Evaluating Virtual Reality Experiences Through Participant Choices

María Murcia-López[*] Facebook
Tara Collingwoode-Williams[†] Goldsmiths University of London
William Steptoe[‡] Facebook
Raz Schwartz[§] Facebook
Timothy J. Loving[¶] Facebook
Mel Slater[‖] University of Barcelona

Abstract

When building virtual reality applications teams must choose between different configurations of the hardware and/or software aspects, and other factors, of the experience. In this paper we extend a framework for assessing how these factors contribute to quality of experience in an example evaluation. We consider how four factors related to avatar expressiveness affect quality of experience: Eye Gaze, Eye Blinking, Mouth Animation, and Microexpressions. 55 participants experienced an avatar delivering a presentation in virtual reality. At fixed times participants had the opportunity to spend a virtual budget to modify the factors to incrementally improve their quality of experience. They could stop making transitions when they felt further changes would make no further difference. From these transitions a Markov matrix was built, along with probabilities of a factor being present at a given level on participants' final configurations. Most participants did not spend the full budget, suggesting that there was a point of equilibrium which did not require maximizing all factor levels. We discuss that point of equilibrium and present this work as an extended contribution to the evaluation of people's responses to immersive virtual environments.

Index Terms: I.3.7—Computer Graphics—Three-Dimensional Graphics and Realism—Virtual Reality

1 Introduction

When constructing a virtual reality (VR) application teams typically have a choice between different configurations of the objective aspects of the design of the experience, or factors, relating to hardware and/or software. Moreover, there may be tradeoffs between these factors in terms of rendering performance, latency, cost of implementation and so on. For example, avatar facial expressions that are highly accurate may be computationally expensive, but in the end make no difference to the experience when compared to more simple facial expression representations.

Teams need a way to evaluate trade-offs between factors that differentially affect the experience. The challenge in doing so lies in that we currently lack a systematic method for assessing experience that does not rely on large multi-factor experiments that elicit participant preferences across all possible configurations of factors. The primary objective of this research is to extend a previously existing method for evaluating presence to inform hardware and/or software trade-off decisions. We illustrate the validity of the approach via a novel application and extension of the method in an avatar presentation context. We briefly review the concept of presence before describing in detail the methodology we adapt and extend.

1.1 Presence

Presence was originally defined as the sense of 'being there' in the place depicted in VR. The concept was derived from its original use in telepresence systems, where people operating in a remote environment through a robot typically had the feeling that they were located 'there' [14]. It was adapted to the similar feeling of 'being there' that people had in VR – e.g. [11, 19–21, 28].

Presence is not the only criteria against which to judge the quality of a VR experience. For example, a participant can have a strong sense of presence, but be quite uninterested in or unconvinced by events that are unfolding. Garau et al. [9] found that participants interacting with virtual human characters would experience those characters more like people when they exhibited some minimal level of response to participant actions compared to treating those characters as part of a computer interface without such responses (see also Steed et al. [29]).

Slater [24] deconstructed presence into two different components: Place Illusion (PI), as the illusion of 'being there', and Plausibility (Psi) as the illusion that events in the virtual environment were actually occurring, in spite of full knowledge that this was just a simulated environment. For a recent review of the field see [22].

1.2 Method of assessment of participant responses

The standard method for evaluating presence is to use questionnaires [13, 27, 30]. Although these provide valuable information, especially in conjunction with behavioral and physiological measures [25], they nevertheless are difficult to interpret - they provide no universal measure since one participant's score of '5' out of a maximum of '7' might mean something completely different than another's. Additionally, answering '7' or at the high end of a scale has no consequences for the respondent, yet decisions that may have costly consequences may be made on that basis. Physiological measures also do not provide a universal solution, given their complexity and utility in a limited number of scenarios. Even if questionnaires or physiological measures alone were suitable, there is still the problem of the explosion of conditions necessary to run a factorial experimental design to test (e.g.) presence across all possible conditions. For example, suppose there were k factors each with just two levels; the factorial design would require 2^k conditions. There are ways to reduce this, through hierarchical designs, but this becomes infeasible for larger k and more than binary levels.

A method that potentially overcomes these methodological challenges was introduced in Slater et al. [26]. This method was based on an analogy with colorimetry, where in order to measure the subjective response to illuminated surfaces participants are never asked to judge how (e.g.) 'red' a color is, but to match their perception of a color produced through manipulation (by adjusting red, green and blue projectors). Carried out over many participants and patches of light, experimenters can calculate on average how much 'red', 'green' and 'blue' went into the makeup of any particular patch.

Similarly, participants in the method of Slater et al. [26] were able to independently manipulate the extent of field-of-view, properties of a virtual body, perspective condition, and illumination quality, in order to match a level of PI or Psi previously experienced with all factors at their 'highest' level. This approach led to the derivation

[*] e-mail: mariamurcia@fb.com
[†] e-mail: tc.williams@gold.ac.uk
[‡] e-mail: willsteptoe@fb.com
[§] e-mail: raz@fb.com
[¶] e-mail: tjloving@fb.com
[‖] e-mail: melslater@ub.edu

of probabilities of how much each of these factors contributed to PI or Psi, without participants having to answer a questionnaire. As in colorimetry, they only had to judge whether a particular experience *matched* or not their experience of the system with all factors at their highest level. The 'matching' is an observable event (it is a fact that they matched) and not something on an ordinal scale, and obviates the problem of knowing the meaning of a score in a questionnaire.

We extend this approach and we use a much simpler criteria for the assessment of participant responses to a VR application: what makes the experience better? - whether they prefer the experience with a specific factor configuration compared to others, and to what extent they would be willing to 'pay' for this configuration.

To illustrate this alternative approach we describe an experiment that shows how a version of this method captures preferences in the context of a VR experience where participants were faced with a virtual human character giving a presentation about how to have a good conversation. Participants evaluated configurations of different levels of four factors relating to avatar facial behavioral realism: eye gaze, eye blinking, mouth animation, and facial microexpressions, and how they contribute to a better experience. Our goal is both to show how the method was improved in its application and analysis, and to present the findings with respect to the four factors above. The main contributions of this paper are:

- We extended a previously used method for evaluating VR experiences, modifying the goal that participants were given from matching previously experienced feelings of presence to "making the experience better". We included a virtual budget component that restricts the number of factors that participants can maximise. We discuss the benefits and limitations of the proposed approach, and suggest directions for future work.

- We present the experimental design, analysis and results of a study to illustrate the aforementioned extension of the method. The design of the study includes a placebo that serves to demonstrate the effectiveness of the approach. A knowledge transfer questionnaire aims to demonstrate that participants were capable of evaluating the factors whilst remaining engaged in the presentation.

In the next sections we give further background, followed by the experimental design, results, discussion and conclusions.

2 Background

2.1 Building a convincing virtual presenter

One of the most compelling experiences in VR is to have a face-to-face encounter with another avatar. This is different to seeing an image or video of a character on a 2D screen as participants share a virtual space. Such virtual interactions require social cues that are central to real face-to-face conversations. Significant research has been conducted to identify the factors that provide these social cues.

One notable factor highlighted by research is avatar behavioural realism. Behavioral realism refers to the extent to which the avatar behaves or moves like a human being [3]. It can be operationalized in the most simple terms by the absence or presence of non-verbal cues, which are a key component of face-to-face interactions. For example, Pan et al. [18] found that participants reported high levels of social presence when a virtual agent blushed after making a mistake during a presentation. Additionally, Bailenson et al. [4] found that virtual agents that mimicked the head movements of the participants were more persuasive and received higher positive trait ratings. Interestingly, the positive effect of behavioral realism is dependent on the understanding of the various factors and their implementation in different use cases. Bente et al. [5] found that having plausible gaze behaviour contributes to social presence but also found that when the duration of the eye contact was too long it led to negative responses from participants.

2.2 Eliciting participant preferences

The method introduced in Slater et al. [26] aims to find an optimal configuration amongst possible factors in a VR application. In the first use of the method, four factors were considered: illumination level (Gouraud shading, static global illumination, global illumination with real-time shadows), field-of-view (small, large), display type (simulated power wall, head-mounted display), and virtual body (none, static, real-time full motion-tracked body). Participants first experienced a scenario with all these factors at their maximum level, and were asked to concentrate on their sensation of either PI or Psi. Starting from a low level for each factor and under simple cost constraints, participants were able to increase one level at a time until they declared that their feeling of PI of Psi matched their original feeling. There were 36 possible configurations, and each change by the participant corresponded to a transition from one configuration to another. By counting the number of times that a change was made from configuration i to j, a 36×36 transition probability matrix (P) was constructed, where entry p_{ij} is the probability of transitioning to configuration j given that the participant was experiencing configuration i. From the transition matrix P, Markov Chain theory was used to compute the k-step transition probabilities (the probability of being in configuration j, k transitions after being in configuration i) [12]. The data also supported computation of the probabilities of choosing a 'match' (i.e. when the participant had stopped through matching their original feeling of PI or Psi) for each configuration. Hence, this method affords computation of interesting probabilities that represent how the 'average' participant behaves in terms of choosing a configuration that matches the level of PI or Psi.

Azevedo et al. [2] closely followed this method augmented with EEG measures of engagement and Azevedo [1] applied the method to auditory environments. Skarbez et al. [23] applied the method to Psi in the context of interaction with virtual characters. Bergstrom at al. [6] applied the method to unravelling how Psi may be influenced by different characteristics of sound rendering, and the responses of musicians to the participants, in the context of a virtual string quartet performance. Gao et al. [8] explored how different factors contributed to the believability of a virtual environment in the context of a rock climbing application. The study involved participants first experiencing a rock climbing environment at the highest levels of each factor: visual appearance of the rocks (3 levels), the appearance of the surrounding scene (from simple to complex, 3 levels), environment sound effects (from no sound to high level windy sound, 3 levels), and environment behavior (none to dynamic changes such as animated leaves, 2 levels). The windy sound and dynamic features were the most important contributors to believability in this setup, and the analysis of the transition matrix showed that to get to the windy sound, dynamics and rock appearance were the transitions that participants made.

Just as this method has been used for PI, Psi and believability, it can be used for any other type of response that is definable and identifiable by participants. The method does not assume an underlying quantitative scale, but only that participants are able to compare the effect of two different configurations and choose one over the other, or conclude that there is no difference in terms of their own experience between them.

In this paper we considered what is perhaps the most straightforward and understandable response by participants to changes in configuration. Given two configurations, we are only interested in the configuration that participants felt made the experience *better for them.* Like previous uses of this method we first let participants experience the 'best' possible configuration in a demo task but we did not then ask them to select changes to move towards that experience since we did not want to impose our notions of what constitutes a better experience. Rather, participants were free to move through the configuration space in any direction, their only criterion being whether they prefer the newly chosen configuration to the previous one. There are other differences with previous uses of this method detailed in the next section.

3 Experimental Design

Participants experienced a pre-recorded avatar presenter delivering a 14-minute presentation, divided into two equivalent trials of seven minutes each, on “How to Have a Good Conversation” in a 1:1 setup in VR. This approach was selected to ensure maximum stimuli uniformity across participants. At fixed points during each trial, the presenter stopped and participants had the opportunity, should they choose to do so, to modify characteristics relating to the presenter through a user interface. The characteristics they were able to modify were the factor levels (described below). All participants were given the same fixed budget, and each transition to higher levels that participants made had an associated cost. We encouraged participants to spend the minimum budget required to achieve what they regarded as the best form of presentation. Note that the budget was virtual and in no way affected participant compensation.

We evaluated four factors related to avatar expressiveness in a 1:1 presentation scenario: Eye Gaze (EG), Eye Blinking (EB), Mouth Animation (MA) and Microexpressions (ME). This is denoted in a property vector of the form $S = [EG, EB, MA, ME]$. Each instance of the property vector was considered a configuration. Altogether there were a total of 81 possible configurations, detailed below:

(EG) Eye Gaze

- (EG = 0) Static centered eyes
- (EG = 1) Dynamic random gaze targeting
- (EG = 2) Dynamic saliency-based gaze targeting

(EB) Eye Blinking

- (EB = 0) None
- (EB = 1) Normal-distribution around mean frequency of 6 seconds
- (EB = 2) Normal-distribution around mean frequency of 6 seconds (note that this level was added as a placebo effect to ensure that participants were only moving to higher levels if this made the experience better for them)

(MA) Mouth Animation

- (MA = 0) None
- (MA = 1) Oculus Lipsync [16]
- (MA = 2) Oculus Lipsync with Action Unit Easing

(ME) Microexpressions

- (ME = 0) None
- (ME = 1) Random triggering of microexpressions
- (ME = 2) Linked to events from Oculus Lipsync and Eye Gaze

All of these factors are variations on the facial animation system built for the Oculus Avatar SDK, and described in detail in our Oculus Connect 6 talk [15]. The highest level in each category is representative of the behavior exhibited in the public release of the Oculus Avatar SDK (with the exception of EB = 2 as noted, which was used as a placebo in this experiment, but is triggered by events in the gaze and speech models related to times of higher blink probability in the Oculus Avatar SDK). The eye gaze model in EG = 1 and EG = 2 conditions both use a physiologically-based kinematic model to generate realistic human saccades, micro-saccades, and smooth pursuits. The difference between EG = 1 and EG = 2 is that the latter uses a saliency model to distribute gaze as opposed to distributing gaze randomly. The saliency model uses a number of factors to estimate the highest probability of where the user is looking, which includes head motion, the array of objects in the current field-of-view and their type, movement, and size, how long an object has been fixated on and ignored, and the normalized distribution of gaze eccentricity from the center. The factors of mouth animation are based on Oculus Lipsync in MA = 1, and our extensions to the animation model that feature in the Oculus Avatar SDK in MA = 2. Natively, Oculus Lipsync generates a probability over 15 visemes (including laughter), and the avatar rig can be set accordingly per-frame. In our animation extension in MA = 2, we further correspond these visemes to their component parts, based on FACS action units [7]. Each action unit has a custom onset and falloff curve, which results in significantly smoother and more natural appearance of mouth movements pertaining to speech. Finally, the microexpression factors operate as a secondary model ME = 2 and in the Oculus Avatar SDK; linked to characteristic events in the gaze and Lipsync models. For instance, an upward gaze may trigger a slight raising of the eyebrows, an end of speech may trigger a subtle smile, and head movement may trigger slight perturbations of the facial state. These microexpression models are designed to be extremely subtle; adding texture and nuance rather than semantic or emotional undertones to the avatar’s performance. In this experiment, ME = 1 is a state in which these microexpressions are triggered randomly at a regular cadence, with no relation to the rest of the facial state.

The following restrictions were built into the system:

1. Participants were given a total budget of 7 in order to encourage them to think carefully about their transitions and avoid the possibility of choosing a maximal configuration [2,2,2,2] (which would not give us any meaningful information). The budget restriction reduced the total number of configurations being evaluated from 81 to 80.

2. The cost of moving to each subsequent higher factor level was equal to 1 budget unit.

3. Factors could only be increased by one level during each transition opportunity. For example, participants could not move from level 0 to level 2 on any of the factors without first making a transition to level 1. This ensured that participants had a chance to experience and assess the factors at all levels. It also reduced the amount of data that had to be collected to populate the Markov transition matrix as some of the transitions became impossible (transitions from level 0 to 2).

4. Participants were able to remove budget units spent (and recover total budget) from any factor each turn by reducing the level, but could not reallocate the budget recovered in the same turn. For example, if a factor was on level 2, participants were able to recover budget and bring the factor back to level 1 or level 0 in one turn. However, they could not reallocate that budget to another factor and increase from level 0 to level 2 on that same turn, respecting rule 3 described above.

5. At the end of each trial, participants were asked to confirm or modify their final configuration choice. This final confirmation turn had none of the previous restrictions in place to allow them to jump to their preferred final configuration.

Each of the trials randomly started in one of four low base configurations, in which three factors were at level 0 and one factor at level 1 ([0,0,0,1], [0,0,1,0], [0,1,0,0], [1,0,0,0]). Participants therefore began with the remainder six budget units to spend.

4 Method

4.1 Participants

A total of 55 participants (31 female, 24 male; average age 35.5 years, SD = 11.3) were recruited from the Oculus user base. All participants signed a consent form and the study was approved through Facebook Research Review. Two participants had no previous VR experience. Twenty participants were broadly classified as gamers (categorized as spending more than one hour gaming a week). Participants were paid £75.

4.2 Materials

The user study was conducted in a lab at Facebook London. An Oculus Rift Consumer Version 1, two Oculus Touch controllers and three Oculus sensors were used. The virtual environment was rendered at scale 1:1 in Unity 2018.2.18f1 at 90FPS in each eye on an Intel Core i7-7700 CPU @ 3.60GHz, with 16GB RAM and Nvidia GeForce GTX 1080 GPU running Windows 10.

The virtual environment consisted of an empty custom built room. A modified version of the Oculus Avatar SDK 1.35 [17] was used to render the presenter and generate the different factor levels. The participants' virtual hands were rendered in a non-human colour (blue) to remove any effect of skin-tone on performance in the task. The Oculus Touch trigger buttons were used to interact. A ray casting method was used in order to point at the user interface in the experience, with a blue reticle appearing upon collision with it. In order to account for handedness, participants were able to switch the interaction from the left or right controller using the X (left hand) or A (right hand) buttons.

The concept of the user interface for participants to make transitions was very simple and straightforward. Four rounds of usability testing with five participants each were completed to iterate on its design prior to the study. The resulting version consisted of a floating panel with a slider showing coloured discrete marks for each factor as well as for the budget at the top. A "plus" and "minus" button were displayed on either side for each factor slider, allowing participants to increase or decrease levels by selecting them. Each factor level increase or decrease would automatically update the budget bar to reflect the units taken or recovered, with a delayed animation to make this obvious for participants. The vertical order in which factors were presented in the user interface was counterbalanced across participants to avoid order effects (with each participant having the same counterbalanced order for the full duration of their session).

4.3 Metrics

There were two sets of dependent variables. The first was the final configuration that the participants chose. The second consisted of the transition data, which depicted the chronological changes made by a participant from configuration i to another configuration j across the two trials. We also included a post-trial questionnaire. This included a 16-question knowledge transfer questionnaire about the content that the pre-recorded presenter delivered. The facts required to correctly answer questions 1-7 were delivered during the presentation in trial 1, and questions 8-16 in trial 2. We also included a question asking participants to rank the factors in order of importance. The questionnaire was delivered outside of VR. Participants also completed a semi-structured interview.

4.4 User interface demo task

To familiarise participants with interaction in the virtual environment, we created a task where participants had to fill in a bar to continue onto a simplified version of the user interface by pointing at and selecting the "plus" button. If participants were in doubt as to how to interact, the experimenter would assist.

Participants were then shown the full user interface containing all factors and levels as part of a full user interface demo task. Here participants were able to experience what was possible within the system by manipulating the factor levels of the presenter. The goal of this task was to allow participants to familiarise themselves with the system and to feel comfortable interacting with the user interface. The budget was set such that the participant could test the configuration where all factors are set to their maximum level. The presenter was not the avatar from the main presentation task but rather a different avatar to encourage the practice of changing the factor levels. These would be the same factor levels that they would be able to modify on the presenter during the main task. The presenter spoke a short looped phrase to allow participants to see the effect of the changes they made. Participants were only able to advance to the next stage if they displayed understanding of how to interact with the user interface, understood the effect of the changes to the confederate avatar, and experienced the system at the highest configuration [2,2,2,2]. They were encouraged to think aloud to help the experimenter assess if they understood how the user interface operated. The confederate avatar for the demo task differed in both appearance and voice to the confederate in the main task. This was to prevent familiarity affecting the choices participants made.

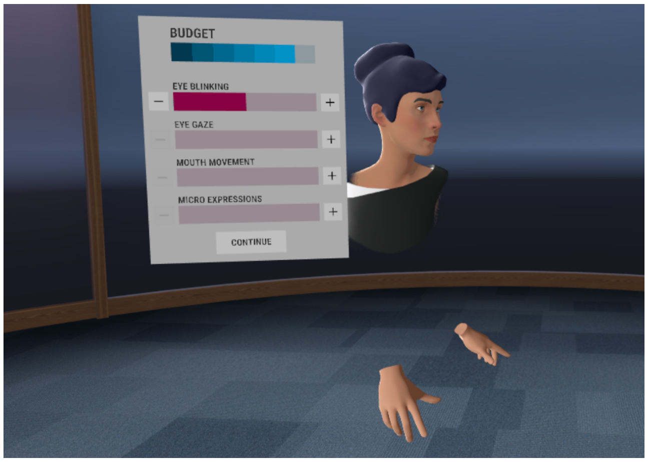

Figure 1: Main task scene with the presenter and user interface.

4.5 Main task

In the main task, participants were faced with a virtual presenter who delivered a presentation to them about how to have a good conversation [10]. This presentation was adapted from an TEDx Creative Coast talk and was selected out of a series of talks in a pilot study because it elicited the highest engagement levels as evaluated via questionnaires.

Participants were reminded of the instructions for the main task and advised that from that moment on there may be options for the experience that they cannot always afford, meaning that going forward the budget restrictions described in Sect. 3 were applied. The main task was split into two 7-minute trials, with each trial corresponding to the first and second half of the presentation. At seven fixed, equally spaced times during the presentation, a dialog box prompt would appear giving the participant the opportunity to remain in the same configuration or make a transition to another configuration. If participants decided to make a change the full user interface would appear, as shown in Fig. 1. An extra dialog box would appear at the very end of each trial to allow participants to confirm or change their final configuration. Participants were not encouraged to think aloud during the main task to avoid any distraction from the presentation and the evaluation of factors. This task was designed to be completed with the participant standing.

4.6 Procedure

Participants were welcomed to the session and escorted to the lab. The experimenter introduced the hardware and the task. The experimenter helped the participant don the headset. After recentering to ensure that participants were facing the correct direction and that the virtual floor was at the correct height, they experienced the user interface demo task. After making sure that the participants understood how to interact with the system, they were reminded of the instructions and completed the main task. Upon completion of the first trial, the experimenter helped the participants remove the headset. The participant was then given a few minutes (no more than

5) to sit and rest, as well as to drink some water. The experimenter then helped participants don the headset. They were reminded of the task before starting the second trial. Participants were asked to stand throughout the main task of each trial. However, in the cases were participants expressed a need to sit down, a chair was provided. The chair was positioned in the same position the participants were asked to stand in and the application was then recentered to account for the height change, to ensure that the presenter's height would always match the participant's. One participant chose the sitting option. After both trials were completed, the experimenter helped the participants remove the headset and they were handed an iPad to answer the questionnaire described in Sect. 4.3. They were also offered water. The questionnaire was completed sitting down. They then discussed their experience with the facilitator.

5 Results

5.1 Method of analysis

Participants completed two trials with each trial starting at different configurations. The results were analysed for each trial independently as well as combined, with all showing similar results. In addition to looking at the transitions from configuration to configuration (Transition Analysis), we also analyze the final configurations that participants chose after they had reached the configuration through following the transitions (Final Configuration Analysis). For Transition Analysis, we denote the set of 80 configurations that a participant could experience by C. Note that the budget restrictions made configuration [2,2,2,2] impossible to reach. The set of all possible transitions from configuration to configuration is therefore a subset of C. Each transition is of the form:

$$[EG_t, EB_t, MA_t, ME_t] \to [EG_{t+1}, EB_{t+1}, MA_{t+1}, ME_{t+1}]$$

denoting the transition from the configuration that a participant was in at time t, to the configuration at time $t+1$.

From the set of all such transitions we can build the probabilities π_{ij} that a participant in configuration $i \in C$ would next choose configuration $j \in C$. This gives us the $m \times m$ Markov transition matrix P, where $m = 80$ is the number of configurations. Fig. 2 shows, for example, the numbers of transitions for each factor separately. The full transition matrix is similar, but includes each of the 80 configurations, and thus is too complex to display.

P^k is the k-step transition matrix, with elements that give the probability that a participant in configuration i would be in configuration j, k steps later. Let u be a 1×80 vector where u_j are the initial probabilities of being in configuration $j \in C$ (i.e. the probability of being in a particular configuration). Then uP^k are the probabilities of being in the configurations after k transitions. All of the above follows from Markov Chain theory [12]. P is constructed from the 770 observed transitions (55 participants $\times$ 2 trials $\times$ 7 transitions).

Markov Chain theory requires that the probability of making a transition to any valid configuration is only dependent on the current configuration and not previous history. We follow this abstraction for the purpose of model building. Using the results of all transitions made by the participants we can estimate the transition matrix: the probability of a transition to a configuration given the current configuration. From the resulting transition matrix we can calculate the probabilities of being in the various configurations after the successive transitions within the set system and budget restrictions.

Suppose that the number of transitions from i to j is n_{ij}. Then the frequency estimate of the probability π_{ij} is $p_{ij} = n_{ij}/N_i$, where N_i is the total number in row i.

From the set of all transitions various other probabilities can be estimated, including the probabilities of each factor level being part of the the final configuration arrived at by participants. We can also compute the marginal probabilities that any particular factor at any level is included in any configuration.

After completing their final transition at the end of each trial participants could choose to make one more change. This was to act as a confirmation or not that they had ended in their desired configuration. Since this final confirmation choice was not regulated by the budget restriction (to allow them to jump onto their preferred configuration regardless or where they were), it is not included in the transition analysis.

Table 1: The four highest probability configurations (C) after each transition (k) with [0,0,0,0] as the starting configuration, and assuming that participants chose the transitions randomly. C is the property vector of the form $S = [EG, EB, MA, ME]$.

Transition	Configuration	Probability	
		Frequency	Random
1	0000	0.333	0.063
	0010	0.333	0.063
	0100	0.333	0.063
	0001	0.000	0.063
2	0010	0.271	0.030
	0100	0.214	0.030
	0000	0.120	0.030
	0110	0.114	0.030
3	0010	0.171	0.026
	0110	0.154	0.026
	0111	0.115	0.026
	0100	0.107	0.026
4	0110	0.146	0.026
	1111	0.128	0.026
	0111	0.126	0.026
	0010	0.101	0.026
5	1111	0.132	0.026
	0110	0.122	0.026
	0111	0.120	0.026
	1110	0.073	0.026
6	1111	0.126	0.026
	0111	0.106	0.026
	0110	0.096	0.026
	2111	0.068	0.013
7	1111	0.114	0.026
	0111	0.091	0.026
	2111	0.087	0.013
	0110	0.074	0.026

5.2 Transition analysis

We added one last transition at the end of each trial that would act as a "confirmation box". This was to allow participants to end each trial in their preferred configuration. Since this last transition was no longer regulated by the budget restriction (to allow them to jump onto their preferred configuration regardless or where they were), we must exclude this last transition from the transition analysis. We therefore calculate the transition probability matrix without the last transition from each trial (110 transitions).

The 55 participants completed a total of 7 transitions in each of the two trials leading to a total of 770 transitions. From this the count matrix N is computed, which represents the number of transitions from configuration to configuration. Note that, due to the nature of the transition restriction, N is a sparse matrix, with 226 non-zero cells (the 80 $\times$ 80 matrix has 6400 cells - the budget restriction reduces the number of valid cells to 4000).

Diving deeper into the question on sparsity, 184 out of the 226 non-zero matrix cells representing transitions were visited four times or less. Results also indicate that 366/770 transitions were from and to the same configuration where $[EG_t, EB_t, MA_t, ME_t] = [EG_{t+1}, EB_{t+1}, MA_{t+1}, ME_{t+1}]$.

The configuration [0,0,0,0] has each of the four factors at their 'minimal' levels. For the purposes of analysis we ordered the con-

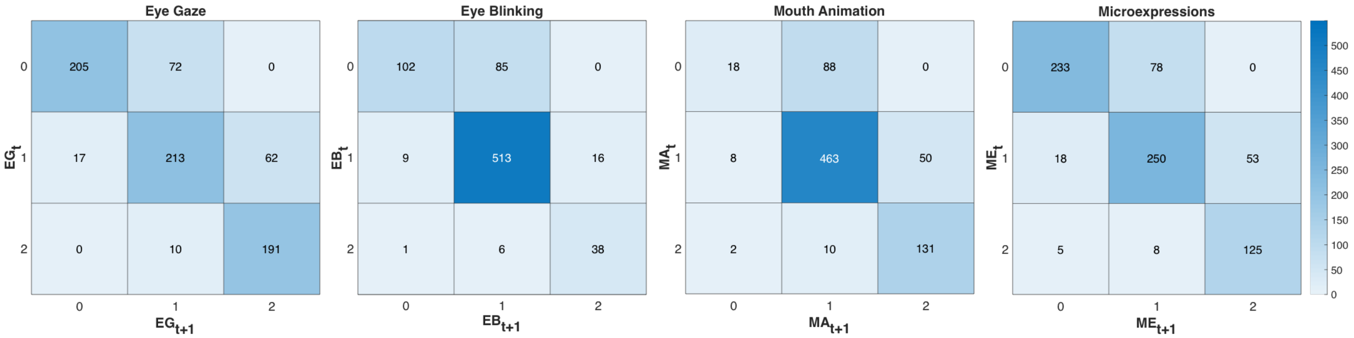

Figure 2: Markov transition matrices showing the number of times participants moved between factor levels.

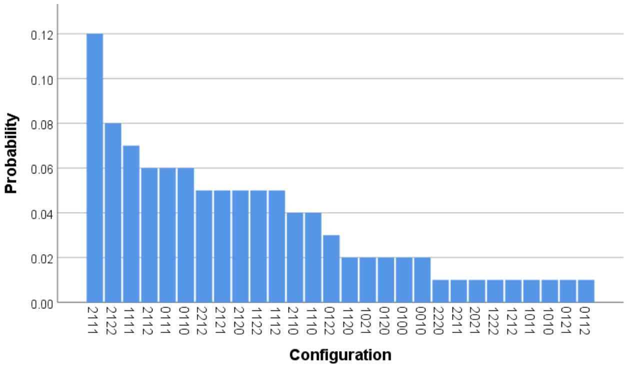

Figure 3: Distribution of final configurations across both trials.

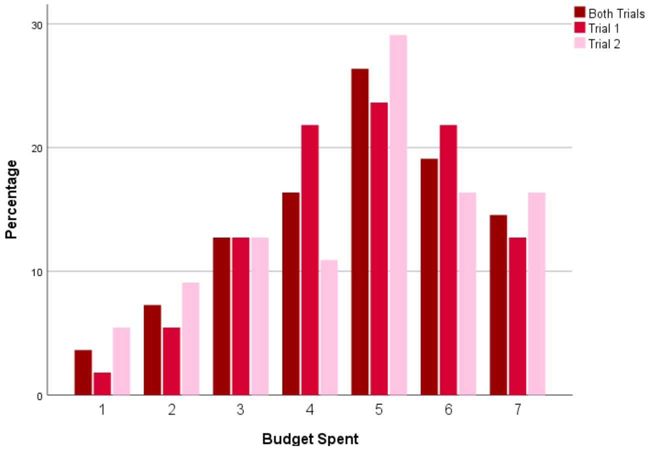

Figure 4: Distribution of budget spent for Trial 1, Trial 2 and both trials.

figurations so that [0,0,0,0] occupies the first place, and therefore the probability vector u = [1,0,0,..,0] (79 zeros) represents this as the starting configuration for a hypothetical participant. Now using $uP^k, k = 1,...,7$ we can find the probabilities for the four highest probability configuration transitions that were more likely to be chosen from this starting configuration. This is shown in Table 1. After the first transition the most likely configurations were the original starting one, or with change in Eye Gaze or Mouth Animation. By transition 5 the most likely configuration had each of the factors at level 1. Note that the probabilities seem to be low, but should be compared with the probabilities assuming individuals selected transitions randomly. Table 1 also shows the highest probability configurations in this case, taking into account that some transitions were impossible. Starting from other randomly chosen low base configurations results in similar transitions as for [0,0,0,0], but more transitions are needed to reach the same configuration.

5.3 Final configuration analysis

For any particular configuration we can estimate the probability of ending in that configuration $P(C|final)$. This is the number of times that participants ended in that configuration over the total number of final configurations, which is 55 participants × 2 trials = 110. The probability distribution is shown in Fig. 3. Overall, the group's most likely final configuration was [2,1,1,1]. Note that 81% of trials had an exact equivalence between the last transition and the confirmation transition (i.e. chosen through the confirmation UI after all the transitions had been completed) and 95% were one level away from their confirmed final configuration. A Wilcoxon signed-rank test showed no significant difference in the distribution of final configurations between the two trials ($Z = -0.115, p = 0.908$). Participants chose their responses non-randomly. If they had, then Fig. 3 should illustrate a fairly uniform distribution among the final configurations. If we carry out a Chi-squared test comparing the resulting distribution with the theoretical uniform distribution, then random choice is an inconceivable hypothesis ($\chi^2(27) = 68.182, p < 0.001$).

5.3.1 Marginal factor probabilities

Starting from [0,0,0,0] we can compute after k transitions the marginal probabilities of a factor being present at a given level (shown in Table 2). For example, after 4 transitions we can find the probability that (e.g.) Eye Gaze would be present at level 1. We consider this after 4 transitions and after 7 transitions. After 4 transitions Eye Blinking and Mouth Animation have the highest level 1 probabilities and these two have the greatest overall probability of having made at least one change. This is also true after 7 transitions. The strong result is that in the case of Mouth Animation there is a very high probability of there being at least one change, though the greater bulk of the probability is at level 1. This is followed by Eye Blinking where again, the greater probability is at a level 1 change. After 7 transitions there is not a lot of difference between Eye Gaze and Microexpressions. It is important to note that Eye Blinking at level 2 was the same as level 1, designed as a 'placebo' to understand if participants were following instructions as they were designed. This is reflected in Table 2, where the probabilities of EB being at exactly level 2 are always small in comparison to others.

5.4 Budget analysis

Participants were able to spend a maximum of seven budget units in each trial. Results show that the mean budget spent by participants across both trials was 4.7 with an $S.D.$ of 1.6. Fig. 4 shows the

Table 2: Probability that the configurations after 4 and 7 transitions would contain the factor at the given level with the probability estimates.

		After 4 transitions				**After 7 transitions**			
		Level 0	**Level 1**	**Level 2**	**At least Level 1**	**Level 0**	**Level 1**	**Level 2**	**At least Level 1**
Frequency estimate	Eye Gaze	0.510	0.382	0.109	0.490	0.289	0.367	0.345	0.712
	Eye Blinking	0.272	0.685	0.043	0.728	0.104	0.819	0.077	0.896
	Mouth Animation	0.086	0.790	0.124	0.914	0.027	0.678	0.295	0.973
	Microexpressions	0.506	0.419	0.075	0.494	0.305	0.446	0.249	0.695

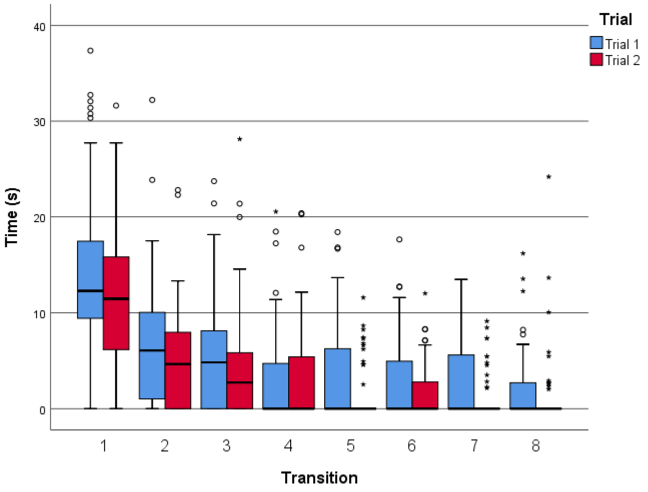

Figure 5: Time taken for participants to complete each of the eight transitions for each trial. Selecting “no change” was recorded as zero. Boxes represent the interquartile ranges (IQR). Whiskers represent either the extreme data points or extend to $1.5 \times IQR$. Outliers are shown by circles. Extremes are shown as asterisks.

distribution of budget spent. A Wilcoxon signed-rank test showed no statistically significant differences on budget spent between Trial 1 and Trial 2 ($Z = -0.801, p = 0.423$). A Mann-Whitney U test showed no statistically significant differences on budget spent based on gender ($U = 1278, p = 0.351$). A Mann-Whitney U test showed no statistically significant differences on budget spent based on gaming experience ($U = 277, p = 0.158$).

5.5 Transition times

The time taken by participants to complete each transition decreased over time and is shown in Fig. 5. A Wilcoxon signed-rank test with a Bonferroni correction applied showed that there were no significant differences between the two trials.

5.6 Questionnaire analysis

5.6.1 Factor ranking

Participants rated the factors from most to least important. Overall, Mouth Animation was ranked as most important, followed by Eye Blinking, then Eye Gaze and then Microexpressions. There was a significant difference in the distributions of importance rankings for each of the factors ($\chi^2(2) = 46.29, p < 0.001, df = 3$). Post hoc analysis with Wilcoxon signed-rank tests were conducted with a Bonferroni correction applied, resulting in a significance level set at $p < 0.008$. Median (IQR) perceived importance for Eye Gaze, Eye Blinking, Mouth Animation and Microexpressions were 3 (2 to 4), 2 (2 to 3), 1 (1 to 2) and 4 (3 to 4), respectively.

5.6.2 Knowledge transfer

A Kruskal-Wallis H test showed that there was no statistically significant difference in knowledge transfer score (total number of correct responses) based on final configuration ($\chi^2(2) = 21.664, p = 0.301, df = 19$). At a first glance at Fig. 6, we can see that the lowest correct responses came from Question 1 and Question 8. Interestingly, the content asked in these questions was located at the start of each trial. This could relate to participants settling into the experience and paying less attention to the presentation at the beginning of each trial.

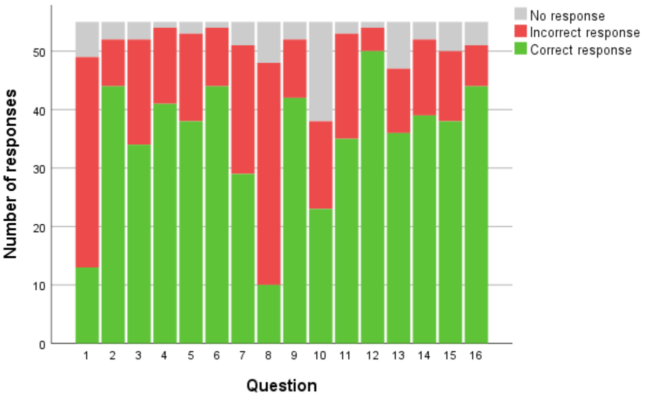

Figure 6: Stacked bar graph depicting participants’ correct answers (green), incorrect answers (red) and blank answers (grey) for each of the knowledge transfer questionnaire questions.

5.7 Interview analysis

Interview responses were coded into the following themes:

Participants felt they had sufficient opportunities to assess the factors and believed that the configuration that they ended in was the best given the options available. This is important for us in validating that the number of opportunities offered to make transitions in our experimental design were sufficient.

Participants tended to mimic real-world behaviour, even though most reported that they knew they were facing a pre-recorded avatar. Some participants laughed, and others stepped back when they felt the presenter was too close.

Participants could articulate what their goals were in making the experience better. They reported that their goal was to make the presenter more ‘real’, ‘natural’, ‘not mechanic’, ‘human’, ‘less distracting’, ‘comfortable’, and ‘life-like’.

The budget was sufficient to find a ‘match’. By this we can conclude that the budget was sufficient.

Some participants had difficulty in interpreting ME levels. Unlike the other factors, ME changes were harder to pinpoint. This led to participants having varied notions on what the effect of this factor was and could explain the results.

6 Discussion

This work extends the methodology introduced in previous work [6, 23, 26] by:

- including a rigorous application of the budget concept,
- using a placebo effect to validate the design,

- proposing a new goal for participants to *make the experience better* where they no longer have to match the initially experienced sensation with all factors at their maximum level,
- allowing participants to move between configurations in any direction before confirming the final configuration, and
- removing the assumption that a given level of a factor is better than another.

For decades VR practitioners have been imposing measurements (e.g. presence) to evaluate quality of experience, but maybe participants have different criteria. Therefore, this approach introduces the idea that results are purely based on participant preferences rather than what teams or experimenters decide in advance is important.

However, the method still presents a number of limitations. On one hand, we do not know the extent to which the results from a study using the method generalise to other scenarios. We also do not yet understand what might be the maximum number of factors and levels for an evaluation using this method. The higher the number of factors and levels evaluated, the larger the Markov transition matrix that has to be populated with data, and the higher the cognitive workload for participants. Another limitation is that it relies on one particular analytical technique (Markov Chain theory - replicating the previous uses of the method). Given that participants were freely navigating through the different configurations during both trials, it is not possible to correlate a single configuration (e.g. the final configuration) for each participant to their knowledge transfer questionnaire score. However, there may be additional ways to approach the analysis that could yield novel insights.

In the study, we evaluated four factors relating to avatar facial behavioral realism: EG, EB, MA and ME. Participants were able to iteratively assess three levels for each of these factors in a 1:1 presentation delivered by a pre-recorded avatar. This allowed for the evaluation of 80 versions or configurations of the system (a virtual budget limited participants from reaching a configuration where all factor levels were maximised).

Overall, the group's most likely final configuration was [2,1,1,1] (Full Model Targeting Eye Gaze, Linear Eye Blinking, Oculus Lipsync Mouth Animation and Random Microexpressions). This is the configuration that we would recommend teams to implement in 1:1 presentation experiences that the scenario we evaluated is representative of and for the studied set of factors and levels. This is not to say that this combination is the 'best' in all circumstances, but is relative to this particular system and presentation type.

Most participants did not spend the full budget, suggesting that there were personal optimal configurations which did not require maximising all factor levels. This is consistent with theory denoted in previous works that have implemented this method [6, 23, 26]. We also found no significant differences in budget spent based on gender and gaming experience. We suggest gathering more granular background data around gaming experience to continue monitoring this result in future uses of this method.

The placebo effect included in factor EB worked well to help us verify that participants were completing the task as it had been designed (to only spend budget if it made the experience better for them) as the probability that participants would end in EB = 2 was low. The expectation was that participants would not end in level 2 for EB as there was no actual increased quality or value, but there was an increase in cost. The placebo effect actually allowed for participants to maximise all factor levels with the available budget. However, this effect was not observed given that most participants did not consume the full budget.

Results indicate that knowledge transfer was generally high but lower at the start of each trial. This is a good indication that participants were involved in the presentation and paying attention to the information that was delivered, beyond evaluating the different factors. Participants may have concentrated on settling into the task and evaluating options towards the beginning of each trial.

Overall, the time taken by participants to make transitions between levels decreased over time. This could indicate that, towards the end, participants had generally found their optimal configuration and decided not to make further changes, whereas towards the beginning there was more exploration and evaluation. This could also be attributed to fatigue; participants may have felt tired and therefore less engaged in the task and more focused on finishing quickly. Another possibility is that participants may have overcome the learning curve, and felt more confident in using the system to achieve the result they wanted. However, for this last potential reason we would have expected to see a significant difference between trials, which we did not observe.

Future work should further explore how different factors may contribute to quality of experience in other applications, extending the range of use cases evaluated. This information will be important to help teams define the best possible configurations for different VR applications, including future hardware that can support those configurations (e.g. face tracking technologies). Even though our goal with the proposed extended method was to model the average user based on the actions that participants took, future work could focus on studying individual differences. The community should equally continue to evaluate other factors and levels in the context of immersive social interactions, and in multi-user scenarios where avatars are driven in real-time.

Extensions to the method should explore other budget restrictions that will force participants into tighter evaluations and, conversely, scenarios in which the budget does allow for maximisation of all factors to understand whether a point of equilibrium can still be reached when there is no tension. Moreover, the budget could reflect real costs, for example, of implementation or production. Other suggestions include different configuration starting points for trials (i.e. completely random or configurations with high levels) to explore whether consistent points of equilibrium are reached. For larger data collection, the research method could also be run as an 'in the wild' study by embedding the experience in public applications and optionally allowing headset owners to voluntarily take part in them. This would allow for more sophisticated machine learning approaches to the data analysis.

7 Conclusions

This paper is based on the framework described in Slater et al. [26] that proposed a method for exploring the contributions of different factors to the illusion of Psi and PI in a VR application. Here we have shown how this work can be extended to account for other objective features of a VR experience relating to avatar-mediated non-verbal communication. Importantly, this method avoids the need for self-report. The only information it is based on is observable - participants chose to make transitions (or not).

We tackle this problem with a novel approach; to explore what participants choose to be acceptable rather than risk imposing preconceived notions of what makes for a better VR experience. In the study we looked at four factors. The results have shown that most participants did not spend the full budget, implying that there was an optimum point reached without having to maximize the factors.

It is important to note that these results should not be taken as an evaluation of the factors themselves but as an exploration of their implementation and influence on participants' preferences on obtaining a better VR experience strictly applied in the context explored. Above we mentioned that MA followed by EB were accepted overall at a minimum of level 1, but there was less agreement in what was the optimal level for EG and ME. This is not to say that EG and ME are not important: in this setup, this is the preference established by participants. This framework hopes to provide teams that are looking to build VR applications with a consistent tool to evaluate the impact of different factors on experience, as well as a way to understand the point of equilibrium across a range of use cases.

Acknowledgments

The authors wish to thank Thomas Figg, Natasha Elliott-Deflo, Simonas Skrebe, Khush Singh, Victoria Kyu, Max Archer and Jessica Cornick.

References

[1] A. S. Azevedo. *3D Sound Enhanced Presence in Virtual Environments*. PhD thesis, Master's thesis. University of Lisbon, 2013.

[2] A. S. Azevedo, J. Jorge, and P. Campos. Combining eeg data with place and plausibility responses as an approach to measuring presence in outdoor virtual environments. *Presence: Teleoperators and Virtual Environments*, 23(4):354–368, 2014.

[3] J. N. Bailenson, K. Swinth, C. Hoyt, S. Persky, A. Dimov, and J. Blascovich. The independent and interactive effects of embodied-agent appearance and behavior on self-report, cognitive, and behavioral markers of copresence in immersive virtual environments. *Presence: Teleoperators & Virtual Environments*, 14(4):379–393, 2005.

[4] J. N. Bailenson and N. Yee. Digital chameleons: Automatic assimilation of nonverbal gestures in immersive virtual environments. *Psychological science*, 16(10):814–819, 2005.

[5] G. Bente, S. Rüggenberg, N. C. Krämer, and F. Eschenburg. Avatar-mediated networking: Increasing social presence and interpersonal trust in net-based collaborations. *Human communication research*, 34(2):287–318, 2008.

[6] I. Bergström, S. Azevedo, P. Papiotis, N. Saldanha, and M. Slater. The plausibility of a string quartet performance in virtual reality. *IEEE transactions on visualization and computer graphics*, 23(4):1352–1359, 2017.

[7] R. Ekman. *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press, USA, 1997.

[8] B. Gao, J.-I. Kim, and H. Kim. Sensory and perceptual consistency for believable response in action feedback loop. In *Proceedings of Computer Graphics International 2018*, pp. 201–210. ACM, 2018.

[9] M. Garau, M. Slater, D.-P. Pertaub, and S. Razzaque. The responses of people to virtual humans in an immersive virtual environment. *Presence: Teleoperators & Virtual Environments*, 14(1):104–116, 2005.

[10] C. Headlee. How to have a good conversation.

[11] R. M. Held and N. I. Durlach. Telepresence. *Presence: Teleoperators & Virtual Environments*, 1(1):109–112, 1992.

[12] S. Karlin. *A first course in stochastic processes*. Academic press, 2014.

[13] J. Lessiter, J. Freeman, E. Keogh, and J. Davidoff. A cross-media presence questionnaire: The itc-sense of presence inventory. *Presence: Teleoperators & Virtual Environments*, 10(3):282–297, 2001.

[14] M. Minsky. Telepresence. 1980.

[15] Oculus. Oculus Avatars: Maximizing Social Presence, Part II. Oculus Connect 5, 2018.

[16] Oculus. Oculus lipsync guide. `https://developer.oculus.com/documentation/audiosdk/latest/concepts/book-audio-ovrlipsync/`, 2018.

[17] Oculus. Oculus avatar sdk. `https://developer.oculus.com/downloads/package/oculus-avatar-sdk/`, 2019.

[18] X. Pan, M. Gillies, and M. Slater. The impact of avatar blushing on the duration of interaction between a real and virtual person. In *Presence 2008: The 11th Annual International Workshop on Presence*, pp. 100–106. Citeseer, 2008.

[19] M. V. Sanchez-Vives and M. Slater. From presence to consciousness through virtual reality. *Nature Reviews Neuroscience*, 6(4):332, 2005.

[20] T. B. Sheridan. Musings on telepresence and virtual presence. *Presence: Teleoperators & Virtual Environments*, 1(1):120–126, 1992.

[21] T. B. Sheridan. Further musings on the psychophysics of presence. *Presence: Teleoperators & Virtual Environments*, 5(2):241–246, 1996.

[22] R. Skarbez, F. P. Brooks Jr, and M. C. Whitton. A survey of presence and related concepts. *ACM Computing Surveys (CSUR)*, 50(6):96, 2018.

[23] R. Skarbez, S. Neyret, F. P. Brooks, M. Slater, and M. C. Whitton. A psychophysical experiment regarding components of the plausibility illusion. *IEEE Transactions on Visualization and Computer Graphics*, 23(4):1369–1378, 2017.

[24] M. Slater. Place illusion and plausibility can lead to realistic behaviour in immersive virtual environments. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 364(1535):3549–3557, 2009.

[25] M. Slater, C. Guger, G. Edlinger, R. Leeb, G. Pfurtscheller, A. Antley, M. Garau, A. Brogni, and D. Friedman. Analysis of physiological responses to a social situation in an immersive virtual environment. *Presence: Teleoperators and Virtual Environments*, 15(5):553–569, 2006.

[26] M. Slater, B. Spanlang, and D. Corominas. Simulating virtual environments within virtual environments as the basis for a psychophysics of presence. In *ACM Transactions on Graphics (TOG)*, vol. 29, p. 92. ACM, 2010.

[27] M. Slater, M. Usoh, and A. Steed. Depth of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 3(2):130–144, 1994.

[28] M. Slater and S. Wilbur. A framework for immersive virtual environments (five): Speculations on the role of presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 6(6):603–616, 1997.

[29] A. Steed, Y. Pan, Z. Watson, and M. Slater. 'we wait'-the impact of character responsiveness and self embodiment on presence and interest in an immersive news experience. *Frontiers in Robotics and AI*, 5:112, 2018.

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