multimodal End-2-End Approach to Accessible Computing. Springer-Verlag, London.

Goodman-Deane, J., Bradley, M., Clarkson, P.J. (2020) Digital technology competence and experience in the UK population: who can do what. Ergonomics and Human Factors 2020, Stratford-upon-Avon, UK, Apr 2020. CIEHF.

Häder, S. (2016). Sampling in Practice. GESIS Survey Guidelines. Mannheim, Germany: GESIS – Leibniz.Institute for the Social Sciences. https://www.gesis.org/fileadmin/upload/SDMwiki/H%C3%A4der_Sampling_in_Practice.pdf (retrieved May 2021).

Hargittai, E., Piper, A. M., Morris, M.R. (2019). From internet access to internet skills: digital inequality among older adults. Universal Access in the Information Society, 18, 881–890.

Hoeke, L. Noteborn, C., Goncalves, M.P, Nesterova, N. (2020) Deliverable D1.1 Literature review: Effects of digitalization in mobility in society. Dignity Project. https://www.dignity-project.eu/ wp-content/uploads/2020/10/200519-D1.1-Literature-Review-Final.pdf (retrieved Apr 2021).

Koch, W., Frees, B. (2016). Dynamische Entwicklung bei mobiler Internetnutzung sowie Audios und Videos. Ergebnisse der ARD/ZDF-Onlinestudie. Media Perspektiven, 9, 418–437. https://www.ard-zdf-onlinestudie.de/files/2016/0916_Koch_Frees.pdf (retrieved Apr 2021).

Kuttler, T., Moraglio, M. (Eds.) (2020) Re-thinking Mobility Poverty: Understanding Users' Geographies, Backgrounds and Aptitudes (1st ed.). Routledge, London.

OECD (2018) Bridging the digital gender divide: Include, upskill, innovate. Available at: http://www.oecd.org/internet/bridging-the-digital-gender-divide.pdf (retreived May 2021)

OECD (undated) Adjusting household incomes: equivalence scales. https://www.oecd.org/economy/growth/OECD-Note-EquivalenceScales.pdf (retrieved Apr 2021).

Office for National Statistics (2017) Internet Access – households and individuals, Great Britain: 2017. https://www.ons.gov.uk/peoplepopulationandcommunity/householdcharacteristics/homeinternetandsocialmediausage/bulletins/internetaccesshouseholdsandindividuals/2017 (retrieved Apr 2021).

Statista (undated) Statistica. https://de.statista.com/ (accessed Apr 2021).

Statistisches Bundesamt (German Federal Statistical Office (undated) Educational attainment. https://www.destatis.de/EN/Themes/Society-Environment/Education-Research-Culture/Educational-Level/Tables/educational-attainment-population-germany.html (retrieved Apr 2021).

Statistisches Bundesamt (German Federal Statistical Office (2021) Risks of poverty have become further entrenched in Germany. https://www.destatis.de/EN/Press/2021/03/PE21_113_p001.html (retrieved Apr 2021).

Taylor, K., Silver, L. (2019) Smartphone ownership is growing rapidly around the world, but not always equally. Pew Research Centre, Feb 2019. Available from: https://www.pewresearch.org/global/2019/02/05/smartphone-ownership-is-growing-rapidly-around-the-world-but-not-always-equally/ (retrieved May 2021).

Appetite for Disruption: Designing Human-Centred Augmentations to an Online Food Ordering Platform

Louis Goffe, Shruthi Sai Chivukula, Alex Bowyer, Simon Bowen, Austin L. Toombs and Colin M. Gray

Open Lab, Newcastle upon Tyne, Purdue University, USA,

louis.goffe@newcastle.ac.uk, *cshruthi@purdue.edu*, *a.bowyer2@newcastle.ac.uk, simon.bowen@newcastle.ac.uk, toombsa@purdue.edu, gray42@purdue.edu*

Online food ordering platforms have changed how many of us purchase takeaway food. They have centralised and streamlined access, providing an opportunity for population-level dietary impact. However, they are currently not human-centred: typically providing limited functionality in support of users' values and dietary considerations; and focused on the provision of food that is broadly characterised as unhealthy. In this paper we explore a redesign of portions of *Just Eat*, an online takeaway food aggregator, building upon theoretical perspectives from public health. We conducted workshops in 2018 and 2019 to identify user behaviours and motivations then designed a human-centric web augmentation template that could disrupt platform provider behaviours and increase functionality to support users' desires and well-being. We provide a template for lightweight end-user appropriations of food ordering platforms that would enable researchers to explore how health information features could improve individual health and satisfaction, and design guidance for disruptively augmenting existing food ordering platforms (or designing new ones) to enable transparency, personalisation, and self-monitoring to empower users and improve their well-being.

Digital economy; Human-centred Design; UX design; Food; Health; Design practices; Web augmentation; Service design

1. INTRODUCTION

The UK is a nation of hot fast food takeaway lovers. The home delivery and takeaway market was valued at £7.9 billion in 2017 (Passport, 2018). Meals from independent traders often provide portions that are excessively large (Jaworowska *et al.*, 2014), and have been linked to significant public health implications including greater body mass index and greater odds of obesity (Burgoine *et al.*, 2014). Online food ordering platforms have grown rapidly in the past decade, with many chain restaurants developing their own websites and apps, such as Domino's and Pizza Hut, as well as aggregator websites such as Just Eat and Grubhub which provide users with the ability to order from a wide (predominately independent) variety of local takeaway food outlets. There has been further adoption of online food ordering platforms as a result of the COVID-19 pandemic (when takeaway food has been the only trading option for restaurants), e.g. 8-9% UK growth in first two months (Shakespeare, 2020) with similar growth worldwide (Watanabe, Omori and others, 2020; McCabe and Erdem, 2021). Just Eat, Uber Eats, Deliveroo, Grubhub, and Doordash each have millions of users, but there is currently little research that describes their role in informing our dietary choices. To date, interventions that aim to support healthier choices have been targeted at activities and practices within the food outlet (Hillier-Brown *et al.*, 2017) and often focused on calorie labelling (Bleich *et al.*, 2017). These interventions are particularly challenging to implement, resource intensive, and require the compliance and engagement with the outlet owner and/or manager (Goffe *et al.*, 2018). Due to the growing number of active users of online food ordering platforms, there is potential for population-level diet (Public Health England, 2020)and subsequent health impact. These platforms are innately user-friendly (Parker, Van Alstyne and Choudary, 2016). However this 'use' relates to enabling users to easily select and purchase takeaway food. From a human-centred design perspective, where consideration is given to how well these platforms support well-being and human flourishing (Buchanan, 2001), these platforms are lacking. Thus, there is an opportunity to go beyond consideration of usability (with narrow definitions of ‘use’) to investigate how food ordering platforms can better support well-being-related factors that influence our food choice, such as health, weight control, natural content, and ethical considerations (Scheibehenne, Miesler and Todd, 2007). A YouGov survey found that 28% of Britons who increased their takeaway usage during the pandemic are still motivated by healthy menu options and freshness of ingredients; the pandemic has not lessened the importance of supporting healthy takeaway selection.

Responding to this disconnect between user-friendly and human-centred concerns, we conducted a sequence of design research activities in 2018 and 2019 to investigate how individual goals for food ordering, including the facilitation of users' informed decision-making regarding healthy lifestyle choices, can be better supported. In framing these activities, we foregrounded health in an *a priori* manner, with the emergent views and findings on how to shape food ordering platforms grounded in our participants'

experiences. This research had three linked objectives. The first was to understand users' processes and experiences of food ordering, which we explored using an interactive workshop in which participants detailed their thoughts, feelings, and actions during food ordering and the role of nutritional information in this process. The second objective was to co-design proposals for idealised and human-centred food ordering services, which we explored in a separate workshop where participants drew upon personal experiences through supported ideation. Our third objective was to produce a technologically feasible design that embodied the idealised feature sets imagined by our participants and researchers. To pursue this third objective, we used an agile UX framework to produce a human-centred web augmentation template for the disruption of the Just Eat platform.

Our work provides an accessible illustration of how online food ordering platforms could improve their structure to support healthier choices and other human-centred considerations by providing better tools to assess food healthiness and outlet hygiene at the point of food outlet selection. We offer a design provocation in the form of a web augmentation template for how an existing platform could support users' values and signpost them to healthier options. Furthermore, we illustrate how the proposed augmentations could impact users' task flow. In doing so, we identify fundamental design aspects of platform design, stemming from opacity. In conclusion, we provide three design features that designers, researchers, and advocates for human-centred design can apply to other e-commerce sites to empower users and provide them with increased control over their platform use.

2. BACKGROUND

2.1 Takeaway Food and Health

More than a fifth of UK residents order a takeaway meal at least once per week, with peak consumption in those aged 19–29 years old (Adams *et al.*, 2015). Frequent takeaway consumption is linked to an increased mean daily energy intake (Goffe *et al.*, 2017) and it has been hypothesised that takeaway food's high energy density can override our appetite control systems and trigger over-consumption (Prentice and Jebb, 2003). To date, observational studies have focused on the geographical pattern of takeaways (Fraser *et al.*, 2010), where there is a positive association between takeaway outlet density and increasing level of area deprivation (Public Health England, 2018). This has led some to conclude that 'the frequency and types of takeaway foods consumed by socio-economically disadvantaged groups may contribute to inequalities in overweight or obesity and to chronic disease' (Miura, Giskes and Turrell, 2012). The Internet has created the ideal landscape for digital platforms to flourish and become the dominant business model, where the value is in the network of producers and consumers (Parker, Van Alstyne and Choudary, 2016). The COVID-19 pandemic has further mediated the mass transition to digital technologies, inclusive of food, both groceries (Grashuis, Skevas and Segovia, 2020) and hot fast food takeaways (Bradshaw, 2021).

2.2 Just Eat

Among takeaway platforms in the UK, Just Eat has the highest number of customers and takeaway outlets. It provides a website enabling users to order from thousands of (primarily independent) hot fast food takeaway outlets across the country. In 2018, in the UK they processed 122.8 million orders from 12.2 million active users (a respective 17.0% and 16.2% increase from 2017) (Just Eat plc, 2019d). It is not known if such platforms are increasing our consumption of takeaway food, or simply providing an alternative method of ordering. What is known, from their annual accounts (Just Eat plc, 2019d), is that Just Eat is hugely popular as it eases access to takeaway food and for many is likely to be the primary means to acquire such food.

As a public limited company, Just Eat is motivated to maximise profits. The majority of its revenue is generated from a commission of 14% plus tax charged to each food outlet on every order processed (Just Eat plc, 2020). Therefore, increasing orders from all outlets is their priority, and issues relating to the public's health do not factor unless there is negative media reporting that has the potential to impact on their revenue. The prominence of media reporting suggests that the results of an independent food safety inspection would be of interest to users to assure them that the food they are purchasing is safe to eat, but Just Eat does not present this information prominently to users. Furthermore, an investigation by the BBC found that Just Eat were routinely listing without warning takeaways that had received a food safety score of zero (Crawford, 2018), which means that urgent improvement is necessary to prevent a risk to public health (Food Standards Agency, 2018). Just Eat announced in July 2019 that it would display the hygiene rating for all businesses listed on its platform (Anon, 2019), however, this is not offered at the point of takeaway outlet selection and requires prior knowledge by the user to locate the rating. This highlights the conflict between Just Eat's stated desire to 'give our customers an amazing experience' (Just Eat plc, 2019c) and the primacy of their profit motive (Just Eat plc, 2019a).

Just Eat presents an appearance of impartiality and places the emphasis on their users' reviews to determine outlet quality. They have carefully balanced the appearance and functionality of the platform with regards to both takeaway owners and

consumers, where the primary outcome of interest is profit. As the end users are largely unaware of these design constraints, there is minimal advocacy for improved functionality and general contentment with what the service provides. To avoid loss of profit, platform development would likely focus on features related to usability such as ease of purchasing and neglect validated measures that are critical of specific takeaway outlets.

As independent researchers absent of responsibility to provide profit to Just Eat shareholders, we can take a human-centred approach to this socio-technical design challenge of going beyond the user friendliness of the takeaway transaction to support human well-being. We view the popularity of online food ordering platforms as an opportunity to identify how platforms might support healthy food choices as for many they may become their primary means to access a hot meal.

2.3 Web Augmentation for Design After Design

The desire to use products in ways that have not been anticipated can be explored through the practice of Adversarial Design (DiSalvo, 2012), whereby design processes are used to challenge the status quo. Storni identified the idea of 'empowerment-in-use', which advocates 'design after design', an application of traditional design techniques to the problem of how users might appropriate their existing technology to different uses that might not have been foreseen by the designers (Storni, 2014). Applying this philosophy, we can consider that a digital platform is simply a product that, like any other, a user might wish to adapt to better suit their needs. When considering how a product might be adapted, it is important to consider the 'seams'—those exposed areas that the user is free to change (Maccoll and Chalmers, 2003). Technology providers often remove seams, reducing the possibility of design after design—ensuring users behave more uniformly, often to limit maintenance costs. Examples of removing seams include manufacturers such as Apple or Samsung making it difficult for users to open their own phones (e.g., to replace batteries), Facebook removing RSS feeds, or Twitter closing its APIs.

When a website is viewed, the loaded web page then exists on the user's local device within the web browser at the point of interaction—creating a seam that cannot be removed and where the service provider's power to influence the user's interaction is reduced. This opportunity is exploited by the mechanism of web augmentation (Díaz, 2012; Díaz, Arellano and Azanza, 2013; Díaz *et al.*, 2014; Díaz and Arellano, 2015), in which a user's experience is modified using a browser extension or plugin to manipulate the loaded web page in order to remove, add, or modify elements of the page before the user interacts with it. Well-known examples are ad-blockers that remove unwanted banners, advertisements, or pop-ups from pages. In research, web augmentation has been used to stop clickbait (Chakraborty *et al.*, 2016), filter explicit words (Suliman and Mammi, 2017), dispute fake news (Ennals, Trushkowsky and Agosta, 2010), and to combat addiction (Pyshkin *et al.*, 2016).

Given the evidence linking exposure and consumption of takeaway food and the rapid growth of online food ordering platforms, Just Eat provides an ideal context to design web augmentations that make a platform more human-centred and support users in making choices that align to their personal values and dietary requirements.

3. RESEARCH DESIGN

Our research consisted of three sequential stages involving design activities to explore current experiences of—and design possibilities for—takeaway food ordering with consumers of such food. Stage one sought to understand participants' experiences of food ordering through a workshop. Stage two sought to generate novel, human-centred design proposals for food ordering services through a co-design workshop. Stage three synthesised findings from stages one and two into a design consisting of proposals for the web augmentation of Just Eat to empower its users, subject to the known capabilities and limitations of web augmentation. Hence, the findings from stages 1 and 2 are summarised alongside their descriptions below. Ethical approval for this study was provided by Newcastle University ethics committee (Reference: 5377/2018). All participants gave written informed consent to take part in the study.

3.1 Stage 1 (S1): Experiences of Ordering Takeaway Food

3.1.1. S1 Setting and Participants

We ran a 90-minute workshop in a Newcastle University (NU) catering outlet that was accessible to the general public, thereby immersing participants in the subject matter. Our objective was to understand and detail participants' experience and processes of ordering takeaway food. We recruited 17 participants through promotional material placed in University catering outlets and academic mailing lists. This was an appropriate target audience as young adults aged 19-29 are the most frequent consumers of takeaway food (Adams *et al.*, 2015). Participants included: six nutrition researchers, four HCI researchers, and seven regular users of University catering services (staff and students from unrelated fields). Participants were split into four groups and guided through the workshop by five facilitators—two of the authors and three user-experience (UX) design students. All participants were given a £20 online shopping voucher.

3.1.2. S1 Activities

The workshop comprised three key activities designed to deconstruct and describe the processes and experiences of food ordering. In the first activity (30 minutes), the participants recounted to their group members a personal recent experience of ordering food in either a restaurant, café, or takeaway. They were asked to consider the various steps in that journey, including what they were thinking before and when they received the food. Each group selected one experience and created a chronological user journey map (Tomitsch *et al.*, 2018) where they detailed each step with regards to what they were *thinking*, *feeling*, and *doing*. For the second activity (15 minutes), we investigated the influence of nutritional information upon participants' meal choices. Participants were asked to select and explain their most and least preferred meal from a selection of three based on an image of the food and a brief description. Following their choice, participants were presented with the following nutritional information for each meal: energy (kcal), protein (g), fat (g), and carbohydrates (g). They were then asked if they would like to revise their first choice and, if so, to state their reason for changing. In the final activity (30 minutes), each group returned to their journey map and detailed at what stage the provision of specific pieces of information, such as ingredients, nutrition, allergens, price, and food outlet reputation, would be of use during the food ordering process.

3.1.3. S1 Findings

The user journey maps included a range of emotions about ordering takeaway food. Many participants considered that takeaway food was broadly a 'treat', and that taste, as opposed to health, was their primary consideration. However, after having eaten takeaway food, they also expressed feelings of regret and a desire for nutritional information to inform their choice. Some participants changed their meal choice to one they perceived as healthier upon provision of calories and other nutritional measures.

The user journey maps consisted of three phases: pre-order outlet decision, order, and post-order reflection. The pre-order phase consisted of feelings of hunger, or those that were a direct result of their hunger, for example 'hangry' [hunger-angry]. This triggered thoughts related to the logistical issues of identifying a suitable food outlet for the individual or party who want to eat. The order phase was the point of direct engagement with the selected food outlet. This was associated with feelings that included confusion, relief, excitement, and impatience. In the post-order phase, participants considered if they were satisfied with their food and outlet choice. Here they listed feelings related to satiation, but also reflective emotions such as regret. Provision of nutritional information motivated four of the 17 participants subsequently changed their meal choice to one they considered healthier. Other participants had their choice affirmed as the nutritional information corroborated their *a priori* nutritional knowledge. As with the user journey maps, the desired point of information provision was split between the three respective phases. During pre-order, users wanted information on the type of cuisine, dietary requirements, rating and reviews, cost, and opening hours. At the point of ordering, users wanted information regarding the cost of specific dishes, distance to outlet, outlet access, ingredients, meal choice of other individuals in their party, and delivery time. Upon meal reflection, users stated a desire for nutritional information.

The ordering, particularly of takeaway food, was instinctive and driven by hunger and a craving for a particular type of food. Environmental factors, including the social setting, were key determinants of what was ordered. Most participants positioned such food as a treat, inferring infrequent consumption, thus placing greater weight on their enjoyment of such food over health considerations.

Health was not a consideration in the pre-order phase. However, during the post-order phase, once participants had had the chance to reflect, some reported feelings of regret and resentment over their food choices. Broadly, the healthiness of the food was not considered when purchasing takeaway food. Some participants reassessed their choices when forced to consider specific nutritional information within a controlled setting, opting for a healthier choice as judged by their interpretation of the presented information, suggesting such information is desirable to inform their meal choice.

3.2 Stage 2 (S2): Novel Design Proposals for Online Food Services

3.2.1. S2 Setting and Participants

A second workshop (two hours) was conducted close to a University catering outlet used by participants. The objective of this workshop was to develop idealised, human-centred design concepts for takeaway food ordering. We recruited 16 participants, including eight participants from S1, with additional recruitment through promotional material placed in University catering outlets and academic mailing lists. As with S1, participants were deemed to be an appropriate target audience (Adams *et al.*, 2015)(Adams et al., 2015) and consisted of: six nutrition researchers, seven HCI researchers, and three regular users of University catering services (staff and students from unrelated fields). Participants were split into four equal groups and subsequently merged into two groups for the final activity and guided through the workshop by three facilitators. A £20 online shopping voucher was offered in recognition of participation.

3.2.2. S2 Activities

This workshop consisted of three phases. Firstly (30 minutes), a critique of a current takeaway food service. Participants were provided with a £5 meal voucher for a NU food outlet and asked to purchase a meal that they considered healthy, but otherwise had autonomy regarding their choice. To initiate critical thinking, we asked them what they bought, why, and how they made that choice. Additionally, we asked them to mentally note what information they looked or asked for, found, or felt was missing. Within their groups, they ate their food together and discussed and detailed the positive ('gains') and negative ('pains') aspects of their ordering experience. Having drawn attention to factors that influence food choice, we moved to the second phase (30 minutes) of supported ideation where each group generated a series of novel food ordering service ideas based on ideation cards which covered the topics: who is this idea for; what is the scenario of use; what food service is being used; what are the user's requirements; and a free-choice 'wildcard' (which included futuristic/fictional technologies such as robots, A.I. or teleportation). In the final phase (50 minutes), the four groups were combined into two equal groups of eight. Here, participants filtered, refined, and combined ideas into one conceptual idealised service, per group, for potential implementation.

3.2.3. S2 Findings

An emergent theme from the ideated food ordering services was the ability of the user to easily identify food that matches one's personal values and dietary preferences. Such ideas were predicated on the availability of data such as ingredients, nutritional information, and allergens. Building upon this, tools that would enable users to monitor and regulate consumption were desirable. Other ideas expressed and developed related to strong feelings regarding the social aspect of food and communal dining.

Clustering into themes the positive and negative aspects of their food ordering experience revealed that participants liked: the availability of made-to-order food; deli customisation; quick and convenient service; availability of healthy options; friendly staff; good value food; and environmentally conscious operation. Limitations of the food outlet were: limited ingredient and nutrition information; poor outlet layout; long time to pay; potential for cross-contamination of deli options; poor complete meal deals; and restricted cuisine diversity.

The groups produced 16 novel service ideas, including preferred service options for: a robot meal delivery service; the monitoring of body functioning in relation to food and outlet experience; an exercise-consumption calorie trading platform; a teleportation service linked to smartphone location; a 'loyalty-carb' tracker to support those on a carbohydrate-restricted diet; a vegan, nutritionally-controlled portions buffet; and an app to support communal dining that accounted for each user's nutritional requirements.

For implementation, one group proposed a model for food ordering similar to the social fitness platform, Strava (Strava, 2021). The main currency would be calories and these would link to their personal level of physical activity. Users would be able to purchase meals and gain credit through the volume of exercise in which they had engaged. The social aspect would allow users to share their exercise statistics and where they had purchased their meals to create a resource of food outlets where meals could be purchased that are conducive for endurance activities. The other group presented a platform to support facilitation of social dining. It would help users to identify and coordinate a suitable food outlet where they could meet and purchase food aligned to their specific dietary preferences with their friends. It would also aim to reduce social isolation through helping users to find a 'lunch buddy' to meet for a meal.

The provision of information was critical to help support participants in relation to their food choices. They liked to know if food aligned to their personal values and dietary requirements. Takeaway food eating was greater than the simple act of satiating hunger. The social experience was viewed as highly valuable, where mealtime provides an opportunity to engage with others in a non-formal setting (Dunbar, 2017). Diversity of cuisine and food component options was seen as a benefit, but participants wanted to be appropriately sign-posted through the ordering process. There was a desire for services to provide a personalised and tailored ordering experience aligned to users' ideals, built on an increased provision of information on food options. The social concepts showed the increased value that participants place upon recommendations from those within their own networks. The novel platform ideas offered the potential to establish dietary goals. With these proposed features users could track, recall, and report dietary habits, with the potential to reward health-promoting activities. Additionally, the social aspects of the platform could provide peer support for users to achieve dietary goals, for example identifying suitable food outlets that adhere to an individual's dietary preferences.

3.3 Stage 3 (S3): Designing an Augmented Just Eat to Empower Users

3.3.1. S3 Setting and Participants

In our third stage, our objective was to design a human-centred web augmentation template that would incorporate those appropriate and applicable findings from S1 and S2. Where participants stated a desire for functionality that would easily enable them to identify food that matched their personal

values and dietary preferences as well as monitor and regulate consumption. We wanted to identify modifications with the potential to support improved well-being of Just Eat users. To this end, a design sprint was conducted over a week (5 hours a day) with a group of five UX experts and practitioners on campus at NU. The sprint was an iterative process based on the agile UX framework (Chamberlain, Sharp and Maiden, 2006), delivered by a team of one postgraduate and four undergraduate UX students from Purdue University, with experience of UX methods and qualitative research methods, and was supported by design and public health academics from Open Lab (NU) and Purdue University.

3.3.2. S3 Activities

The design sprint was divided into four definitive phases to capture the design elements currently absent on Just Eat that could be implemented through web augmentation tools to create a more human-centred platform. Firstly, a review of the internal reporting, findings, and workshop illustrations from S1 and S2 was carried out by the design researchers. Details about the workshops conducted and insights were leveraged by the design sprint members to ideate. Secondly, the design researchers delivered an affinity analysis (Hartson and Pyla, 2012) of 'customer reviews' and ratings from existing online food ordering platforms: Just Eat, Deliveroo, Uber Eats (all UK-based), and Grubhub and Doordash (US-based), as well as their Google Play Store app listings; to get access to user complaints and needs. The affinity analysis of 'negative' experiences shared through the comments resulted in broad themes such as: 1) late delivery; 2) incorrect information; 3) packaging problems; 4) requests to present hygiene ratings on the website; 5) staff's disrespectful attitudes; and 6) Complaints about Payment Options. This provided user insight and identification of a range of problems and needs at both the outlet- and platform-level, and we used these insights to create a user journey map (Tomitsch *et al.*, 2018) as shown in grey in Figure 1, to detail the existing task flow for Just Eat users. The third phase was the ideation and subsequent wireframing of web-augmentable human-centred concepts applicable to the Just Eat platform (as seen in red in Figure 1). The final phase consisted of modelling the potential impact of our proposed new features on users' task flow on Just Eat by plotting concepts on to the user journey map and running user scenarios.

3.3.3. S3 Findings: Web Augmentations of Just Eat

The S1 and S2 findings recognised users' diverse needs and requirements for online food ordering platforms. Whilst the novel ideas generated in S2 had specific functionality (though not all were applicable to web augmentation of Just Eat), they were all predicated on an increased availability of information on both the outlet and their food offering. Based upon this, the design sprint members developed their ideas using the existing nutritional information requirements and considered the different stages at which that information might be expected by the user. This information was paramount to allow for an ordering experience tailored to the user to help them select food items that were aligned to their personal values and dietary preferences. This human-centred approach, whilst not restricting choice, would support easier identification of healthier food. The affinity analysis revealed that, in terms of features, generic components included: location, delivery time, delivery updates, door delivery, pick-up service, and menu item listings.

For specific dietary considerations (e.g., allergens), this information was the responsibility of the food outlet and users were directed to contact the outlet directly. The user reviews on the food ordering platforms were food outlet specific. While this provided an understanding with the main themes that users liked or found frustrating about ordering food—for example timely delivery, poor customer service, and packaging problems—they provided little critical insight as to the functioning of the service offered by these platforms. Data aligned to our design goals was found in Google's Play Store. This highlighted users' frustration regarding lack of

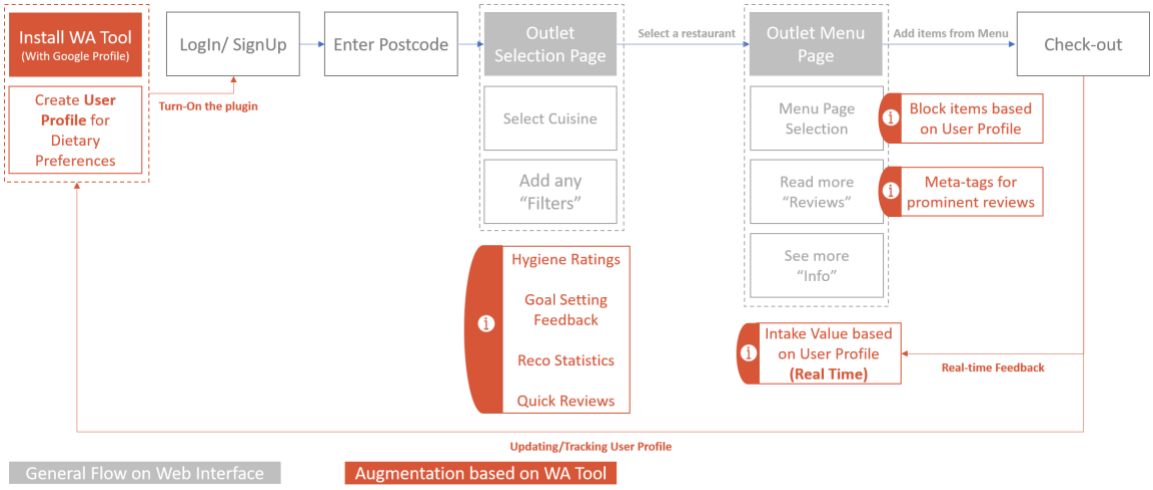


Figure 1: *Just Eat user journey map with proposed augmentations in red*

information from the platform on a food outlet's hygiene, ingredient and allergen information, and payment problems. The current users' journey map on the Just Eat platform is illustrated in Figure 1 in grey as 'General Flow on Web Interface'. This figure also illustrates the experience of the user after incorporating the designed feature augmentations subsequently detailed in Section 4.3. S3's ideation phase produced a design whereby new features could be augmented onto three specific parts of the current the Just Eat task flow: *User Profile* (which is not currently part of the Just Eat ordering flow), *Outlet Selection Page*, and *Outlet Menu Page*. The positioning of these three touchpoints, the moments where a user directly interacts with a platform, in the user's journey map (Figure 1) illustrates the sequence of the user's step-by-step interaction with the platform through these three key points of interaction.

User Profile. The *User Profile* would be an addition to a users' Google Account's existing 'Personal information' page—attaching this to Google's profile makes sense in the context of the features being delivered as a Google Chrome extension—and link to their Just Eat account. Here, as shown in Figure 2, the users would specify their personal requirements in relation to: dietary preferences, for example vegan; any allergens they should avoid; cultural or personal values, for example halal; foods disliked; and the ability to set consumption targets, for example limiting the number of weekly purchases for either health or financial reasons.

Outlet Selection Page. These are proposed modifications at the outlet selection page, which presents a list of outlets filtered by proximity to the user's specified postcode (ZIP code). As illustrated in Figure 3, this included augmenting each outlet's listing with: its hygiene rating; a personalised outlet recommendation metric based on the User Profile; keywords from outlet reviews and nutritional information, with both metrics having the associated ability to filter by value; and quick reviews, which would highlight keywords associated to a given outlet that were applicable to the user based on their User Profile. The page would also be augmented with goal setting feedback, to set limits and track their takeaway food consumption.

Outlet Menu Page. On a given outlet's menu page, a user can choose their meal items from the menu and choose to see more details about the selected outlet. As shown in Figure 4, the proposed feature augmentations could: block items not suited to the user's profile; provide real-time feedback on the nutritional information of a food item and suggest alternative options; and highlight pertinent reviews of the selected outlet by selecting meta-tags—similar to existing functionality for review filtering seen on websites such as TripAdvisor. The proposed feature augmentations are shown in red in the user's journey map in Figure 1, providing an overview of the user's augmented experience while interacting with the service platform.

4. DISCUSSION

4.1 The Role of Food Ordering Platforms to Support Healthier Choices and Human-Centred Considerations

Our design of a web augmentation template for Just Eat highlights clear modifications that could be made to the platform design to support users in making healthy, personal value-led food choices. Though our insights predate the pandemic, they are timely given COVID-19 has significantly accelerated the role that online food ordering platforms play in our food purchasing routines (Shakespeare, 2020) and diets. Our design prominently positions user-desired information at critical point of decision-making in relation to food choice. For example, the provision of an outlet's hygiene rating enables direct comparison of two or more outlets with a comparative cuisine offering. In addition to the provision of supplementary information, users could

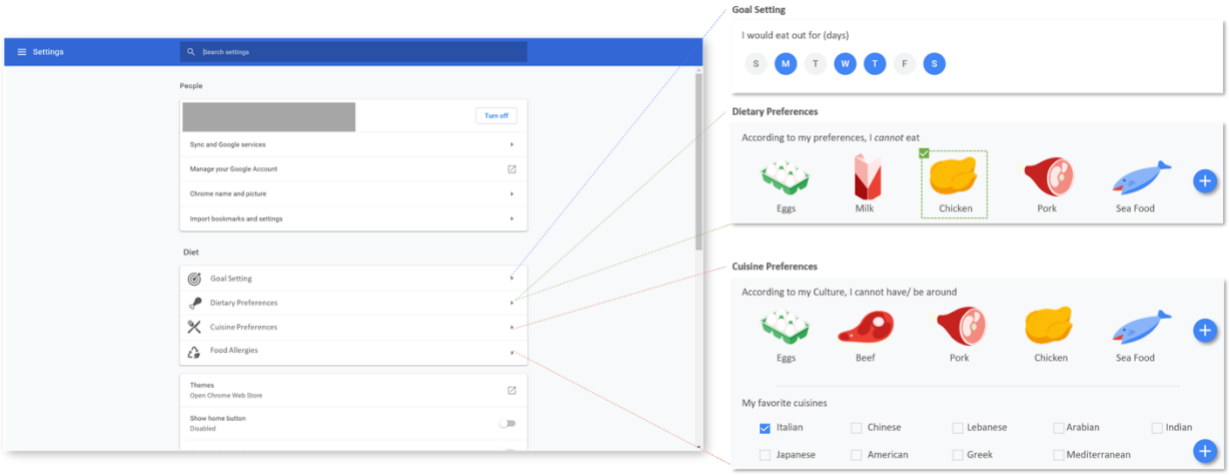


Figure 2: *User profile: proposed personal dietary preference settings*

enable filtering features that would obscure certain food choices from their view and set consumption goals, enabling them to track their progress. Such additions could potentially dampen the more instinctive and impulsive behavioural aspects of takeaway food ordering ('*Smells good, I want it*'), and support more deliberate and considered thinking and food choices ('*Does it contain too much salt?*') (Evans and Stanovich, 2013). Future implementation through web augmentation tools would be required to evaluate the value of the proposed features to Just Eat users and any behavioural impact. As the ideas originated from our user workshops and are supported by prior empirical studies on meal choice heuristics (Scheibehenne, Miesler and Todd, 2007), there is sufficient formative evidence to suggest they have the potential to constructively disrupt the established user journey on the Just Eat platform.

4.2 Pathway to Impact

Food ordering platforms are regularly used by millions of people to buy meals. In 2018, Just Eat reported almost a quarter of the UK population purchasing food from the platform (Just Eat plc, 2019a; Park, 2019), and have shown a substantial upward trajectory since 2014 (Just Eat plc, 2015). Given the accepted concept of ubiquitous computing and the rise of dark kitchens (outlets that prepare food solely for food delivery services accessed through online food ordering platforms (Butler, 2017)), computers and smart devices will most likely be our primary point of food selection for takeaway food in the near future. Therefore, the earlier we understand how these platforms influence our choices the sooner we can put in place measures that are supportive of healthier behaviours and reduce the substantial costs to healthcare and the wider economy (McPherson, Marsh and Brown, 2007). Our design provocation is a first step. It makes visible (Klein, 2000) the profit-optimised system design of Just Eat's platform, which in itself could increase agency with regards to their food choice. It provides a clear illustration that is comprehensible beyond an academic audience of what more could be done by online food ordering platforms to empower users to make healthy choices, stimulating the question "why are they *not* doing this?" However, as a design exercise we must suitably situate how it could lead to a population impact with regards to our dietary habits.

Public health research is concerned with population-level health and the translation of science into action. We see two pathways to potential public health impact stemming from our web augmentation template: (i) the ability to model and measure platform behaviour for the purposes of policy development, and (ii) to support advocates and media reporting to exemplify how platforms could be more human-centred. The first step along either pathway is the development of a browser extension. Browser plug-ins such as Takeaway Hygiene Ratings UK (Richard Hodgson, 2017) demonstrate that adding existing supplementary information such as hygiene rating using web augmentation is technologically viable. Other aspects, such as nutritional profiling, are currently not possible due to a lack of data provided by the platform and the outlets, however we have identified a feasible method to create objective outlet-level healthiness indicators through automated content analysis of available online menu text (Goffe *et al.*, 2020). While browser-side augmentation is not novel, its use in public health research is. For the first proposed pathway, such tools must go beyond provision of extra functionality and be effective and GDPR-compliant data collection tools. This would convert a commercial platform into an experimental environment to understand user behaviour. Ensuring that recruitment to future studies are appropriately stratified, they would be able to evaluate different site and extension cues and functionality to identify and understand the mechanisms that promote healthier purchases across different socio-economic and demographic groups.

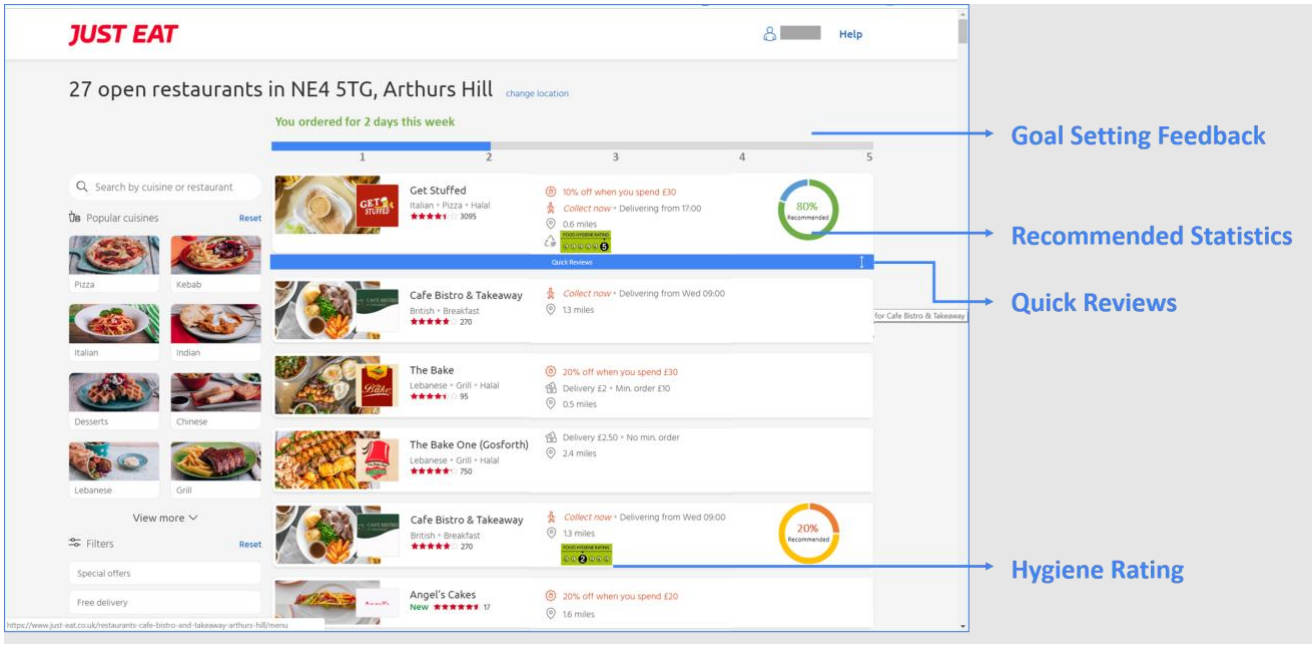


Figure 3: *Just Eat's outlet selection page: additional information presented to users that is pertinent to their outlet choice*

Researchers and policymakers would be able to trial and simulate the impact on behaviour of considered policies that targeted changes information provision such as meal calorie labelling (Department of Health and Social Care, 2018).

The second pathway using the extension would provide advocates, reporters, and action researchers, with an operational tool that shows how Just Eat, as an exemplar platform, could be used to support improved health and well-being. This could stimulate further ideas and generate similar media reports that may well have played a role in Just Eat's recent notable but low-key announcement of its decision to inconspicuously display an outlet's hygiene rating (Just Eat plc, 2019b). If either pathway is successful, this would have clear implications and applications to other digital platforms that have notable public health issues, such as gambling (Wardle *et al.*, 2019), social media (Seabrook, Kern and Rickard, 2016), and gaming (Király *et al.*, 2014).

4.3 Human-Centred Design Features for Online Food Ordering

Our work also highlights what the important design factors are in satisfying people's desires for healthy eating support tools. We have identified three key human-centred design features that would empower food platform users to support their health goals: *transparency*, *personalisation*, and *self-monitoring:*

Transparency relates to making information available, meaningful, and easily accessible to the users. People need more food and outlet information in context to inform their ordering decisions. Just Eat's lack of transparency includes using metrics other than outlet hygiene or food healthiness to order outlet listings, hiding bad user reviews and poor hygiene ratings from users and not incentivising or supporting food outlets to present ingredient, nutrition, or allergen information (see for example the Food Safety Authority of Ireland's MenuCal (Food Safety Authority of Ireland, 2019)). This lack of transparency is 'undermining trust in food' (Crawford and Benjamin, 2019).

Such transparency would enable users to judge not only whether a food option is objectively good but whether it is suitable for them specifically. *Personalisation* features would allow users to tailor the food ordering experience to their specific values. If a user were able to select a personal nutritional meta-tag such as 'low-fat' which would filter out or label high fat menu items, they would be able to judge which choices would meet their health needs, a particularly useful capability for users with dietary preferences or food allergies. As well as increasing nutritional comprehension, providing a better user experience and a better meal, this could improve data literacy by allowing the user to *use* not just read health metrics. Enabling users to reflect more deeply on their choices from a better-informed position would enable them to adopt healthier habits, which could be supported through the introduction of *self-monitoring* features - the ability for the user to track and regulate platform use to avoid excessive takeaway ordering. This could be as simple as setting a monthly target limit on takeaway orders. Such an approach has already been applied in other areas such as smartphone screen time (Zimmermann, 2021) and online gambling (where it has been shown to have a lasting positive effect on well-being (Auer, Hopfgartner and Griffiths, 2020)). A human-centred platform would support these

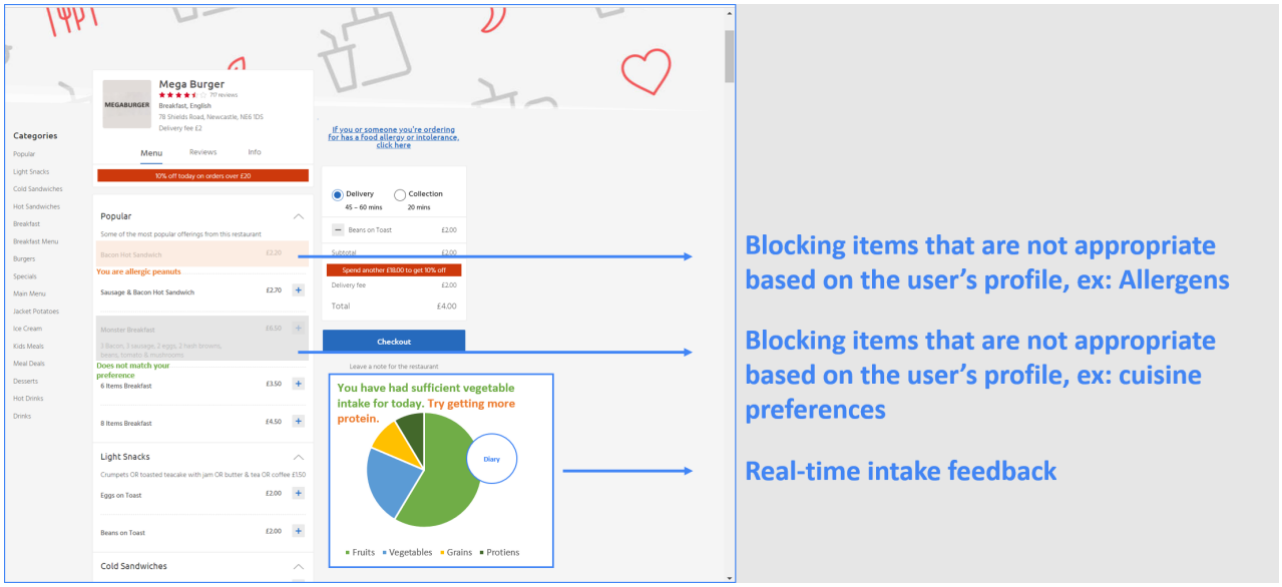


Figure 4: *Just Eat's outlet menu page: personalising menu item listing in accordance with user's dietary preferences*

healthy human values; it would respect its users by showing them all pertinent information during ordering; it would be considerate of and provide tools to avoid potential harm, even where that mean a slight loss of profit. However offered, such tools could increase agency, control, and satisfaction. If provided by the platform itself, this could increase the user's trust in the brand. The proposed design guidelines can act as *design heuristics* to build new services or augment existing platforms and as *evaluation criteria* that could be applied to other lifestyle platforms to explore their potential to improve well-being and support healthy behaviours.

4.4 Limitations and Future Work

The formative nature of this research meant that the concepts and ideals were from a limited sample size, reflecting our goals of generative engagement in the spirit of adversarial design instead of seeking generalizability. While our participants were predominately from the age group that is known to be the highest consumers of takeaway food, they did not reflect the views of all members of society who consume takeaway food (Adams *et al.*, 2015). Further research is required to implement the proposed web augmentations presented in our study to understand which design modifications would be of value to different socio-demographic and vulnerable groups. It is also important to acknowledge the intention-behaviour gap, the discrepancy in the translation of intention to action (Sniehotta, Scholz and Schwarzer, 2005). We tried to minimise this gap through the use of both food ordering and purchasing as part of S2, but at this stage in the research it is not known whether the proposed design changes would be used in practice.

5. CONCLUSION

We created a design provocation in the form of a web augmentation template for a popular online food ordering platform informed by design workshops with potential users. It exposes design elements of Just Eat that could embrace a human-centred perspective on food ordering behaviour, e.g., identifying healthly options and regulating consumption. In this work, we have shown that it is possible to design improvements to an existing food ordering platform rather than having to create and promote a new e-commerce site. Furthermore, we detail the potential pathways to impact using web augmentation technologies to convert a commercial platform into an experimental environment to evaluate behaviour change. This highlights a new mode of delivery for public health improvement research. Interventions such as our proposed feature augmentations, can surface the existing disempowerment being enacted by platforms and model how those platforms could better support users and their ideals. If such feature augmentations are preferred by platform users, this would provide evidence that could be used by health and well-being advocates and policy makers to influence service providers towards platform design that improves user well-being.

6. ACKNOWLEDGEMENTS

We would like to thank all participants for their time and personal insight. Thanks to workshop organisers Kaela Disney, Elizabeth Finley and Rhea Manocha (Objective 1) and Emma Simpson (Objective 2) and to Deanna Bell, Andrew Chu, Kevin McDonald and Leo Qu for their expertise in running the design sprint (Objective 3). Research was funded by the EPSRC Digital Economy Research Centre (EP/M023001/1) and the *'Digital civics goes abroad'* (Gray *et al.*, 2019) programme.

7. REFERENCES

Adams, J. *et al.* (2015) 'Frequency and socio-demographic correlates of eating meals out and take-away meals at home: Cross-sectional analysis of the UK national diet and nutrition survey, waves 1-4 (2008-12)', *International Journal of Behavioral Nutrition and Physical Activity*, 12(1). doi: 10.1186/s12966-015-0210-8.

Andreassen, C. S., Pallesen, S. and Griffiths, M. D. (2017) 'The relationship between addictive use of social media, narcissism, and self-esteem: Findings from a large national survey', *Addictive Behaviors*, 64, pp. 287–293. doi: https://doi.org/10.1016/j.addbeh.2016.03.006.

Anon (2019) 'Just Eat to show food hygiene ratings for all 30,000 restaurants on the app', *The Sun*. Available at: https://www.thesun.co.uk/money/9512165/just-eat-food-hygiene-ratings-all-restaurants-app/.

Auer, M., Hopfgartner, N. and Griffiths, M. D. (2020) 'The effects of voluntary deposit limit-setting on long-term online gambling expenditure', *Cyberpsychology, Behavior, and Social Networking*. Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New~…, 23(2), pp. 113–118.

Bleich, S. N. *et al.* (2017) 'A Systematic Review of Calorie Labeling and Modified Calorie Labeling Interventions: Impact on Consumer and Restaurant Behavior', *Obesity*, 25(12), pp. 2018–2044. doi: 10.1002/oby.21940.

Buchanan, R. (2001) 'Human dignity and human rights: Thoughts on the principles of human-centered design', *Design issues*. MIT Press, 17(3), pp. 35–39.

Burgoine, T. *et al.* (2014) 'Associations between exposure to takeaway food outlets, takeaway food consumption, and body weight in Cambridgeshire, UK: population based, cross sectional study', *BMJ*. BMJ Publishing Group Ltd, 348. doi: 10.1136/bmj.g1464.

Butler, S. (2017) 'How Deliveroo's "dark kitchens" are catering from car parks', *The Guardian*. Available at: https://www.theguardian.com/business/2017/oct/28/deliveroo-dark-kitchens-pop-up-feeding-the-city-london.

Carmichael, J. (2014) 'Google Knows You Better Than You Know Yourself', *The Atlantic*. Available at: https://www.theatlantic.com/technology/archive/2014/08/google-knows-you-better-than-you-know-yourself/378608/.

Chakraborty, A. *et al.* (2016) 'Stop Clickbait: Detecting and preventing clickbaits in online news media', in *2016 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. IEEE, pp. 9–16. doi: 10.1109/ASONAM.2016.7752207.

Chamberlain, S., Sharp, H. and Maiden, N. (2006) 'Towards a framework for integrating agile development and user-centred design', in *International Conference on Extreme Programming and Agile Processes in Software Engineering*, pp. 143–153.

Clement, J. (2019a) *Google's revenue worldwide from 2002 to 2018 (in billion U.S. dollars)*. Available at: https://www.statista.com/statistics/266206/googles-annual-global-revenue/.

Clement, J. (2019b) *Number of monthly active Facebook users worldwide as of 2nd quarter 2019 (in millions)*. Available at: https://www.statista.com/statistics/264810/number-of-monthly-active-facebook-users-worldwide/.

Crawford, A. (2018) 'Just Eat listings include takeaways given zero ratings for hygiene', *BBC News*. Available at: https://www.bbc.co.uk/news/uk-45888709.

Crawford, A. and Benjamin, A. (2019) 'Trust "undermined" by food delivery firms over hygiene', *BBC News*. Available at: https://www.bbc.co.uk/news/uk-48705066.

Department of Health and Social Care (2018) 'Consultation call: Calorie labelling for food and drink served outside of the home'. Available at: https://www.gov.uk/government/consultations/calorie-labelling-for-food-and-drink-served-outside-of-the-home.

Díaz, O. (2012) 'Understanding Web Augmentation', in, pp. 79–80. doi: 10.1007/978-3-642-35623-0_8.

Díaz, O. *et al.* (2014) 'End-User Browser-Side Modification of Web Pages', *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 8786, pp. 293–307. doi: 10.1007/978-3-319-11749-2_23.

Díaz, O. and Arellano, C. (2015) 'The Augmented Web: Rationales, Opportunities, and Challenges on Browser-Side Transcoding', *ACM Trans. Web*. New York, NY, USA: ACM, 9(2), pp. 8:1----8:30. doi: 10.1145/2735633.

Díaz, O., Arellano, C. and Azanza, M. (2013) 'A Language for End-user Web Augmentation: Caring for Producers and Consumers Alike', *ACM Trans. Web*, 7(2), pp. 9:1----9:51. doi: 10.1145/2460383.2460388.

DiSalvo, C. (2012) *Adversarial Design*. The MIT Press.

Dunbar, R. I. M. (2017) 'Breaking Bread: the Functions of Social Eating', *Adaptive Human Behavior and Physiology*, 3(3), pp. 198–211. doi: 10.1007/s40750-017-0061-4.

Ennals, R., Trushkowsky, B. and Agosta, J. M. (2010) 'Highlighting disputed claims on the web', in *Proceedings of the 19th international conference on World wide web - WWW '10*. New York, New York, USA: ACM Press, p. 341. doi: 10.1145/1772690.1772726.

Evans, J. S. B. T. and Stanovich, K. E. (2013) 'Dual-Process Theories of Higher Cognition: Advancing the Debate', *Perspectives on Psychological Science*, 8(3), pp. 223–241. doi: 10.1177/1745691612460685.

Food Safety Authority of Ireland (2019) 'MenuCal - helps you put allergens and calories on your menu'. Available at: https://menucal.fsai.ie.

Food Standards Agency (2018) 'Food Hygiene Rating Scheme'. Available at: https://www.food.gov.uk/safety-hygiene/food-hygiene-rating-scheme.

Fraser, L. K. *et al.* (2010) 'The Geography of Fast Food Outlets: A Review', *International Journal of Environmental Research and Public Health*, 7(5), pp. 2290–2308. doi: 10.3390/ijerph7052290.

Goffe, L. *et al.* (2017) 'Relationship between mean daily energy intake and frequency of consumption of out-of-home meals in the UK National Diet and Nutrition Survey', *International Journal of Behavioral Nutrition and Physical Activity*, 14(1). doi: 10.1186/s12966-017-0589-5.

Goffe, L. *et al.* (2018) 'The challenges of interventions to promote healthier food in independent takeaways in England: Qualitative study of intervention deliverers' views', *BMC Public Health*, 18(1). doi: 10.1186/s12889-018-5096-3.

Hartson, R. and Pyla, P. S. (2012) *The UX Book: Process and guidelines for ensuring a quality user experience*. Elsevier.

Hillier-Brown, F. C. *et al.* (2017) 'A description of interventions promoting healthier ready-to-eat meals (to eat in, to take away, or to be delivered) sold by specific food outlets in England: a systematic mapping and evidence synthesis', *BMC Public Health*, 17(1), p. 93. doi: 10.1186/s12889-016-3980-2.

Jaworowska, A. *et al.* (2014) 'Nutritional composition of takeaway food in the UK', *Nutrition {\&} Food Science*. Emerald Group Publishing Limited, 44(5), pp. 414–430.

Just Eat plc (2015) *Full Year Results*. Available at: https://www.justeatplc.com/investors/results-reports.

Just Eat plc (2019a) '2018 Full year Results', *Press release*. Available at: https://www.justeatplc.com/news-and-media/press-releases/2018-full-year-results-announcement.

Just Eat plc (2019b) 'Just Eat starts displaying official food hygiene ratings of all UK restaurants', *Press release*. Available at: https://www.justeatplc.com/news-and-media/press-releases/just-eat-starts-displaying-official-food-hygiene-ratings-all-uk-restaurants.

Just Eat plc (2019c) 'Our technology'. Available at: https://www.justeatplc.com/about-us/our-technology.

Just Eat plc (2019d) 'Results and reports'. Available at: https://www.justeatplc.com/investors/results-reports.

Just Eat plc (2020) 'Join the UK's leading food delivery provider'. Available at: https://restaurants.just-eat.co.uk/.

Király, O. *et al.* (2014) 'Chapter 4 - Problematic Online Gaming', in Rosenberg, K. P. and Feder, L. C. (eds) *Behavioral Addictions*. San Diego: Academic Press, pp. 61–97. doi: https://doi.org/10.1016/B978-0-12-407724-9.00004-5.

Klein, N. (2000) *No Logo*. Flamingo, London.

Lambton-Howard, D. *et al.* (2019) 'WhatFutures: Designing Large-Scale Engagements on WhatsApp', in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. New York, NY, USA: ACM (CHI '19), pp. 159:1----159:14. doi: 10.1145/3290605.3300389.

Maccoll, I. and Chalmers, M. (2003) *Seamful and seamless design in ubiquitous computing A Population Approach to Ubicomp System Design View project Information Environments Program View project Seamful and Seamless Design in Ubiquitous Computing*. Available at: www.equator.ac.uk.

McCabe, S. and Erdem, S. (2021) 'The influence of mortality reminders on cultural in-group versus out-group takeaway food safety perceptions during the COVID-19 pandemic', *Journal of Applied Social Psychology*, (December 2020), pp. 363–369. doi: 10.1111/jasp.12740.

McFarlane, N. (2005) ‘Fixing web sites with Greasemonkey’, *Linux Journal*. Belltown Media, 2005(138), p. 1.

McPherson, K., Marsh, T. and Brown, M. (2007) *Tackling obesities: future choices: Modelling future trends in obesity and the impact on health*. Citeseer.

Miura, K., Giskes, K. and Turrell, G. (2012) ‘Socio-economic differences in takeaway food consumption among adults’, *Public Health Nutrition*. Cambridge University Press, 15(2), pp. 218–226. doi: 10.1017/S136898001100139X.

Park, N. (2019) *Population estimates for the UK, England and Wales, Scotland and Northern Ireland: mid-2018*. Available at: https://www.ons.gov.uk/peoplepopulationandcommunity/populationandmigration/populationestimates/bulletins/annualmidyearpopulationestimates/mid2018%7B%5C#%7Dageing-number-of-over-65s-continues-to-increase-faster-than-the-rest-of-the-population.

Parker, G. G., Van Alstyne, M. W. and Choudary, S. P. (2016) *Platform revolution: how networked markets are transforming the economyand how to make them work for you*. WW Norton \& Company.

Passport (2018) ‘Fast food in the United Kingdom’. Euromonitor International. Available at: https://www.euromonitor.com/.

Prentice, A. M. and Jebb, S. A. (2003) ‘Fast foods, energy density and obesity: a possible mechanistic link’, *Obesity Reviews*, 4(4), pp. 187–194. doi: 10.1046/j.1467-789X.2003.00117.x.

Public Health England (2018) *Fast food outlets: density by local authority in England*. Available at: https://www.gov.uk/government/publications/fast-food-outlets-density-by-local-authority-in-england.

Pyshkin, E. *et al.* (2016) ‘Striving with online addiction with a self-control chrome extension’, in *2016 IEEE Symposium Series on Computational Intelligence (SSCI)*. IEEE, pp. 1–4. doi: 10.1109/SSCI.2016.7850190.

Richard Hodgson (2017) ‘Takeaway Hygiene Ratings UK’. Available at: https://chrome.google.com/webstore/detail/takeaway-hygiene-ratings/bkmnhmkibfcgcddfkgmgnecchilhbgmi.

Scheibehenne, B., Miesler, L. and Todd, P. M. (2007) ‘Fast and frugal food choices: Uncovering individual decision heuristics’, *Appetite*, 49(3), pp. 578–589. doi: https://doi.org/10.1016/j.appet.2007.03.224.

Seabrook, E. M., Kern, M. L. and Rickard, N. S. (2016) ‘Social Networking Sites, Depression, and Anxiety: A Systematic Review’, *JMIR Ment Health*, 3(4), p. e50. doi: 10.2196/mental.5842.

Shakespeare, S. (2020) *Changing Consumer Landscape: Dining and quick service restaurants*, *YouGov*. Available at: https://yougov.co.uk/topics/consumer/articles-reports/2020/05/13/changing-consumer-landscape-dining-and-quick-servi (Accessed: 23 April 2021).

Storni, C. (2014) ‘The problem of De-sign as conjuring: Empowerment-in-use and the politics of seams’, *Proceedings of the 13th Participatory Design Conference on Research Papers - PDC ’14*. New York, New York, USA: ACM Press, pp. 161–170. doi: 10.1145/2661435.2661436.

Strava (2021) ‘Strava’. Available at: https://www.strava.com/.

Suliman, N. A. B. and Mammi, H. B. K. (2017) ‘Explicit words filtering mechanism on web browser for kids’, in *2017 6th ICT International Student Project Conference (ICT-ISPC)*. IEEE, pp. 1–6. doi: 10.1109/ICT-ISPC.2017.8075322.

Tomitsch, M. *et al.* (2018) *Design. Think. Make. Break. Repeat. A handbook of methods*. Bis Publishers.

Wardle, H. *et al.* (2019) ‘Gambling and public health: we need policy action to prevent harm’, *BMJ*. BMJ Publishing Group Ltd, 365. doi: 10.1136/bmj.l1807.

Watanabe, T., Omori, Y. and others (2020) ‘Online consumption during the covid-19 crisis: Evidence from Japan’, *Covid Economics*, 38(16), pp. 218–252.

Zimmermann, L. (2021) ‘Your screen-time app is keeping track: Consumers are happy to monitor but unlikely to reduce smartphone usage’.

Appropriate Value-based ICTs in support of Frontline Peacekeepers

Lynne Hall, Samiullah Paracha and Gill Hagan-Green

University of Sunderland, UK

lynne.hall@sunderland.ac.uk, samiullah.paracha@sunderland.ac.uk, gillian.hagan-green@sunderland.ac.uk

This paper reports a mixed methods study with frontline peacekeepers that aimed to explore values in relation to effective peacekeeping and ICTs. A quantitative study and field visit identified that even in peace keeping areas with poor infrastructure there is considerable access to the Internet with ICT in regular and frequent use. 86 civilian and military peacekeepers participated in 11 focus groups that discussed potential ICT improvements and innovations for peacekeeping at a United Nations base. Analysis identified 4 horizontal themes (User Experience, Integration, Connectivity and Privacy) across 3 use contexts (work performance, personal physical safety and well-being). Core values were being safe, maintaining relationships, doing work well and being cared for by their organisation. Recommendations highlight the urgent need to deploy existing apps on everyday ICTs rather than any real requirements for innovation or significant R&D spend.

Digital Peacekeeping, ICTs for Peace, Military and Civilian Peacekeepers

1. Introduction

United Nations Peacekeeping began in 1948 and is a unique and dynamic instrument developed by the UN as a way to help countries torn by conflict to create the conditions for lasting peace (UN Peacekeeping, 2021). More than a quarter of the world's population live in fragile, violent, and conflict-stressed environments. Peacekeepers monitor and observe peace processes in post-conflict areas and assist ex-combatants in implementing the peace agreements including legitimacy, burden sharing, and an ability to deploy troops and police from around the world.

Military peacekeepers are integrated with civilian peacekeepers to address a range of mandates set by the UN Security Council and General Assembly. Although ICTs could clearly support and add value, peacekeeping suffers from several significant problems (Van Wie 2020) such as the absence of intelligence-gathering and information-processing capabilities between the field and field headquarters and the UN headquarters in New York (Salun 2019), as well as insufficient access to, and use of digital technologies (Fidler, 2015; Stauffacher et.al., 2005).

The United Nations has on average launched one peacekeeping mission a year since 1948 (Jett, 2019). Currently, 14 peacekeeping operations are underway that employ nearly 100,000 people at an annual cost of almost $7 billion (UN Peacekeeping, 2021). Despite being backed by rich and powerful countries, the UN missions have mostly failed on their mandates (Mugabi, 2021). Since, the Brahimi Report (Brahimi et al., 2000), which argues peacekeeping has to be brought into the information age, operations have used ICTs, but struggled to capture their full capabilities (Fidler, 2015). All too often when UN peacekeepers are deployed, peace is waged by primitive or obsolete methods and devices (Dorn 2021; 2016; Shaker 2015). For example, shortcomings in modern techniques of information-gathering and early warning have accounted for many failures in UN missions (Salaün, 2019; Sigri & Basar, 2014).

According to Wählisch (2019), the UN is still in the early stages of exploring data-driven and new technology–based solutions. Despite the benefits, the use of data and technology faces technical and operational challenges to support the peace process or crisis management (Garber & Carrette, 2018). Limited internet access and restrictions at UN base camps impede digital sentiment analysis or opinion mining. Data privacy has not matured for peacekeeping, which poses ethical dilemmas (Wählisch 2019). As with many sectors, the value of data – collection and usage – is only just being recognized and applied.

Field missions often lack peacekeeping simulations to help train their soldiers (Dorn & Dawson 2020). Nor have advances in monitoring and surveillance technology been leveraged significantly by the UN, resulting in a distinct disadvantage for the world body responsible for the maintenance of international peace and security. Further issues include inadequate training for UN peacekeepers to fulfil their mandates in counter-terrorism (Curran, 2016) with little know-how of digital technologies. Figure 1 briefly outlines this challenging context.

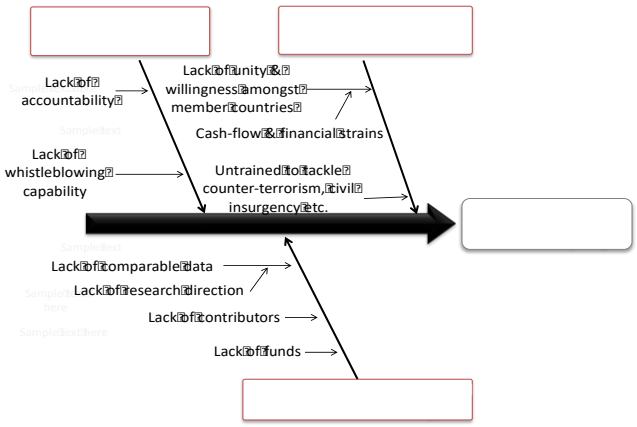


Figure 1. Fishbone Analysis: Cause and Effect

Dorn (2016) aptly remarked that the UN's power to protect depends on its power to connect. In an age when peace operations are mandated for the protection of civilians, it is essential to connect with them. Friedman & Kahn Jr, (2003) argue that digital technologies can no longer stand apart from human values which reside with the user. Value Sensitive Design focuses on questions of human welfare, security and productivity providing a pertinent approach to consider peacekeeping. Values are at play in all spheres of envisioning, designing, developing, implementing, deploying and reinvention of ICT. Values Centred Design aims at making human values a part of technological design, research and development (Friedman, 1997; Friedman & Kahn Jr, 2003; and Van den Hoven, 2007). It is a theoretically grounded approach to designing technology (Friedman, Kahn, & Borning, 2006) that brings human values to the forefront of the technical design process; providing technologists, designers, and others involved in developing technology with strategies for identifying and incorporating human values into the design and development process.

By basing future ICT inventions and interventions for peacekeepers on extant human and technical values, there is a considerably greater chance for effective systems that ultimately improve peacekeeping operations. The intention of this project was ultimately to design and co-develop novel ICT solutions for effective peacekeeping, through prioritising these technical and human values. This initial study assessed the potential to introduce technologies and concepts to the field missions and to determine what values frontline peacekeepers held in relation to effective peacekeeping and ICT design.

2. ICT FOR PEACEKEEPING

Digital technologies can enhance people's capacity to acquire truthful information (Van Wie 2020); strengthening their resilience to cope with conflict (Bertschek, Polder & Schulte (2019); illiteracy (Khan 2019); poor health infrastructure Tran Ngoc et al., (2018); and discovering means towards reconciliation (Al-Dajani 2020), community building, and empowerment (Ullah 2017). However, Shaker (2015) shows how outdated UN technology used to be by saying: "…*if villagers wanted to alert troops that they were in danger, they had to bang their pots and pans together*." According to Dorn (2021) when UN peacekeepers are deployed today, peace is waged by technologies of the 1980s or older. Advances in monitoring and surveillance technology have so far been unleveraged by the UN, resulting in a distinct disadvantage for the world body responsible for the maintenance of international peace and security.

This should not be the case in our modern globalized world with cost-effective technologies available to increase the efficiency and effectiveness of military operations so they can better achieve the ambitious mandates set out by the Security Council. Of course, innovation is not just about technology, but about people and processes as well (Dorn 2016). Ideas must percolate continually. Research and development (R&D) need to be carried out. Field testing and pilot projects complete the R&D cycle before procurement and deployment, but the UN has very little experience in researching, developing, and testing new technologies (Dorn 2016).

In considering the ways to maximize technology and innovation in peacekeeping, the Expert Panel on Technology and Innovation in UN Performance Peacekeeping (2014) has prioritized how technology could be leveraged for mandate implementation, including the protection of civilians; interoperability, as a prerequisite for effective operations; federated mission networks, to enable information sharing; medical support; camp and installation security; and mobile communications and information platforms. However, the UN is still in the early stages of exploring data-driven and new technology–based solutions (Wählisch 2019). Despite the benefits, putting data and technology to work for peace process and crisis management continues to face technical and operational challenges (Garber & Carrette, 2018). Limited internet access and restrictions at UN base camps impede digital sentiment analysis or opinion mining. Data privacy has not matured for peacekeeping, including ethical dilemmas (Wählisch 2019).

The UN field missions lack peacekeeping simulations to help train their soldiers (Dorn & Dawson 2020). As with many sectors, the value of data – collection and usage – is only just being recognized and applied. Such innovation is expected to be a game changer for peace operations (Hansen 2020). To encourage collaborative learning and innovation capacities

across peace operations, a 3-phase research design was developed.

Peace is an important value for the Human-Computer Interaction (HCI) research community, yet it has not resulted in the development of a research sub-community or even a research agenda (Hourcade & Bullock-Rest, 2011). This is due to the fact that the space technology in-habits are still being debated, and the ways in which it is and can be used for peace-building and development are in flux (Firchow et al., 2017). There is a need to understand peacekeepers, in relation to how they are supported, augmented or constrained by technology, and how this may have an impact on the way we design human computer interactions. In this paper, we seek to address this void by motivating the need for HCI research in peacekeeping space.

As the field of Human Computer Interaction has matured, an increasing trend of HCI research has concerned itself with human values (Mahamuni, Kalyani & Yadav, 2015; Borning & Muller, 2012). At the same time, a number of approaches for systematically considering human values in information technology have also emerged (Brey, 2015; Van den Hoven, 2007). A more principled approach that can clarify issues of both theory and practice is Value Sensitive Design (Friedman, Kahn & Borning, 2006), It is an established theory and method for addressing issues of values in a systematic and principled fashion in the design of digital technologies. While some projects have employed Value Sensitive Design (VSD) in the military space, there is a paucity of research applying VSD to design issues in peacekeeping.

The paper reports a preliminary study that assessed the potential to introduce technologies and concepts to the field missions and to determine what values frontline peacekeepers held in relation to effective peacekeeping and ICT design. The investigation aimed to design and co-develop novel ICT solutions for effective peacekeeping, through prioritising these technical and human values. The insights gained into important value dimensions of different peacekeepers and the subsequent value framework can help in closing the information gap between the system developer and peacekeepers by offering the relevant values that coincide with the value desired by the potential user. Thus, providing technologists, designers, and others involved in developing technology with strategies for identifying and incorporating human values into the design and development process. By basing future ICT inventions and interventions for peacekeepers on extant human and technical values, we hope there is a considerably greater chance for effective systems that ultimately improve peacekeeping operations.

The best peacekeeping research addresses, both, practical problems confronted by the peacekeepers and advances the development of scientific theory (Castro, 2003). This project was partly factual (thus practical), as it dealt with what peacekeepers were experiencing with ICTs and how they were responding or adapting (Harris & Segal, 1985) to change and technological innovation. Furthermore, it also rested on normative theory as it sought to introduce change into the existing situation, either totally or partially, in order to improve the well-being of the peacekeepers and the success of their mission (Bartone et al., 1998). These two general models apply to all successful peacekeeping research (Castro, 2003); it is impossible to recommend changes for improvement unless one knows the facts on the ground. Likewise, Maslow's Hierarchy of Needs Theory (Maslow, 1943) offered valuable insights into the inner dynamics of peacekeeping, sources of conflict, and thus possible resolutions, see table 1.

TABLE 1: KEY CHARACTERISTICS

Factual	- *Describe Reality* - *Intuition* - *Description, Categories and Classification*
Normative	- *Change or Improve Reality* - *Intuition* - *Description, Categories and Classification*
Maslow's Hierarchy of Needs	- *Physiological, Security, Information,* - *Social, Motivational*
Communication	- *Simplicity, Generality and Quant ability*
Design	- *Pragmatic, Grounded & Interactive* - *Iterative, Flexible. Integrative & Contextual*

Applying this to the specifics of ICTs entailed a communication frame that examined the: (i) channels of communication flows between the different entities; (ii) tools or platforms; (iii) spheres of activity; and (iv) functions that ICTs can play in promoting peace and preventing conflict (Communication for Peacebuilding: Practices, Trends and Challenges, 2011; Weaver and Shannon, 1963). In addition, design thinking or design theory (Brown & Wyatt, 2010) provided

guidance to collaborate with the stakeholders in order to innovate high-impact solutions, rigorous creativity and critical inquiry that bubbled up from below rather than being imposed from the top.

3. METHOD

This research was conducted with the Multidimensional Integrated Stabilization Mission in Mali (MINUSMA) established by Security Council resolution 2100 (UNSC 2013) to offer support in political and security processes, for the stabilization of the country. The mandate of MINUSMA (MINUSMA Fact Sheet, 2013) included among others: security-related stabilization tasks, protection of civilians, human rights monitoring, support to the extension of state authority in northern Mali and the preparation of free, fair and inclusive elections. MINUSMA is the fourth-largest UN operation with a personnel strength of 14,321 with 12,815 uniformed personnel and 1,342 civilian and 164 volunteer personnel (MINUSMA Fact Sheet, 2013).

The research question to be explored was ***"What values are important to frontline peacekeepers in relation to effective peacekeeping using ICTs."*** A mixed methods approach was taken including an initial quantitative survey and a MINUSMA field visit followed by 11 Focus Groups.

The quantitative survey explored demographic characteristics (age, gender, nationality), role and experience (uniformed/civilian roles, length of service, participation in prior peacekeeping missions). Participants were to be asked about their access to the internet (on base / off base / via mobile, need for internet access for work) and ICTs (generic devices and their uses (e.g. laptop, mobile) and specialist comms devices in use at MINUSMA: DECT (Digital Enhanced Cordless Tele-communications) and TETRA (Terrestrial Trunked Radio).

Focus groups were drawn from frontline peacekeepers, uniformed and civilian staff at MINUSMA. Participants were to be selected by management and engagement criteria which included being well-informed on ICT issues such as resources, ICT needs, challenges etc., as well as on the importance of digital technologies, innovation and learning in the peacekeeping space. The focus groups were semi-structured, with the following questions used to start discussions:

How do peacekeepers envision the role of ICTs in peacekeeping operations?

What are the most promising areas for innovation and experimentation in the peacekeeping space?

What are creative ways in which ICTs for peacekeeping can be designed, test deployed, experimented with and scaled?

How can UN Peacekeeping institutions be best organized for innovation and experimentation?

The Focus Groups were to be recorded and transcribed. They would be analysed using Template Analysis (King & Brooks, 2017), an approach to thematic analysis that involves the development of a hierarchical coding template from initial data analysis that can be further refined as it is applied to the full data set (Brooks et al. 2015). Template Analysis offers the following features that made it suitable for this analysis: (i) the use of initial templates and building up; (ii) lack of prescription or hierarchical coding; (iii) ability to use a priori themes; (iv) iterative focus on trying to develop the template.

Template Analysis follows a process of reading and conducting preliminary coding on a subset of transcripts and from surrounding evidence. Critically, Template analysis allows for the definition of a *priori themes* and these were based on the ICT for Peace literature and from the United Nations University's online portal - Pelikan. As the initial template is applied to the data, it can be modified and reorganized as needed through repeated readings of literature and transcripts. Quality checks are included with researchers coding independently and revising their codes with the project leader and team.

4. RESULTS

47 participants responded to the survey. Most respondents were aged between 28-42 (72.1%) and were male with only 10.6% of the sample being female. 61.7% were from Africa and 23.4% from Asia. 59.6% of respondents were uniformed peacekeepers and 40.4% civilian staff. The mean length of service at MINUSMA was 22.89 months for civilian peacekeepers and 9.39 months for uniformed peacekeepers. With longer employment, civilian peacekeepers had typically engaged in more missions than their uniformed colleagues. 95.7% of respondents had internet access via their mobile. 76.6% of respondents had access to the internet on base although only 34.04% of respondents required internet access for working purposes, with staff using facilities such as the cyber cafe. 83% were able to access the internet outside of MINUSMA.

There was good accessibility of ICTs at MINUSMA. 63.83% of respondents had laptops available to them. Laptops were used for research, work, self-development and entertainment. 29.78% used DECT and 51.06% used TETRA with significantly less usage than of the mobile phone with 93.61% using the mobile phone, clearly the most popular ICT device for peacekeepers at MINUSMA. The results from the survey highlighted that the user group demographics, particularly age, with most users between 28-42, and tech-savvy. From the results, staff can already be seen to be significant ICT and internet users. The most used and thus, presumably

the preferred device is the mobile phone. The survey also highlighted that the internet is available and device access ubiquitous with clear potential for providing innovation via ICTs with concerns about access removed.

Eighty-six participants took part in 11 focus groups. Five groups composed from the 30 civilians and six groups from the 56 uniformed peacekeepers participated in the discussions (with quotes identified as MC and MU respectively in the results). These two groups were heterogeneous in composition (incl. African, Asian, European and American) and served in different contexts and echelons of the UN missions. 10 of the civilian peacekeepers were female. All of the military peacekeepers participating were male.

As detailed in the following sections, the template analysis of the 11 sessions resulted in the identification of four horizontal themes (User Experience, Integration, Connectivity and Privacy). These were sub-themed through three contexts where staff felt that ICTs could have an impact – supporting work, personal physical safety and staff wellbeing. These results are summarised in Table 2 and further discussed with illustrative quotes below. Figure 2, at the end of the results section, presents a Venn diagram depicting Design and Human values emerging from this analysis with red arrows indicating where both categories are intertwined whereas, grey arrows represent other values important for respondents.

4.1 User Experience

Participants identified several common attributes of positive user experiences in peace keeping contexts including simplicity, small-sized, appealing, mobility, portability, automation, smart, predictive etc. The preferred device was the mobile phone, confirming the survey results.

4.1.1. Work Performance

Civilian participants highlighted that *"people use these [smart devices] for everything, official work and side by side… everything every feature I think we can do it with this smart device (MC-17)."* Military participants criticised hard to use devices comparing them unfavourably to everyday technologies *"TETRA which is very complex … it would take you 14 to 15 minutes [to set up]..... Whereas, if you buy a phone, somebody who doesn't even know how to use the phone, is able to use it (MU-19)."* Participants did recognize the value of the TETRA phone in allowing communication in emergency situations, however, it was clear that issues with its size and usability inhibited full use: *"TETRA radios*

Table 2: Summary of Results from Focus Groups

	Work/Performance	**Personal Physical Safety**	Wellbeing
Positive User Experience	• *Automation of processes* • *Ease of existing systems: COSMOS and FSS but* • *Challenges with Umoja impacting on morale and staff productivity*	• *Size and portability issues with TETRA* • *Simple devices for emergency reporting*	• *Personal development* • *Connection with family* • *Medical emergency reporting*
Integration	• *Consolidation of different platforms* • *Shared info across departments*	• *Centralized dashboards* • *Situational awareness* • *Quicker, more accurate responses*	• *Simple integrated health records* • *Crowd-sourced info for basics around the city* • *Easier onboarding*
CONNECTIVITY	• *Remote Office* • *Network issues*	• *Situational awareness* • *On the ground communication*	• *Connecting with family and friends* • *Boosting morale* • *Network issues*
Privacy	• *Access issues* • *Data sharing*	• *Cyber-security* • *Data protection*	• *Confidentiality* • *Issues with misuse of information*

are good, they are best … but, if it could still be a little bit smaller (MC-21)".

Participants, like any users, wanted a simple system that would be easy to setup and intuitive to use: "a

solution where you just have a radio with the features of a satellite instead of having all these gadgets installed (MC-19).” Automation of basic processes was proposed in all groups, from troubleshooting “you just click one button and then it executes all these commands every time and it fixes it. So, it saves 2 to 30 minutes to sending out security alerts (MC-2).” Some participants already had positive experiences of automated alert systems with one participant no longer receiving “phone calls ten times a day from ten different people for the same question (MC-22).” The medium in which alerts were provided was also raised by military participants: “When you are driving like 5km and you have this broadcast after two minutes of your departure, how are you going to read it? (MC-9)”. ICTs were highlighted in some groups as having potential for the transfer of medical information and related gains in health staff performance, with participants keen to extend this to support work: “scan the fracture in the ambulance, send … so that they'd know earlier that this case is coming and they could respond to it properly (MUG-6)”.

4.1.2. Personal Physical Safety

Several groups discussed how having simple, portable and small ICTs would enhance physical safety in risky situations. Participants provided examples of what they would prefer: “Simple powerful handsets … go on patrol everything secured … having a facility that gives us emergency like the Tetra radio frequency (MU-27).” Wearables were proposed by some groups as a potential way to enhance physical safety: “you don't even have to press it [watch] to send an alarm … it has to sense your level of anxiety (MC-27)”. A key innovation to improve physical safety was by ensuring all staff were aware of alerts, such as assaults on the base, was proposed in several groups: “everybody, disregarding level, disregarding the contract structure, disregarding the type of work should be in a single loop, to at least to be alerted at the same time these types of alert come in (MC-27)”.

4.1.3. Wellbeing

ICTs were reported as contributing to peacekeepers’ wellbeing in different ways. A key factor for staff was being able to easily and regularly connect with families and friends via ICT: “I have a little child … she gets to feel that she is actually seeing her parent, because here it is a non-family duty station (MC-27).”

Others highlighted their use of ICT for recreation and enjoyment: “For me mobile phone is my fun ICT (MC-14).” Participants also highlighted the potential for ICTs to support self-development: “with new technologies’ scope we never stop learning … improving personal development (MU-23).” ICTs were also viewed as having a key role in supporting staff health and welfare: “it would be really cool, if we had an ICT system that cares about me as an individual... and even if I'm moving across missions, this is the information that the doctors have (MC-27).”

4.2 Integration

Integration, or rather the lack of integration, emerged as a key theme, with a need for systems to be centralized, for information to be shared and applications consolidated.

4.2.1. Work Performance

The productivity and holistic performance of individuals and the organization can be improved through an increase in system integration. The recent launch of Umoja system, designed to help the UN Missions streamline recordkeeping, workflow, and communications among its myriad departments, infuriated some participants who were struggling to master the complex system: *“who says its easy or user-friendly? Rather, I’d say it is fundamentally a flawed system... It is damaging both our morale and productivity (SUB-2).” “I find Umoja system non-intuitive, labour-intensive and full of glitches and distractions (NICTP-3).”* The majority of civilian participants who used back-office systems raised issues related to a lack of integration of basic information, requiring duplication of effort. Participants identified that integrated facilities would improve the productivity and efficiency of the UN mission. Participants explained how different offices often use different applications so when people move from duty station to another: *“they struggle to again learn another application to do the same job (MC-20).”* Groups also expressed frustration over the current on-boarding process: *“your information is supposed to be transferred but you are still required to do all your input again. There is lot of forms and you have to fill it up, have to scan it, you have to load it (MC-21)”.*

Some groups discussed the need for integration between departments, *“so the work of human rights could be integrated with the work of political affairs and this means sharing information on you know just basic [information] (MC-35)”.* However, in addition to the lack of information sharing, there was a recognition that participants were unaware of what was available: *“we don’t have a comprehensive solution because we don’t understand the problem because we can’t see even what we have (MU-15).”* Were staff able to share information and data participants felt that it would enhance their problem-solving capabilities at work: *“Imagine the power … quantifiable to say that this area needs more police assistance, more guys going to help them or I have more human rights violations; maybe I need more legal aid (MU-15).”*

4.2.2. Personal Physical Safety

Many participants were of the view that integrated data and access to real-time data can improve personal physical safety: *"we need a tool that help us integrate information as quickly as possible and make us able to synchronize (MU-23)."* Proposals included centralised dashboards to highlight incidents and show no-go zones or to track the current situation of a convey and *"in case of incident they would send feedback automatically and alert the respective departments involved (MC-24)."* Participants had seen and were positive about existing systems: *"[Track 24] can be integrated into whatever system that personnel are given, what vehicles are given. It alerts you, okay you are in the no-go zone, better get out of there (MC-7)."* Similarly, proposals for planning and making UN security information more accessible included an app that was *"a map and then the actual security of UN updates - information saying that this particular road is off limits today (MC-14)."*

4.2.3. Wellbeing

Groups also discussed ways in which integrated systems could help with common issues they face that relate to but are not exactly work. Several expressed the need for consolidated medical records and a system to help keep track of appointments and medical history: *"having a kind of repository let's say with all the[health] information that the person [can] carry with him also (MC-24)."*

Another common suggestion from participants was a mobile application or resource centre that could easily crowd-source information from the staff on basics for life in the city: *"this kind of system to support the human life, if it can be developed it would be great (MC-25)."* Participants identified that this could be especially helpful for new employees while they are on boarding: *"…where to eat, where to get staff, these kinds of basic things for example it can be connected with an information package and it should be available from day one of the mission (MC-20)."* Again, some participants had already experienced systems that provided support: *"in Darfur it was very easy because we used to have an app like that (MC-19)."* Participants from the military-side of the mission raised challenges in accessing common cultural knowledge held by the civilians: *"[civilian staff] have been here for three years in the mission. They will have a breadth of knowledge, but that breadth of knowledge is here (pointing to his head), it's not on a document, it's not shared (MU-15).*

4.3 Connectivity

A key issue for participants was connectivity. Whether it was for communicating with the rest of their force or with their families, having a stable way of connecting was seen as essential for their work performance, personal physical safety and wellbeing. Currently, network issues remain a big hurdle whether for work or in allowing members to communicate with their families.

4.3.1. Work Performance

Connectivity brought significant communication benefits in the work context, allowing staff: "to chat all over the world with other UN missions and I can also connect via internet to contact, to chat with my projects (MC-16)." The UN like other organisations is moving to the cloud: "The UN is becoming more accessible via the internet you have more and more apps that are in cloud that's the direction we are heading (MC-7)." Several civilian participants brought up the idea of a "remote office" in which they would be able to connect to the office and solve issues remotely. This in turn would allow them to handle work more efficiently while providing them with flexibility in terms of location. Proposals included: "manage systems remotely… you can monitor how the parameters, how the packets are going to and fro, instead of you rushing back to the office, just log in … and you can fix them remotely (MC-19)."

Connection issues were seen as limiting work performance: *"if you have network problems, it is not possible to share your documents (MC-36)"* and also were often a harbinger of significant issues: *"sometimes we have emergency situations and we don't know if the government want to kill the network (MC-36)."* Participants were positive about systems that could work offline as well as relying on connectivity such as the Field Suit Support system, *"which is really very user-friendly even without network it always works through most of the time (MC-34)."*

4.3.2. Personal Physical Safety

For some participants being *"able to communicate [emergencies] all the time" was important "for the sake of our security (MC-3)."* Participants identified that physical safety could be enhanced by greater integration of information when calling in physical alerts: *"the other information that they need from you, what is your location, who you are and all those things, the technology could help to give them immediately (MC-24)."* Military participants highlighted the need for instant communications with troops on the ground or security forces at the base: *"we have patrol but in addition to that, we could have CCTV, … moving of personnel from this point to that point could be monitored through the CCTV and if something, they will come to action rapidly (MC-9)."*

ICTs were seen as a way to provide greater support in action, reducing the dangers and uncertainty: *"there are situations you forget everything... [if] they are monitoring these things from the room and then*

they will come out for the action immediately ideally (MC-9)." Military participants felt that all the different troops in mission should be connected with a universal network, whether the person was on foot patrol or vehicle patrol: *"We are completely blind at certain moments, which is really uncomfortable position to be in. You have to make crucial decisions, you know, time critical on information that is not there. … we are talking about people's lives here (MU-13)."*

4.3.3. Wellbeing

When speaking about the need to stay connected, both civilian and military participants brought up video calling their families as a way of alleviating some of the strains of working in the mission. This need to ensure connection with families was highlighted by those who managed others, *"they need to make a real good effort in making sure that these people are able to reach out back to their families (MU-13)."* And *"the mission affects family men and if you are in a mission you should stay connected to your family. The mission should think or UN should think how to keep people connected to their families (MU-25)."* Connecting with families and friends was seen as key to boosting morale: *"because of this ICT we are able to constantly get in touch with our family that making life a little better than if we hadn't heard from them (MU-51)."* For many participants, connecting with their family members was listed as one of their most favourite uses for ICTs. And the most popular technologies were smartphones: *"I carry it all the time, so that's the most used item in ICT. I use it for emails, I use it to for the news, I communicate with family and friends and also for work related (MC-14)."*

4.4 Privacy

With regards to privacy, cyber security and secure data, connections were seen as lacking in current ICTs. Several participants were concerned about being hacked or having information intercepted. However, there were no actual experiences of such problems reported.

4.4.1. Work Performance

With information sharing, there were several issues with transparency and privacy settings. However, concerns lay not in the use of ICTS but rather on the protocols about future information use and sharing: *"Basically, I work on information. Some sections give information you don't want to share with other sections. It's quite sensitive because the information you give becomes national (MU-45)."*

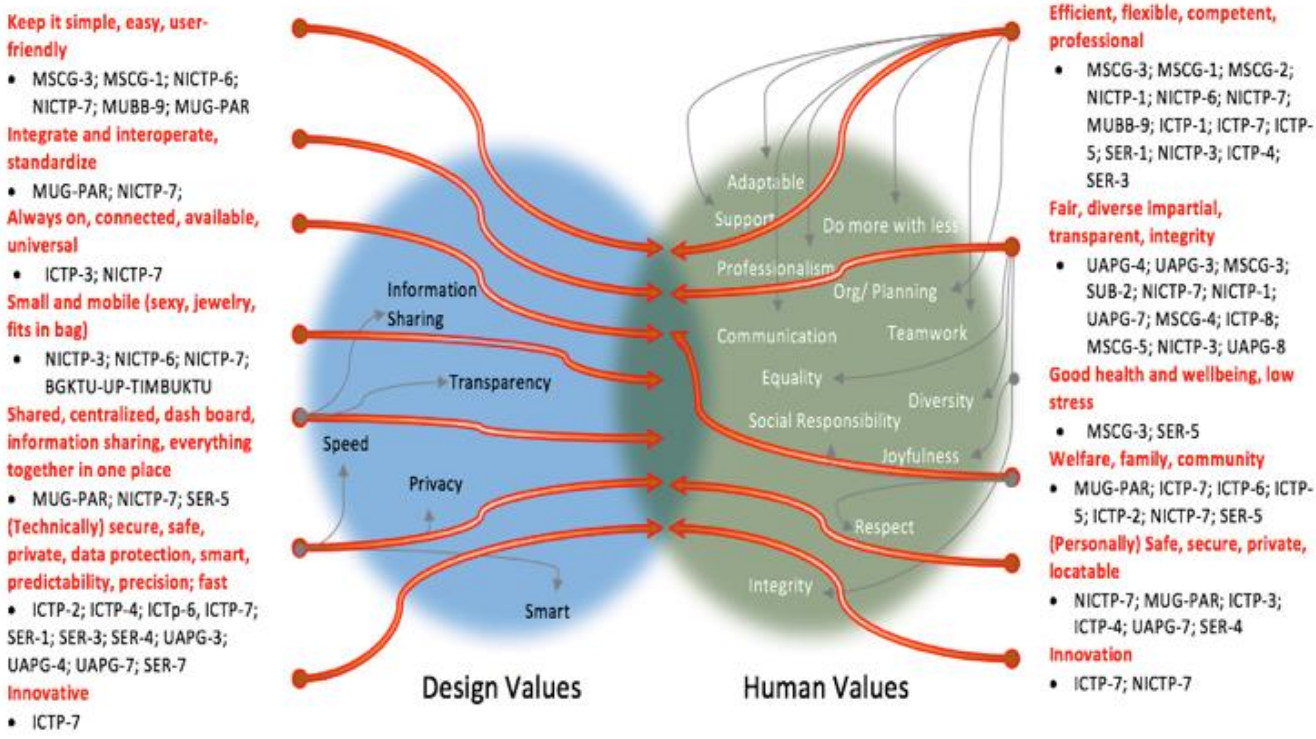


Figure 2 - Venn Diagram of intertwined values

Cybersecurity was raised in all groups: "*Without data protection there's no way to have fairness. Because, if I send information to someone that, surprise, is lost and someone is there to hack this information, it's not fairness in the ICT* (MU-46)." A lack of awareness was also flagged: *"[major issue] work wise, is ICT security. I have to teach my clients I have to tell them the basics of how to protect their data* (MC-4)."

4.4.2. Personal physical safety

Although both civilian and military participants could see the benefits of monitoring and surveillance, these benefits also raised issues: *"But then, tracking everyone, I think it is a bit of a privacy, a human privacy issue (MC-25)."* The transition to online and digital raised issues for some participants who were concerned about the security of their personal data: *"yeah, from a cyber protection or security point of view… a big gap that could be filled (MC-7)."*

4.4.3. Wellbeing

The most common concern from participants when discussing information and surveillance was the importance of confidentiality and making sure that information was not misused: *"But it is not only the confidentiality of the things I have seen in the past that there has been a misuse by managers (MC-24)."*

5. VALUES & RECOMMENDATIONS

The main values that emerged from the Focus Groups that should help to tailor ICT adoption and use strategies for missions such as MINUSMA are provided in figure 2 and summarised into four main themes as detailed in the following sections, ending with recommendations.

5.1 Being Safe

Safety was the most important and a core value for self, others and the mission. Multiple layers of safety enabling ICTs were identified, from the automatic panic button to the provision of up-to-date and timely geographical information in the field. Safety was a ubiquitous value, of importance at all times and in all places. A key facet was the constant sense of danger that emerged with UN missions based in unsafe environments, in contexts that evoke fear. This, in itself, is profoundly unhealthy for staff. However, it could be reduced through providing all staff with several means that made them safer, such as the sending out of all-person alerts on everyday devices. With safety the core value for the majority of staff, it is critical that existing and proven technologies, some already in use in other missions in the UN are deployed in all contexts thereby saving lives.

As well as physical safety, there was also the quasi-traditional values and unsubstantiated concerns about cybersecurity. As with physical safety it is the sense of imminent and potential danger that pervades the value. Informing staff in induction about excellent ICT security, etc. and the protocols for information sharing should aim to establish a valid sense of security. Notably, values related to ICT/data security were limited to the work context, with almost no concerns about personal data security. Even the focus on health records and appointments was fundamentally related to work, as being regularly medically assessed is part of the job.

5.2 Maintaining Relationships

The potential to maintain relationships was one of the main values for participants. ICTs were seen as already adding value to maintaining relationships, meeting the need to communicate with family, friends and colleagues. Beyond ensuring access to internet and devices (almost all staff have a smartphone) there is no need to develop technologies or applications to support this, as they are all mainstream, already available and being used.

5.3 Doing Work Well

Working well was intrinsically linked with safety and increasing automation that impacted on safety, such as health information provision across bases and all-staff alerts. Beyond safety, working well, was frequently frustrated by the lack of integration of information and systems. There was also a lack of training via online or digital approaches. Engaging with everyday apps and technologies highlights what should be possible and has raised participants expectations.

5.4 Being Cared for by UN

This sense of care could be manifested through targeted onboarding and in particular, information continuity, so that personal information, such as health records, followed staff across missions and bases. In additional organisational care requires the establishment of appropriate practices for information sharing, data access and training.

5.5 Recommendations

Based on this analysis, the following recommendations were made to inform future design decisions in ICT development and deployment at the MINUSMA base and in other similar contexts.

1. To adopt individual personal mobile devices as the prevalent platform on which to develop ICTs for staff. This should include the use of mobile apps for work purposes as

well as for information provision, health, onboarding, communication, and most importantly as a way to receive alerts.

2. To improve personal physical safety through increased situational awareness, achieved through increased integration of information and data sharing further supported through surveillance ICTs such as sensors and cameras. Providing centralized dashboards indicating safe/unsafe zones, allowing ways of reporting emergencies in a simple and efficient manner, and providing easier on-the-ground communication are seen as key. Wearables that provide information on where staff are, that can be 'pushed' to register concerns and that can sense if you need support and are unable to ask for help would seem an appropriate future direction.
3. To integrate and consolidate information and systems, increasing automation where possible, with automated approaches to health information provision across bases and all-staff alerts essential. In parallel, to establish appropriate practices for information sharing, data access and use with training via mobile phones.

Although ICTs are becoming visible in peace-building literature, yet there has not been any overarching account that hold out human values with ethical import as a central design criterion for peacekeeping. In this study, we have offered such an account, emphasizing VSD theory and method to enhance the digital peacekeeping in which values arise, encompassing not only the wellbeing of peacekeepers, but also their work-related productivity and security. As mentioned in Britt & Adler (2003): '*the proper study of peacekeeping is the peacekeeper…...the human dimension factors may either improve or slow down the wellbeing or performance of the peacekeepers.'* It implies that we should focus on peacekeepers' values vis-à-vis technical design which are determinants of improved performance. The goal of Digital Peacekeeping is to create an enabling environment for peacekeepers that maximally improves their well-being as well as their capacity to function efficiently and effectively in the conflict-zones. This can only be achieved if the human and value dimension of the frontline peacekeepers is given due consideration.

6. CONCLUSION

Understanding how technology can be used for sustainable peace and social change is essential (Firchow et al., 2017). With technology, an "amplifier of human intent" (Toyama 2011) expanding the base of knowledge increases the understanding of the circumstances under which technology amplifies peace supporting a holistic discussion of the ways that technology can impact contentious social and political processes. The study reported in this paper highlights that the core values for frontline peacekeepers are: safety, relationships and the reciprocity of doing work well for an organization that cares for them. Providing ICTs that meet all these values does not need to wait for technological advances, rather everything needed already exists. Appropriate, value-based ICTs will increase user ability to focus on their peacekeeping roles, feeling safe, loved and cared for, something clearly essential in a conflict zone.

10. ACKNOWLEDGEMENTS

We sincerely thank UNDFS-ICTD and the UNU-CS for assisting this field research at the UN MINUSMA in Mali. We are also immensely grateful to the frontline peacekeepers, both uniformed and civilian, deployed at Bamako and Gao bases for sharing their ideas with us to improve peacekeeping.

This work was partially supported by the Creative Fuse North East project exploring how technology can be used to support diverse sectors.

11. References

Bartone, P.T. et al. (1998). Dimensions in psychological stress in peacekeeping operations. *Military Medicine*, 163(9), 587- 593.

Borning, A., and Muller, M. (2012). Next steps for value sensitive design. In: Proc. CHI '12, ACM, 1125- 1134.

Brahimi, L., et al. (2000). *Report of the Panel on United Nations Peace Operations.* A/55/305-S/2000/809. New York: United Nations General Assembly.

Brey P. (2015) Design for the Value of Human Well-Being. In: van den Hoven J., Vermaas, P., van de Poel I. (eds) Handbook of Ethics, Values, and Technological Design. Springer, Dordrecht. https://doi.org/10.1007/978-94-007-6970-0_14

Britt, T.W., & Adler, A.B. (Eds.), (2003). The psychology of the peacekeeper: Lessons from the field. Praeger Publishers, Westport, Conn.

Brown, T., Wyatt, J. (2010). Design thinking for social innovation. *Stanford Social Innovation Review*, (Winter). Retrieved from https://ssir.org/articles/entry/design_thinking_for_social_innovation

Bushe, G.R. (2013). The appreciative inquiry model. In E.H. Kessler, (ed.) *Encyclopedia of Management Theory*, (Volume 1, pp. 41-44), Sage Publications.

Castro, C. A. (2003). Considerations When Conducting Psychological Research during Peacekeeping Missions: The Scientist and the Commander. In B. T. Litz, & A. B. Adler (Eds.), *The Psychology of the Peacekeeper*. Westport, CT: Praeger, Pp. 11-27.

Communication for Peacebuilding: Practices, Trends and Challenges. (2011). Search for Common Ground with support from USIP.

Curran, D. (2016). *More than Fighting for Peace? Conflict Resolution, UN Peacekeeping, and the Role of Training Military Personnel*. New York: Springer.

Dorn, A. W., & Dawson, P. F. (2020). Simulating Peace Operations: New Digital Possibilities for Training and Public Education. *Simulation & Gaming*, *52*(2), 226–242. https://doi.org/10.1177/1046878120968605

Dorn, A.W. (2021). *A Technology Innovation Model for the United Nations: The "TechNovation Cycle".* UN Unite Paper 2021(1). <https://walterdorn.net/pdf/Tech-Innovation-Model-for-UN_UnitePaper-2021-1_Dorn_2021-01-27.pdf>

Dorn, A. W. (2016). *Smart Peacekeeping: Towards Tech-Enabled Operations.* IPI, Providing for Peacekeeping No.13.

Fidler, D. (2015). Can UN Peacekeeping Enter the Digital Age? *Council on Foreign Relations*, 2015. Retrieved from http://blogs.cfr.org/cyber/2015/07/02/can-un-peacekeeping-enter-the-digital-age/

Firchow et al. (2017). PeaceTech: The Liminal Spaces of Digital Technology in Peacebuilding. *International Studies Perspectives18*(1): 4-42.

Friedman, B., Kahn, P. H., Jr., and Borning, A. (2006). Value Sensitive Design and information systems. In P. Zhang and D. Galletta (eds.), *Human-computer interaction in management information systems: Foundations*, 348-372. Armonk, New York; London, England: M.E. Sharpe.

Garber, K., Carrette, S. (2018). *Using technology in fragile, conflict, and violence situations: Five key questions to be answered*. Washington, DC: World Bank.

Gowan, R. & Andersen, L.R. (2020). Peacekeeping in the shadow of Covid-19 era: Short-term responses and long-term consequences. DIIS Policy Brief. https://www.diis.dk/en/research/peacekeeping-in-the-shadow-of-covid-19-era.

Harris, J.J and Segal, D.R. (1985). Observations from the Sinai. *Armed Forces & Society* 11:235-248.

Hourcade, Juan Pablo, Bullock-Rest, Natasha E. (2011). HCI for peace: a call for constructive action. In: Proceedings of ACM CHI 2011 Conference on Human Factors in Computing Systems, 2011,. pp. 443-452. https://dx.doi.org/10.1145/1978942.1979005

Jett, D. (2019). Why UN Peacekeeping Missions Fail. The Globe Post. https://theglobepost.com/2019/08/01/un-peacekeeping

King, N. & Brooks, J. (2017). Doing template analysis: a guide to the main components and procedures. In *Template analysis for business and management students* (pp. 25-46). SAGE Publications Ltd, https://www.doi.org/10.4135/9781473983304

Maslow, A.H. (1943). "A theory of human motivation". *Psychological Review.* 50 (4): 370–96.

Mahamuni, R., Kalyani, K., Yadav, P. (2015). A simplified approach for making human values central to interaction design. Procedia Manuf.3, 874–881.

MINUSMA Fact Sheet (2013). Supporting political process and helping stabilize Mali. *UN Peacekeeping.*

https://peacekeeping.un.org/en/mission/minusma

Mugabi, I. (2021). Why UN peacekeeping missions have failed to pacify Africa's hotspots. Mad for Minds. https://www.dw.com/en/why-un-peacekeeping-missions-have-failed-to-pacify-africas-hotspots/a-57767805

PIERSKALLA, JAN, AND FLORIAN M. HOLLENBACH. 2013. "Technology and Collective Action: The Effect of Cell Phone Coverage on Political Violence in Africa." American Political Science Review 107: 207–24.

Salaün, N. (2019). *The Challenges Faced by U.N. Peacekeeping Missions in Africa. The Strategy Bridge.* https://thestrategybridge.org/the-bridge/2019/10/14/the-challenges-faced-by-un-peacekeeping-missions-in-africa

Sigri, U and Basar, U. (2014). An Analysis of Assessment of Peacekeeping Operations. *The Korean Journal of Defense Analysis*. Vol. 26, No. 3, pp. 389-406.

Shaker, N. (2015). *UN peacekeeping needs a major technological update. QUARTZ*. <https://qz.com/509351/un-peacekeeping-needs-a-major-technological-update/>

Scaturro, G. (2016). Tech for peace: Facts and figures. SciDevNet. https://www.scidev.net/global/features/tech-for-peace-facts-and-figures.

Toyama, K. (2011). Technology as amplifier in international development. Paper presented at the i*Conference* 2011, February 8-11, 2011, Seattle, WA, US.

Tran Ngoc, C., Bigirimana, N., Muneene, D. et al. *Conclusions of the digital health hub of the Transform Africa Summit (2018): strong government leadership and public-private-partnerships are key prerequisites for sustainable scale up of digital health in Africa*. BMC Proc 12, 17 (2018). https://doi.org/10.1186/s12919-018-0156-3

UN Peacekeeping (2021). *What peacekeeping does?* <https://peacekeeping.un.org/en>

UNSC. (2013, April 25). *Resolution 2100* (2013). http://www.un.org/en/peacekeeping/missions/minusma/documents/mali%20_2100_E_.pdf

Van den Hoven, J. (2007). ICT and value sensitive design. *The Information Society: Innovation, Legitimacy, Ethics and Democracy in Honor of Professor Jacques Berleur* SJ, 67–72.

Wählisch, M. (2019). Big Data, New Technologies, and Sustainable Peace: Challenges and Opportunities for the UN. *Journal of Peacebuilding & Development*, *15*(1), 122–126. https://doi.org/10.1177/1542316619868984.

Weaver, W., and Shannon, C.E. (1963). The Mathematical Theory of Communication. Univ. of Illinois Press.

WEIDMANN, NILS B. 2015. "Communication Networks and the Transnational Spread of Ethnic Conflict." Journal of Peace Research 52: 285–96.