UNIVERSITY OF COPENHAGEN

Master of Computer Science

The Social Effect of Virtual Body Representations on the Observer

Empirical Observations from a VR Tug-of-War Game

Andreea A. Muresan
zph748@alumni.ku.dk

Supervisors

Henning Pohl
henning@di.ku.dk

Kasper Hornbæk
kash@diku.dk

September 2019

Contents

1 Introduction

VR has been shown to alter people's perceptions and bring about realistic physical and behavioural reactions to situations that are virtual and fabricated. For example, virtual environments (VEs) can decrease the perception of pain [30], aid phobia treatment [36] and activate stereotypical social responses ([14]). From arachnophobia [18] to fear of heights [29], virtual contexts have induced realistic fear-related physiological changes. In these cases, people reacted implicitly to VEs in congruence with their real-world experience. What is more, some researchers have noted that adding a physical objects greatly increased presence and realism in a virtual pit situation [29]. Apart from phobias, attitudes and behaviour are the most common perceptual changes that have been studied in virtual environments. Physical performance and related perceptual incongruencies have scarcely been investigated in such contexts. A notable exception here is the work of Peña and colleagues, which explores performance changes in a Wii tennis game, varying opponents' and players' appearances [38].

Virtual humans are a powerful driver for people's behaviour changes in VR. Apart from reactions to fear-inducing stimuli, most research exploring the transformative powers of VR has focused on virtual bodies. In this field of study, we identify two main themes: agency and viewpoint. Agency relates to the nature of the operator behind the virtual representation — an actual human or an algorithm. Viewpoint refers to the point of view assumed over the virtual representations. From a first-person perspective, theories of body ownership rooted in the *rubber hand illusion* [10] have shown that varying one's avatar determines behavioural and perceptual changes. Many experiments done by Slater [48, 45], Yee and Bailenson [58] shed light on the *transcending* power of virtual reality. From the *Proteus Effect* to theories of embodiment, numerous frameworks have been put forward to explain behavioural changes that occur when adopting virtual personas. However, it often happens that we are observers of these virtual bodies as much as we embody them. Researchers have often overlooked the third-person perspective when exploring perceptual changes in observers. An example of this is Pan and Slater's work on interactions between males and virtual females [35].

Considering the focus on embodiment, we aim to investigate perceptual changes occurring in users of virtual reality from a third-person perspective. To this end, we leverage the social power of virtual bodies and design a competitive VR rope-pulling game. We look at whether different levels of perceived strength would lead users to change their performance. To inform the design of the avatars, we first ran a survey to review perceived strength and intimidation in virtual humans. After establishing a baseline for strong and weak looking avatars, we ran an empirical study in which users played tug-of-war with five avatars representing various levels of perceived strength. We used a real physical rope which maintained

the same resistance across all trials. Our main hypothesis was that participants would perceive the rope as having various degrees of resistance. Furthermore, we expected them to use more force for stronger-looking avatars. To validate these assumptions, we measured the challenge and perceived force of each rope-pull instance. This set up allowed us to measure users' physical and perceived performance while seamlessly varying avatar appearances.
Our results show that users did indeed perceive variations in the force acting on the rope. However, we observed notable individual variations in users' ability to succumb to this illusion and sustain it throughout the game. Furthermore, the manner in which they pulled the rope was not consistent with our assumptions. Many participants reported an expectation to use more force with stronger opponents. Despite this, our quantitative results are inconclusive and suggest users were interacting with the system in unforeseen and unrealistic ways. While most VR experiences elicit realistic responses, this has not always been the case in our study. In the following, we present various incongruencies between quantitative data and qualitative feedback and give an overview of our findings. Our results inform the design of realism in virtual reality encounters and shed light on challenges designers might face in order to produce and sustain presence and plausibility illusions.

2 Related Work

In the following, we give an overview of the technical background that informed the design of our study.

2.1 The Transformative Power of Avatars

The threshold model of social influence put forward by Blascovich [8] implies that digital actors, who are perceived to be human have more influence than their computer counterparts. Both perceived agency and agency have been shown to play a role in eliciting social influence, especially for tasks such as persuasion [22, 15]. We use the term *avatar* to refer to a user's digital and embodied representation in virtual reality and the term *agent* for computer-driven digital entities. Additionally, people have been shown to respond socially to computers even when they are not necessarily embodied in avatars, known as the Computers As Social Actors paradigm [31]. In a virtual reprise of a well known social-psychological study - Milgram's obedience experiment - Slater et al. observed that people had realistic responses when shocking a virtual learner [45]. This occurred despite participants knowing that their activities had no effect in reality.
Digital self-representations have far-reaching implications from altering people's

behaviour in the virtual world, to shaping perceptions in reality. *The Proteus Effect*, postulated by Yee and Bailenson, refers to how users change their behaviour, conforming to the perceived behaviour of their digital representation [56]. They observed that users in more attractive avatars got closer to confederates and showed more self-disclosure than participants with less attractive avatars. Similarly, in a negotiation task about monetary splits, people with taller avatars were more confident and did more splits in their favour, while owners of shorter avatars were more likely to accept unfair deals.

The Proteus Effect is framed around self-perception theory [7] in a context in which users are deindividuated [60]. In such cases, users rely more on identity cues and behave in a way which conforms with the stereotypes of their virtually displayed body. Peña and colleagues provide an alternate explanation for this phenomenon. They suggest priming as an underlying cause of the Proteus Effect and raise concerns about users role-playing in their new identities [37]. They measure participants' awareness about the true scope of the study and use clothing as a means for priming. Avatars were shown from a third person perspective in a 3D distributed desktop virtual environment. In their first experiments, participants were given a negotiation task to solve in groups of 3 either, having all white or black robes. Results showed that participants in black presented more aggressive intent and lower group cohesion than those in white. In their second experiment, participants had either a Ku Klux Klan (KKK), doctor or transparent avatar and were tasked with writing 2 stories. Those with a KKK type avatar wrote more aggressive stories compared with the other groups.

To decouple the effects of priming and embodiment, Yee and Bailenson ran a study in which users were given an attractive or unattractive avatar in an immersive virtual environment (IVE) [57]. Users were looking at themselves in a mirror inside the IVe, or, alternatively, they saw a playback of someone from before. Participants interacted with a confederate of the opposite gender who was blind to the condition, and then completed a dating website task. The authors found that virtual embodiment resulted in more behaviour change with users. In the attractive condition, people chose more attractive dates and got closer to the confederates. In the opposite condition, users were more likely to increase their height in the dating profile.

Extending their work on the Proteus effect, Yee, Bailenson and colleagues found that in online communities, an avatar's appearance was a predictor of performance. Furthermore, changes in behaviour have been observed to last outside of IVEs. For example, participants with tall avatars had more aggressive negotiations with confederates [58]. The authors further investigated this effect in a conversation between 2 opposite-gendered participants in a distributed medium [52]. In this case, the findings, were inconsistent with the Proteus effect. Females who had an unattractive avatar were reported to behave more friendly, affectionate

and intimate. The authors explain these findings through the behavioural compensation effect [9].
The Proteus effect does not always benefit the user. In similar experiments, women having highly sexualized avatars objectified themselves more [17], embodiment in black avatars increased implicit racial bias [21]. Users were also more prone to persuasion by avatars that mimicked them [3] or consistently gazed towards them [4]. More recently, virtual social exclusion had similar negative effects with real-world exclusion, leading to less prosocial behaviour outside of VR [26].
In a series of experiments, Slater et. al put forward the idea of body semantics to explain how body ownership illusions can generate behaviour and attitudes in users [48]. Some examples are reducing implicit racial bias for owners of black avatars or affecting the perception of object sizes in the case of child avatars. Furthermore, the effects of reduced racial bias were observed for at least a week outside VR [6]. This paradigm differs from the Proteus Effect by attributing this change of behaviour to the generation of the body ownership illusion. While people have mostly experienced positive outcomes due to the empathy-driven by embodiment, highly stereotypical contexts may have opposite outcomes.
Virtual reality has shown its potential as an efficient tool for framing social and psychological studies. However, most research has focused on behaviour and attitude changes stemming from one's avatar. Few studies examine changes induced by the appearance of another's avatar or look at performance. In a notable example, Peña, Khan and Alexopoulos vary avatar and opponent body size in a Nintendo Wii tennis exergame and explain their findings through social comparison theory and priming [38]. They found that participants with obese avatars had less physical activities than those in normal avatars and showed that this effect was mediated by the appearance of their opponents. When the opponent had a more obese avatar, participants performed less. In the same experiment with women, they found that participants made the most effort when both avatars were normal and, furthermore, the appearance of obesity decreased performance [39]. In a series of experiments for health and behaviour change in VEs, Fox and Bailenson observed that participants made more exercise when their avatar lost or gained weight according to their movements [16]. From a first-person perspective, Christou and Michael explore the effects of avatars on performance. They created a game in which users deflected incoming objects while embodying a human or, stronger, alien avatar. Users had fewer misses and males used more force in the alien condition [13]. We contribute to research on performance changes in IVEs, by looking at whether participants use more force when faced with a stronger opponent than with a weaker opponent. We implement a rope-pulling game in virtual reality and allow participants to see their hands on the rope. However, we do not vary their body representation.

2.2 Embodiment

The effects of appearance have also been studied with respect to ownership illusions. Lin and Jorg investigate the influence of six different hand models on ownership [28]. In two experiments, participants played a game in which they had to block white spheres and experienced a threat scenario where a knife slashed their hand in the VE. They note that all these hands generated ownership for at least some participants, with various levels of strength. However, the effect was strongest for the realistic model and weakest for a *non-anthropomorphic* bloc. Despite this, the most realistic hands do not always receive the highest ratings. In their study Argelaguet et. al [1] evaluate ownership and agency for hands in 3 different realism condition that offer different degrees of freedom. In their study, the most realistic hand with the most accurate tracking was rated highest in ownership. Conversely, agency was stronger for less realistic hands. They conclude that the incongruence in ratings occurred because the highly realistic hand was often mismatched with the actual hand. This was not the case for the less realistic hands providing fewer degrees of freedom.
The rubber hand illusion[10] (RHI) is an experiment in psychology to demonstrate the formation of body ownership through congruent visual and tactile stimulation. Usually, a rubber hand is placed where a participant's hand should be, and both hands are stroked or tapped in a synchronized manner. The illusion created by the synchronized tactile and visual feedback generates a feeling of ownership of the fake arm for the participant. This illusion has been successfully replicated in mixed [23] and virtual reality [46], where researchers have also used avatars to generate feelings of ownership over virtual bodies. Slater and colleagues explore in an extensive body of work the parameters and effects of this illusion [47]. They introduce the term, *sense of embodiment* to capture the broader experience of embodiment in VR, which has three distinguishable levels: body-ownership, agency and self-location [25]. To evaluate the occurrence of RHI, researchers usually measure the perceived location of participants' hands and use questionnaires for subjective data. When movements are synchronized and the illusion occurs, usually there is a displacement of the perceived location closer to the rubber hand called *proprioceptive drift*. [41].
In an experiment combining embodied cognition and body ownership, Bailey, Bailenson and Casasanto [5] explore *space-valence associations* as a result of mirroring participants' hand movements. They conclude that multiple senses have to be engaged in order to obtain an effect.

2.3 VR Illusions

The illusions that enable the feeling of presence have been extensively studied in literature. *Presence* has many dimensions ranging from location, simply *being there*, to various aspects of computer-mediated communication such as social presence or co-presence — *being there together* [59]. For IVEs, Slater and colleagues propose a separation of presence in *Place Illusion* for location-related presence, and *Plausibility Illusion* to denote the perceptual override that occurs when users perceive what they know cannot occur. He defines *sensorimotor contingencies (SCs)* as actions users perform in order to achieve perceptual clarity. Some examples are as bending to see below, or touching a virtual object and receiving haptic feedback. He contends these VR illusions occur *as a function* of possible SC. The author defines plausibility illusion as"*the illusion that what is apparently happening is really happening (even though you know for sure that it is not)*" [44]. He emphasizes being the target of events is an important element in producing this illusion. Moreover, motor and visual synchronicity are essential to produce place illusion. Additionally, Slater posits that a correlation between a user's sensations and events they have not caused is important to bring about plausibility illusions (Psi). In his experiments, the author notes that Psi also occurs in low physical realism conditions, such as in the reprise of the obedience experiment [45]. To illustrate this illusion, most of the examples refer to interactions between virtual humans, such as reacting to the gaze of avatars. However, he also mentions people's realistic responses to a virtual pit, despite their knowledge there are no such objects in real life [49]. Commercial VR applications are able to reproduce this illusion with relative easy [1]. When haptic feedback is added to, the response is even more realistic, with significantly increased heart rate [29].
We identify three key elements in Slater's framework: synchronicity, correlation and realism. The body is, of course, the main focal point of these illusions. The author emphasises that these illusions take place despite people having knowledge of the actual environment. However, their reactions to the environment seem to be automatic. As such, some level of perceptual override takes place in order to generate these illusions. We posit, however, that these illusions can occur even in ambiguous settings, where participants do not have knowledge of their actual environment. It seems that eliciting realistic reactions in VEs is highly dependent on sustaining body ownership and meeting users' expectations. In our study, for the tug-of-war game participants were represented by an avatar showing both of their arms. Participants used VR gloves and held a real, physical rope that corresponded to the virtual rope. The opponents did not respond to participants' pull, and there was no force activating on the rope. However, in order to meet users' expectations

[1] https://store.steampowered.com/app/517160/Richies_Plank_Experience/

and to maintain realism, we used a spring and an elastic band to give some resistance when participants pulled. Furthermore, we used animations, game physics and sound to give participants continuous feedback about the state of the game. Overall, we paid particular attention to timing and synchronization in the design of the game. It has been observed that congruence between visuomotor actions gives rise to agency [25] and ownership is determined by synchronized haptic and/or visual feedback. We combine and leverage these synchronicities to match the expected outcome of peoples actions. Such *sorimotor contingencies* [44] allow us to create a context in which VR perceptual illusions could occur. We use spring resistance to give participants some motor sensations as the rope moving towards them is something they would expect to happen in a rope-pulling game. Realistic interactions that allow visual and motor synchronization seem to contribute to the formation of these illusions [44]. The ambiguous source of the motor feedback and its magnitude is what participants will have to discern. While we do not vary their appearance, we acknowledge that studies exploring a one-sided view of this interaction may fall short. However, avatar appearance is outside the scope of the present work. Despite this, we hope our findings can inform future research that looks at the relationship between user experience, ownership and realism in sustaining realistic behaviour and giving rise to more complex VR illusions.

3 Methodology

To investigate whether an agents' appearance can determine people to pull stronger, we ran a study where participants played tug-of-war with an opponent in VR.
The experiment has two parts: a survey and a user study. We ran a survey in order to determine perceptions of strength and intimidation in avatar design. We chose the avatars for the user study based on the results from the survey, where we measured perceived strength and intimidation. Please see section 4.1 for more details about the survey and choice of avatars. The user study represents the actual empirical study of our experiments. We presented this research as a VR gaming study. This setup allowed us to vary the opponents' appearances in a natural way. Realism is important to sustain the illusion that some force could be activating on the rope. Furthermore, it allows us to take objective measures of performance without interfering with the users' experience.

3.1 Research Aims

We aimed to give participants the illusion that the rope is being pulled back harder by stronger opponents. Equivalently, we expected people to feel the rope being pulled less by weaker opponents. If participants assumed these expectations, we

hypothesized they would also perceive pulling harder for strong opponents, and less for weaker ones. Additionally, we measured the actual force of the pull to verify performance differences.
Our independent variable is the appearance of the opponent. Each opponent was randomly assigned a condition of this variable, from weak-looking to strong-looking. We investigated if participants perceived any change in rope-pulling from one trial to another. We state our first hypothesis:

> **H1**: *Participants will perceive changes in rope-pulling force between opponents.*

To measure perceived changes in rope pull, we introduce a dependent variable, **challenge**. We asked participants to rate how challenging each rope-pull was after the respective trial. Additionally, we left it to the participant to interpret what *challenging* means.
We introduce **perceived pull** as a subjective measure of how much participants thought they pulled. We measured both variables on a 5-point Likeart scale. For each condition, we also took objective measurements of force. We used a force meter to detect the maximum force each participant pulled per trial. This constitutes the the third dependent variable, **force**. We expect users to react to the appearance of their opponent and pull harder for stronger-looking opponents. We state the hypotheses of these variables:

H2: *Participants will use more force for stronger-looking opponents.*

H3: *Participants will report pulling harder for stronger-looking opponents.*

H4: *Participants will find stronger-looking opponents more challenging.*

For H2, we use the maximum pull per trial as a measure of *more force*.

3.2 Experiment Design

We had a gender-matched, within-group experimental design for the user study. We independently varied the agent's appearance on a 5-point scale from weak-looking to strong-looking. Participants played five trials of rope-pulling with the five avatars chosen from the survey. The avatars were meant to display various degrees of strength and intimidation, compounded in a final weighted score. We chose the weakest (condition 1) and strongest (condition 5) male and female avatars. For the remaining three, we chose avatars in a low-average, average and high-average strength condition. Further details about the avatars chosen for the

experiment are presented in section 4.1.4. For each trial, we measure the maximum pull force in kilograms for each participant, which constitutes our dependent numeric variable. We further measure perceived pull and perceived challenge on a 5-point Likert scale, with constitute our two ordinal dependent variables.
Participants completed a post-experimental survey, gave feedback for each rope-pulling trial in-game and had a short chat with the experimenter at the end.
Between each rope-pull, participants rated four statements about the previous round on a 5-point Likert scale. They were presented with a panel of these questions in VR, and were asked to read the questions out-loud and give their answer to the experimenter. These questions are the following:

- **Q1**: I felt the virtual rope was realistic. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: It looked and felt like I was the one holding the rope. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q3**: How much did you pull the rope? Rating: 1 (*Not at all*), to 5 (*Very much*);
- **Q3**: How challenging was this round? Rating: 1 (*Not at all*), to 5 (*Very challenging*).

Q1 refers to rope realism, Q2 captures rope agency, Q3 refers to perceived pull and Q4 refers to challenge. We will use these terms to refer to the categories of these questions in the results section. Q3 and Q4 are two dependent variables. We present results of Q1 and Q2, however their main purpose was to give participants the impression we are evaluating rope performance.

For the post-experimental survey, we measured participants' subjective experience on three levels: body ownership, presence and co-presence measures. To measure presence, we retained 5 items from the igroup presence questionnaire (IPQ): [2]

- **Q1**: How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)? Rating: 1 (*Not aware at all*), to 5 (*Extremely aware*);
- **Q2**: How real did the virtual world seem to you? Rating: 1 (*Not real at all*) to 5 (*Completely real*);
- **Q3**: How much did your experience in the virtual environment seem consistent with your real-world experience ? Rating: 1 (*Not consistent*) to 5 (*Very consistent*);

[2] http://www.igroup.org/pq/ipq/index.php

- **Q4**: I felt present in the virtual space. Rating: 1 (*Fully disagree*) to 5 (*Fully agree*);
- **Q5**: In the computer generated world I had a sense of "being there". Rating: 1 (*Not at all*) to 5 (*Very much*).

The IPQ questionnaire measures presence on three levels: spacial presence, involvement and experienced realism. From our retained questions, Q1 refers to involvement, Q2 and Q3 capture realism, Q4 spacial presence and Q5 refers to general presence.
The survey items for body ownership were taken from [1]. For our study, we replaced the term *hand* with *arm*, and changed the 7-point rating scale to a 5-point one. The following questions were retained:

- **Q1**: I felt as if the virtual arm moved just like I wanted it to, as if it was obeying my will. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: I expected the virtual arm to react in the same way as my own arm. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q3**: I felt that the interaction with the environment was realistic. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q4**: I felt like I controlled the virtual arm. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q5**: I felt as if the virtual arm was part of my body. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q6**: I felt as if the virtual arm was someone else's. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*).

With respect to co-presence/social presence, we retained three questions from [34], changed their scale to a 5-point one and replaced the term *interaction partner* with *opponents*. They questions are:

- **Q1**: My opponents were intensely involved in our interaction. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: To what extent did you feel able to assess your opponents' reactions?. Rating: 1 (*I was unable*), to 5 (*Their reactions were clear*);
- **Q3**: To what extent was this like you were in the same room with your opponents? Rating: 1 (*Did not feel in the same room*), to 5 (*Felt completely in the same room*).

Additionally, the participants had to rate the appearance of the avatars on the same scale as in the initial appearance survey. We did this in order to verify our assumptions with respect to the hypothesis, that the opponents users faced were indeed

perceived as strong/intimidating. These questions were in the final part of the post-experimental survey:

1. This avatar looks attractive.
2. This avatar looks strong.
3. This avatar looks intelligent.
4. This avatar looks intimidating.

The avatars were randomized for each participant at the start of the experiment. This study was ran alongside another study by researchers within the department. We made a common call for participants to take part in a *VR Games Study*. Their task was to assess two virtual reality games and give researchers feedback about their experience. The two games were Tug-of-war VR and Whack-a-mole VR. At the end of the experiment, participants had a short recorded chat with the experimenter. Participants were asked to give feedback for the games and, in addition, we checked their awareness about the true purpose of the experiment.
We frame our research in the context of a competitive strength task between a user and a perceived agent in virtual reality. A competitive set up allows participants to face their opponents directly in a physical task. Furthermore, it constitutes a realistic interaction for a game setup. Players were told they were testing a rope pulling VR game and they would be asked for feedback and suggestions to improve the game. Their reported task was to face several different *opponents* in VR, compete by pulling a rope and win. The player was able to see their hands in the virtual environment holding the rope. They were instructed to keep their hands on the rope at all times. Their fingers were not animated and the grip was fixed on the rope, as such letting go of the rope would result in breaks of presence and ownership. We do not make the aim of our research transparent to avoid any possible biases, such as the observer expectancy effect, or induce powerful demand characteristics.
Participants see a countdown accompanied by sounds for each visual element being displayed. They are told to start pulling when they see *Start*, and stop pulling when they see *Stop*. When participants start pulling, they also see their opponent pulling and feel the resistance increase on the rope as they go on. Through this flow of events we synchronize, visual, audio and haptic feedback in order to give users the impression of agency and realism.

4 User Study

4.1 Survey

Through the survey, we aim to make an informed choice when selecting the final agents for the experiment. The main goal is to provide a range of designs that elicit realistic perceptions of intimidation and strength.

4.1.1 Survey Questions

Participants were told we were interested to see how people perceive traits of avatars based on their design and looks. The goal of evaluating strength and intimidation was not made completely transparent to avoid any possible bias. We gathered demographic data like gender and age. For each avatar, participants had to answer the following questions on a 5-point Likert scale, from 1 — *Strongly disagree* to 5 — *Strongly agree*:

1. This avatar looks attractive.
2. This avatar looks strong.
3. This avatar looks intelligent.
4. This avatar looks intimidating.

We are not interested in measuring intelligence, however it serves the purpose of distracting participants from the specific goal of the survey. Our main interests are strength and intimidation, as these variables may have the effect of changing user performance. Attractiveness is tangentially related because of the Halo Effect [32], whereby participants could be determined to assign more positive traits to the avatars. A further consideration with respect to the survey responses is agency. Since social responses can vary with agency [15], when answering the scale on intimidation, thumbnails of virtual humans may elicit lower responses that actual people.

4.1.2 Agent Design Coding

The agents were chosen through stratified sampling to display various levels of strength on a scale from *weak* to *very strong*. In total 36 agents were designed, 18 female and 18 male. The design procedure and implementation for these agents is described in chapter 4.2.1. Section 9.9 contains an enumeration of the images that users had to rate in this survey and represents the designed look of the agents' upper body. In section 9.4 we present tables containing a mapping of a unique ID for female (10) and male avatars (11) to an agent design and its respective condition. We use this ID to reference the agent ratings in the following chapters.

4.1.3 Participants and Design

To measure perceived strength in our agent designs, we ran a survey with 31 participants (15 female), aged 21-60 (mean 25.5). Participants were recruited from the university and through snowball sampling. The survey was gender matched and each participant rated 18 avatar thumbnails on a 5-point Likert scale measuring perceived strength, attractiveness, intelligence and intimidation. The order of the agent thumbnails was the same for all participants, but they were randomly selected from an initial ordered set. The survey took 10 to 15 minutes to complete.

4.1.4 Results and Discussion

Please see section 9.9 for a complete overview of all ratings given to all agents and their respective thumbnail. Tables 12, 13 from section 9.5 show the mean ratings given female and males. The tables are ordered ascending according to the **Weighted** column.

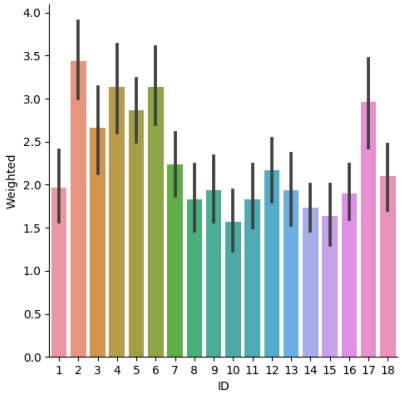


Figure 1: Female avatars weighted ratings.

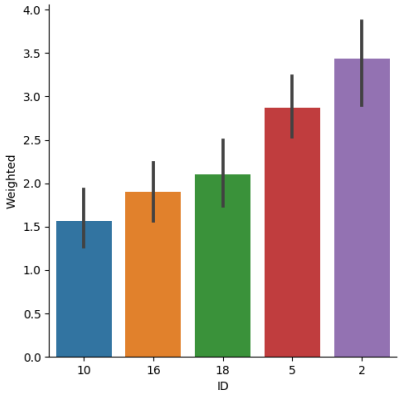


Figure 2: Chosen female avatars weighted ratings.

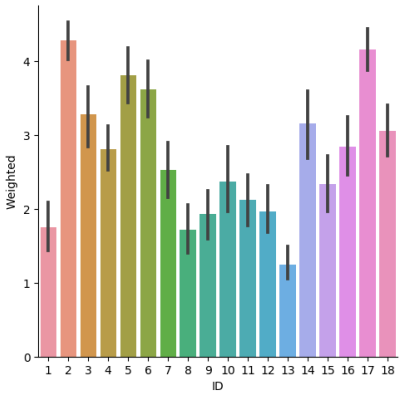


Figure 3: Male avatars weighted ratings.

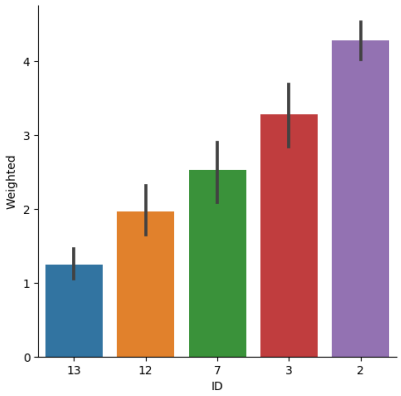


Figure 4: Chosen male avatars weighted ratings.

To determine the final avatars for the user study, we compounded the strength and intimidation rating and computed a final weighted score according to the following formula: $0.5 \times strength + 0.5 \times intimidation$. This value can be found in the **Weighted** column. As a reference for the other chapters, the chosen UMAs are marked in the table in the **UMA** column. The first part of the column value represents a unique identifier for the UMA design and the second part is the condition shorthand, where C1 denotes the weak condition, C2 average and C3 strong. The unique identifier is shown on the x axis in the above table as ID. Users generally found UMA designs unattractive. Female ratings for intimidation and strength were lower than male ones.
As expected, female ratings were overall lower than male ratings. This is a limitation of the UMA DNA ranges and difference in character enhancements magnitude between genders and races. Female upper and lower body muscle mass could be increased less than for males. Males had more defined muscle, while for female muscle tone decreased with muscle mass. Other downsides of increasing body mass for females was that leg muscles grew non-uniformly and appeared unnatural. Some other disadvantages are mentioned in the previous chapter. We consider the Morph Character System (MCS) as an alternative to UMA, however support for current Unity versions has been discontinued for this library.

4.1.5 Conclusion

For more control over avatar variation, we can consider a custom design for the avatars using a 3D modelling tools. UMA has the advantage of providing a simplified and fast method of generating many avatars with some degree of control over the body-related changes. Conversely, 3D tools are have a high learning curve and can be difficult for inexperienced users. To allow the same magnitude of change

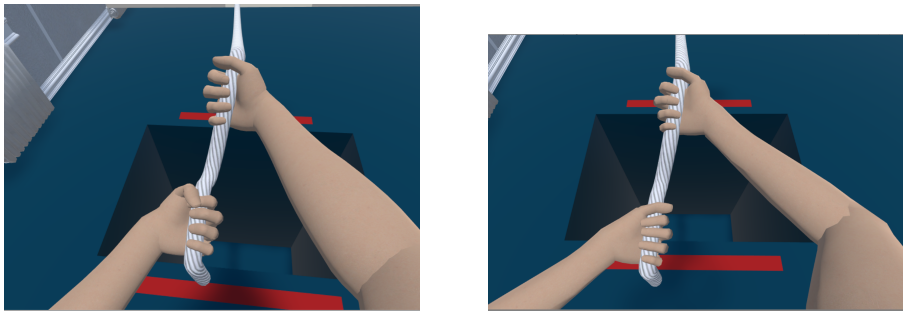


Figure 5: First person view of male (left) and female (right) arms.

for all body modifications, we would ideally implement a VR character creator with a simple user interface. With respect to appearance modification, other body enhancements such as piercings ca be considered.

4.2 Implementation

The tug-of-war game was implemented in Unity3D and integrated with VR through SteamVR. Auditory feedback was provided for the countdown before the rope-pulling in order to increase the appearance of a gaming situation. Sounds were provided from the laptop speakers which was located behind the black curtains. To give users a sense of depth and maintain performance, we use mixed lighting methods[3].
The avatars of the players were designed from the Morph Character System (MCS) female and male humanoid avatars [4]. The models were edited in Autodesk Maya [5] to remove parts of the mesh so that only the upper torso and arms were displayed. The original skin material was replaced with a custom design with less detailing to remove any uncanny valley effects [19].

In the game players could see their left and right hand holding the rope, which comprised their embodied avatar (seen in figure 5. The virtual grip of the hands on the rope was taken from a real hand grip of the rope using the HI5 gloves. The finger movement was disabled as the gloves were easily magnetized by the set up and would malfunction shortly after the beginning of the experiment. To prevent loss of ownership from random finger movements, players were told to maintain the same grip on the rope at all times during the experiment. Before starting, users would undergo the calibration of the gloves provided with the HI5 glove set up to maintain accurate hand orientation and position. Despite calibration efforts and manual adjustment trials, the undesirable finger movement and often inaccu-

[3]https://docs.unity3d.com/Manual/LightMode-Mixed.html
[4]https://connect.unity.com/p/morph-3d-morph-character-system-mcs
[5]https://www.autodesk.com/products/maya/overview

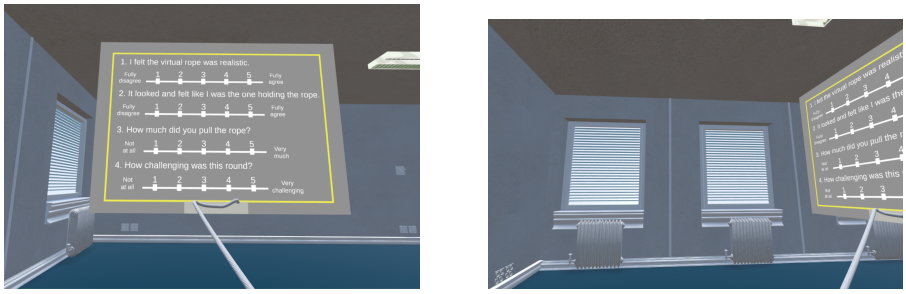


Figure 6: Panel with questions in VR, side and left view.

rate hand orientation and position proved to be the most difficult set back of this project. This was most challenging when integrating the gloves within an inverse kinematics VR algorithm to display user arm movement in a natural manner.
We used Final IK VR [6] to display naturalistic avatar body movements. The positions of reference for the algorithm were the left and right hand trackers, together with the headset position. This implementations does not always congruently display the position of the virtual arm and user arm. It is, however, an acceptable trade-off in to the lack of elbow tracking and full body tracking. The virtual rope was implemented using the Obi Rope asset [7]. Its settings were adjusted manually to simulate real rope movements through a trial-and-error approach.

Between rope-pulls we display a vertical panel with questions 6. We used TextMesh Pro[8] for the font, to allow for better readability considering the resolution of the headset. Between trials, participants always hold the rope in their hands. In order to still display the rope in users' hands when agents were not holding it, we placed it on the panel as if hanging there. To allow for natural transitions between experiment states, we used a black-out animation lasting one section. This animation fades to black, then fades back into the room scene. During this time, we spawned the questionnaire panel or the opponents. We designed the experiment scene to be as close as possible to the room where the experiment was taking place.
Every second the state of the experiment was logged in a file. Among others, we log users' the gaze target. This was implemented with a Raycast [9] originating at the center of the headset. Gaze targets are detected as objects with colliders. Since UMA's and some special parts of the setting (such as the hole) cannot have colliders, we added transparent panels with colliders on top of these objects.

[6] https://assetstore.unity.com/packages/tools/animation/final-ik-14290
[7] https://assetstore.unity.com/packages/tools/physics/obi-rope-55579
[8] https://assetstore.unity.com/packages/essentials/beta-projects/textmesh-pro-84126
[9] https://docs.unity3d.com/ScriptReference/Physics.Raycast.html

4.2.1 Agent Design

The virtual agents were created in Unity using the Unity Multipurpose Avatar 2 (UMA) asset [10] and o3n UMA Races [11]. We chose this library because the UMA character system has a high degree of avatar customization allowing avatar generation with relative ease. Programmers are able to vary body part sizes such as ear and brow rotations.
The manipulations described below serve to provide users with a more dynamic experience and make their opponents appear more life-like. We aim to generate high behaviour realism in order to make up for the lack of perceived agency. This serves the purpose of eliciting more realistic responses from people in accordance with the threshold model of social influence [8]. We placed UMAs close enough to participants so they noticed the changes in expression and other animation changes.

4.2.2 UMA Dna

We varied perceived strength in the design of the agents and used UMA DNA settings to generate a distribution of agents varying from strong-looking, to average and weak looking. UMA Dna represents a dictionary used to customize Unity multipurpose avatars (UMAs). Appearance can be changed on a scale of 0 to 1, with higher values being used to make adjustments more salient. A downside of this character system is that the magnitudes of these features do not vary linearly and can differ between races and genders. For example, female muscle size cannot be increased on the same scale as male muscle size. Due to this, male agents could have more muscle mass than females. Another example is that increasing lower muscle mass for women inflated the leg muscles in an unnatural way. Furthermore, there were color differences due to mesh materials and textures.
UMAs in the strong conditions generally had strength-signalling Dna parts closer to 1. Their values decreased to 0.5 for agents in the normal conditions and below 0.5 for agents in the weak condition. We considered the default 0.5 Dna values as average. We manually adjusted these values through trial and error to give UMAs the most human-like appearance and prevent them from looking exceedingly uncanny.

[10] https://assetstore.unity.com/packages/3d/characters/uma-2-unity-multipurpose-avatar-35611

[11] https://assetstore.unity.com/packages/3d/characters/humanoids/o3n-male-and-female-uma-races-102187

4.2.3 Strength Cues

As strength cues we used muscle tone and fitness, together with adjusted facial proportion as seen in [53]. Skin, eye and hair color, together with outfit designs were kept as constant as possible between races and genders. For the upper body clothing, the agents had a sleeveless shirt to increase the visibility of the arm muscles. With respect to facial manipulations, we varied facial width and jaw size which are known to be associated with testosterone in men [27]. We decreased the size of the lips and eyes in the strong conditions to increase face-width ratio. While height and age are the best predictors for female perceived strength [42], we applied the same variations to the female avatars. We did not adjust height for the agents displayed in the survey images. However, we manipulate height in the user study, increasing height for agents in the strong condition to 0.8 and average to 0.5, keeping avatars in the weak condition at 0.5.
As strength is closely related to dominance and aggression, we added variables meant to increase intimidation such as tattoos, military hairstyles and manipulated outfit colors. As in previous Proteus effect literature and associated body of work related to social psychology [58, 37], we used black clothing to elicit more aggressive attitude perceptions. Agents in the strong conditions had black upper-body clothing, those in medium conditions had gray and for weak agents we used white. While male tattoos are dominantly viewed, female tattoos generate mixed responses from observers [54]. As with previous research [52], agents were gender matched to avoid any tensions or similarity biases that would arise from cross-gendered evaluations. Please see section 9.9 for the agents chosen for the surveys.

4.2.4 Animations

In addition, we used the UMA Expression Player to even out neutral facial expressions between races and genders. The Expression Player was also used at the moment the UMAs were pulling the rope to give the impression that the opponents were exerting themselves or seemed angry. Their facial expressions were exaggerated to be more visible to participants in VR. Additionally, we use the blinking mechanism provided by the UMA Expression Player to make the avatars blink throughout the trials. We interpolate[12] between the changes in facial expression in order to blend them in a natural way.
We used the Mecanim animation system provided by Unity for animation layering and blending. At the start of each trial the opponents are shown with the rope in their hands in an idle posture animation, with their hands moving in the air and their chest moving as if breathing. This animation overlays all other animations

[12]https://docs.unity3d.com/ScriptReference/Vector3.Lerp.html

and plays throughot the trial. We use the *MORRO MOTION* Idle MoCap for this purpose [13]. After a 3-second countdown, UMAs are triggered with a rope-pulling animation. This animation is derived from the rope pulling animation on Mixamo [14].When UMAs reach their maximum-exertion posture, a tugged animation is triggered showing the agents moving their hands back and forth as if the rope was being pulled from them. For the remaining pull time, the UMAs maintain maximum pull posture. This serves the purpose of giving users the impression that their opponents react to the rope being pulled. Please see the figure in section **??** for participant's view of their opponents. From the end of the countdown, the rope-pulling lasts 10 seconds. At the beginning of each trial participants have 20 seconds before the countdown starts. This is to allow users to inspect their opponent and look around the room if desired. After the rope pulling ends, there is a 10 second delay before the screen fades the black and the question panel is presented.

4.3 Initial Design

Before starting the empirical study, we had a series of pilots to determine rope-pull quality, force measuring feasibility and opponent realism. We also collected feedback about rope-pulling duration and animation styles. For the piloting, we had a gender-matched, within-group experimental design. Initially, the game consisted of three of rope-pulls with questions in-between. The rope was tied to a box with a force meter, and some elastic bands were placed between them, to give the impression of resistance. The box was set on a table to prevent damage to the force meter. Additionally, this would prevent participants from feeling a weight on the box between rope-pulls. The VR room was made as similar as possible to the experiment room. There were two rounds of piloting. For this, we recruited participants from the university campus. Each round resulted in several modifications to the original design of the experiment and implementation of the game. We detail these changes below.

4.4 Piloting

For the first piloting session we had 2 participants, 1 female, aged 23-25. Participants had observations with respect to the way their opponents were holding the rope, mentioning that one looked like it was "holding a rod" in their hand. They also remarked on the distinction between the avatars' strength was too obvious and showed high awareness of the experimental purpose.

[13] https://assetstore.unity.com/packages/3d/animations/idle-mocap-28345
[14] https://www.mixamo.com

Animation wise, initially opponents pulled the rope at the start of the countdown and resumed initial position at the end of the trial. One participant mentioned that they "expect her to react when [...] pulling".
After this session, we decided to increase the number of rope-pulls to 5, and add 3 additional agents to represent low-average, upper-average and average strength. With this, we wanted to remove any obvious strength separation between the agents. We redesigned the grips the opponents had on the rope and added two fixed points on the opponents' thumbs and pinky finger, to make the grip seem more natural. Additionally, we added a tug animation after the agents start pulling the rope, and overlayed an idle breathing and moving animation throughout the trials. The sequence and design of these animations are further explained in section 4.2.4.
Regarding the rope set up, we noticed the force meter was placed inadequately inside the box and participants would rotate it when pulling. This could affect their perception of the forces acting on the rope and it was undesirable. We subsequently used a bigger box and placed the force meter and camera in the middle, to balance the weight remove rotations. Furthermore we noticed the elastic bands were stretching and breaking. Therefore, we used a spring and an elastic rope for the final experiment.
A second pilot was ran with 2 males participant, aged 28. Seeing the rope being pulled back, one participant mentioned that "[the opponents] reacting to me is nice". In this session, one participant asked if they were pulling on a box. We noticed the sound of the box hitting the table was drawing their attention to the rope setup. To remove this, we placed a blanket on the table to dampen any sounds from the rope or other attachments. The participant also mentioned that they wished the game lasted longer, and so we increased rope-pulling duration to 10 seconds from 6 seconds. We also increased the time spent looking at avatars before starting the countdown to 20 seconds from 10 seconds. One participant also mentioned that "[starting he competition] felt too sudden" and would like "more time to adjust and look at hands, take in the surroundings." As a consequence of this, we made an additional scene that participants would see before starting the actual game. This was the questionnaire part, between rope-pulls. Participants would first see the panel of questions, their hands and the rope for 1 minute before starting the competition.
Additionally, we noticed that participants were pulling the rope too hard and moving the set up, if allowed to use their legs. The first set of pilots were not allowed to use their legs and the second one was allowed. This was problematic because, upon noticing that they were pulling too hard, participants would restrain themselves and the first trial would have a higher force distribution than the rest. We decided to tell participants not to move their legs, if they can.
To test if the elastic addition to the rope makes any difference, we asked a partic-

ipant from the last pilot to play the game two times, one with a rope and elastic resistance, and one with the rope tied to the meter directly. The participant mentioned that it was “much better” with the elastic band because it gave the impression of “some resistance”.
Other miscellaneous set up decisions were taken:

- For the glove calibration, we changed the instruction panel to appear in front of the users, so they would not need to turn around to calibrate the gloves. This change was needed because participants were turning around and moving away from the preset experimental location and interfering with the base stations;
- We placed the desks in the room away from the rope-pulling location;
- We relocated the laptop containing the post-experimental survey away from the base stations;
- We relocated the base stations and placed them higher, as we noticed taller participants would get between them and cause loss tracking issues;
- We placed papers with directions to the experiment room around the experiment location.

4.5 Setup and Measurements

For the experiment we used a Predator Helios 300 Acer laptop with an Nvidia GeForce GTX 1060, VRREADY graphics card. VE immersion was achieved by using a HTC Vive headset with a resolution of 2160x1200 and a refresh rate of 90Hz. Players’ hand movements were tracked by the Noitom HI5 Vive tracker gloves[15]. The glove uses a wrist-mounted Vive tracker to generate users’ presumed hand position. In addition to the subjective measures presented in the previous section, we gather quantitative data namely the maximum pulled force for each rope-pull. We measure the force of participants’ pulling through a digital force meter with a seven-segment display. The force meter was fastened inside a box and placed on a table. We recorded the force by filming the digital display with a webcam, Intel RealSense VF0800 Developer Kit Digital Depth Camera[16].
To hide the measuring mechanism, participants pulled a rope between two black sheets which hid the whole set up of the experiment. This setup is shown in figure 7.
The force meter was placed on a table and it was tied to a heavy piece of furniture. To dampen the sound of the box hitting the table, a piece of cloth was placed under

[15]https://hi5vrglove.com/
[16]https://www.intelrealsense.com

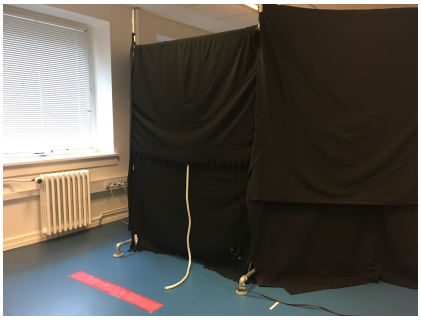


Figure 7: Participants' view of the setup.

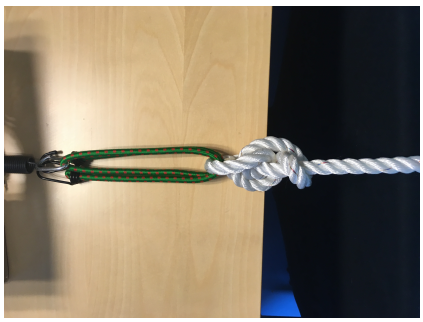


Figure 8: Rope and spring tied to the meter.

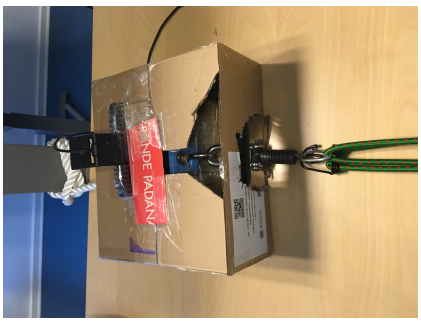


Figure 9: Box with force meter and camera.

the box. Various amortizing materials were placed inside the box to absorb sound and prevent the force meter from being damaged. The rope was tied to an elastic cable connected to a strong spring. The spring was connected to the force meter. The elastic band offered less resistance and, as participants pulled, the spring increased resistance. This spring mechanism was designed to increase realism and offer users some resistance in order to give the impression of a force acting on the rope. The box and spring mechanism can be see in figures 8 and 9.

4.6 Procedure

We greeted participants and briefly explained the broad purpose of the study. After that, we gave them two consent forms to fill in and further explained the details and constraints of the tug-of-war game. Before giving participants the VR headset and gloves, they filled in a short survey with demographic data. The experimental procedure is fully explained in section 9.2.
Before starting the game, we calibrated the gloves for participants' hands. For each rope-pulling trial, participants were told to start pulling after the end of the countdown, when they see *Start*, and keep pulling until they see *Stop*. They also

received audio feedback for the countdown. Participants were told to follow two constraints: always keep their hands on the rope and try pulling the rope without moving their legs. Between rope-pulls, participants answered the questions on the quiz panel (figure 6) from within the VE. After 5 rope-pulls, the game ends and participants are asked to fill in the post-experimental survey. The survey questions of the avatar appearances were gender-matched. Before participants left, we collected feedback about the game and then debriefed users about the true scope of the study.
For each trial, users pulled the rope for 10 seconds. Before that, they were in the VE with their opponent for 20 seconds. After the pulling ended, they remained in the same VR state for an additional 10 seconds. During this time, users could inspect the appearance of the avatars. Only after that, participants were shown the quiz panel. Transition between the game states (rope-pulling and quiz panel) was realized through a black splash screen. When the screen faded to black, the objects in the room changed. The screen fades back when the changes are complete. We did this in order to maintain a natural state of the room and prevent breaks in presence and immersion. Further details about the implementation are presented in section 4.2.
Throughout the piloting sessions, many of the difficulties of running this study were revealed. The most challenging task would be synchronizing data capturing with starting the game and giving all necessary information to participants. If VR set up errors occurred, such as base station loss of sync or tracker issues, the experimenter would need to ad-lib until starting data capturing, resume experimental procedure and repeat important instructions. We incorporate all this knowledge in an *Instruction Sheet* the experimenter must follow to give participants the same experience of the study. This sheet can be seen in section 9.2. Additionally, participants were required to sign a consent form (presented in section 9.3) which broadly described the experimental setup, method and purpose. In it, we ask for their consent to store recordings and other experimental data in an anonymized form. Both forms were modified from their initial structure as a result of piloting. The purpose of evaluating strength of pull based on avatar appearance was not mentioned in the consent form. Instead, broader terms such as *virtual reality interactions* were used.

4.7 Participants

In order to give the appearance of a gaming study, we designed a poster to recruit people for the experiment. We placed this poster (9.10) around the university campus, and used it for recruiting messages. For the user study, we recruited a total of 28 participants (17 female) from the university campus and local Facebook groups. They were between 21 and 34 years old ($M = 24.42, SD = 3.06$).

Participants were unable to take part in the study if they had previously completed the avatar appearance survey. We discarded the results of 2 participants (1 female) due to invalid data. Of the remaining ones, 9 participants had never used VR before and 16 had never played real-life tug-of-war. Participants were rewarded with a gift valued at 100 danish kroner for taking part in the whole study.

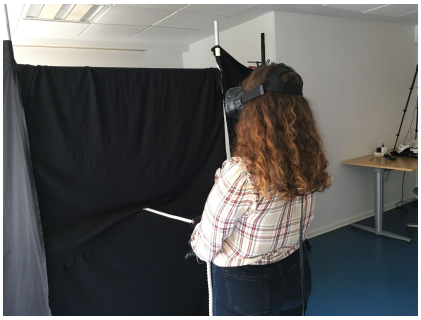



Figure 10: Participants playing tug-of-war VR .

5 Results

To measure the maximum force, we looked at the force meter data from the end of the countdown, at the beginning of the *Start* message, until the end of the *Stop* message. All force data presented is measured in kilograms.
In the plots below, we present data in terms of ordering and condition. By ordering, we mean the order in which participants saw the appearances, labelled as *Trial*. By condition we mean the condition of the opponent participants faced. The condition captures the independent variable and represents a numerical mapping from 1 to 5 to the appearance variable, increasing by strength/intimidation. Condition 1 refers to the weakest-looking avatars, according to the weighted score (see 2,4). Condition 5 refers to the strongest looking avatars. We also display our results by gender.

5.0.1 Conditions per Trial

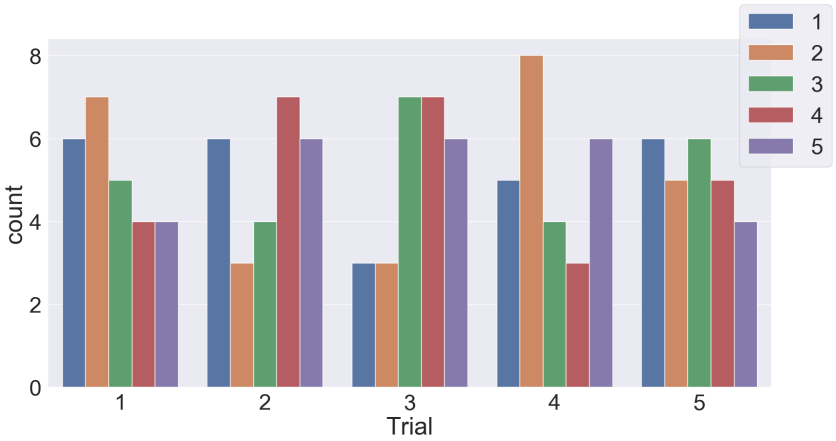
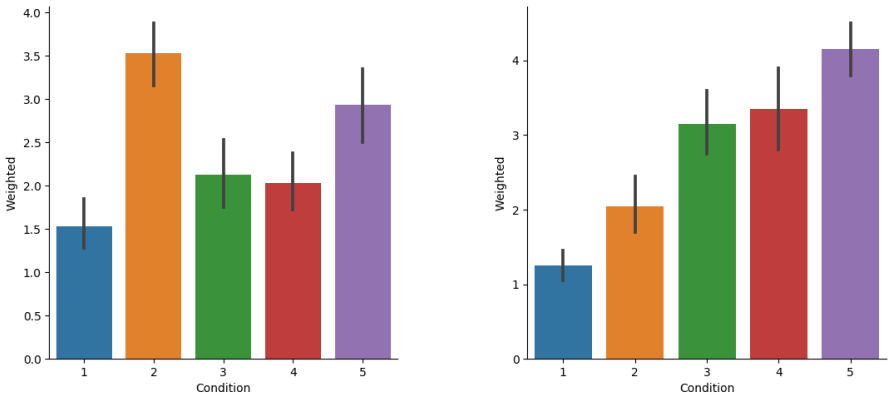


Figure 11: Number of conditions per trial.

In the above pictures we show the distribution of each condition for each trial. Since the conditions were randomized, there may be an effect of over-representation. The fifth trial seems to be the most balanced, representation wise. We had more female participants and thus more female avatars were evaluated. For an overview of the distribution of conditions per trial for males and females please section 9.1 figures 32 and 31.

5.0.2 Appearance Ratings



(a) Ratings for female avatars.

(b) Ratings for male avatars.

Figure 12: Weighted ratings by condition for the opponent avatars in the user study. Error bars show 95% confidence interval.

Participants rated the appearance of same-gender avatars on attractiveness, intimidation, intelligence and strength. In the above figures we present the weighed results of those ratings for female (12a) and male (12b) avatars. The weighted value is computed as: $weighted = 0.5 \times strength + 0.5 \times intimidation$. Male participants rated the avatars as expected, increasing in intimidation and strength by condition. For females, the avatar in condition 2 (low-average) was rated as highest in intimidation and strength, while the avatar in condition 5 (strong) came second. All ratings for these avatars and their thumbnails can be found in section 9.8.

5.0.3 Qualitative Feedback

Our observations and qualitative feedback show that people experienced various degrees of physical illusions. Participants initial comments suggest that the illusions were stronger in the first few trials. As the game went on, they seemed to become aware that there is no force acting on the rope. Realism was often quoted as a reason for the breaks in illusion. With respect to this, participants suggested improving the animations and making their opponents react to their pull. The virtual rope was seen as highly realistic and holding the rope seemed to contribute to this belief.

In total, 4 participants showed high awareness of the purpose of the experiment ("how people react when they see different people in front of them"— P22). Among them, a few made comments about their own behavior, which seemed to support our research question ("for the people that looked stronger I had to pull more"— P2).

Participants generally seemed to have an expectancy that the stronger the avatar looks, the more they should pull ("if you expect a stronger person you will pull stronger"—P14). P16 mentions " I felt like the body shape was according to how to pull. [I was] pulling more strong for someone bigger and less for someone smaller". In the case of P26, they, mention intimidation: "I was expecting I would feel the need to pull harder because the characters were more intimidating from their look".

With respect to the challenges, some participants clearly state feeling a distinction: "[for one avatar] felt like it was easier"—P19, "felt actually pulling like there was another person on the other side"—P16. Furthermore, our qualitative feedback suggests the illusions were breaking as trials went on. In the case of P13, her feedback in the first trial was that: "It's weird - I can feel something pulling". Eventually, in the post-experimental interview she mentions that "sometimes they were pulling without me feeling". P4 remarks: "the machine did not make much effort pulling back" and that "[it] was not consistent with the guys [referring to

opponents]". However, in the end they state:"[it] lost realism when I realized it isn't pulling back". P15 remarks feeling a difference in the first trials, but not in the last ones — "didn't see a difference in challenge in the last 2 ones". P9 comments: "I can tell the difference between the challenges". However, by the end they conclude referring to the opponents that it "didn't feel like they were pulling". One participants adds that the appearances made him think about his own abilities: "I think it was the same strength but, in a way, since you are not prepared for that, maybe you feel weak" (P18).

Some participants were unsure if they had felt any difference and used ambiguous language to suggest they felt little to no variation: "can't tell if there was a difference or not in pulling"—P15, "more or less the same strength"—P24, "did not feel that much of a difference in enemies"—P28. P8 then adds that it seemed like they were "receiving same amount as putting into it". They admit feeling "resistance", however it was "difficult to tell". Acknowledging some perception incongruities, P25 says that: "I wasn't sure if I should feel the pressure, but I could see them pulling so you created that impression a bit".

Others mentioned not feeling any difference throughout the experiment: "didn't feel physical challenge of a counterattack" (P17), "feels like it's the same challenge" (P12). Often, participants referred to lack of realism when pointing out downsides of the game: "when I was trying to pull harder, [I] didn't have my body being pulled"—P17. They were mostly disappointed that their opponents were not reacting properly to their pull: "always leaned back no matter how much I pulled"–P23, "if you pull harder you expect them to come closer, lose balance or pull harder"—P14. They suggested "make[ing] the other avatar move or fall" (P7) or adding sounds: "avatars made it look more unreal, it looked silent; make them talk with some sounds" (P16). When their expectations about the game were not met, the realism of the experience decreases ("didn't feel real because they were not pulling stronger when they looked stronger"—P9). Some participants noticed that the avatars were animated in the same way, which also had an impact: "avatars doing the same, kind of the same move, doesn't feel realistic"—P11. Additionally, users desired their opponents to be more *expressive* (P27). Beyond avatar design, a participant mentions that "being able to also move your feet would be good" (P11). One participant disliked that the textures and shading in the game were not detailed enough (P6). In their opinion, these design enhancements were required to generate a realistic setting.

In contrast, there were some participants who found the experience highly realistic "Felt real like really against an opponent" (P1). The rope was often mentioned as a good source of realism for the game:"Felt more real because I was holding something"–P9. The haptic feedback seemed to made the virtual rope even more real, one participant mentioning: "I didn't look as much at my hands as I thought [...] I could feel the rope ". P16 remarks: "I played VR before, but never like

pulling an object and seeing it in VR [...] It helps that you feel the rope and see the rope". Few mentioned any downsides of the rope ("rope was too elastic"—P15).
Participants did not generally make comments about their arms. Some noticed inaccuracies in their avatars: "Arms seemed off when I looked straight at them"—P4. For example, P1 had a highly realistic experience ("I really had a feeling like I was in that room") , but mentions the arms *twisted around*, and "everything else felt more responsive than the arms".
We also noted that sometimes participants dynamically changed their standards for the challenge ratings: "I'm getting tired more so I will give it a 5"—P11. P10 mentions that he begins to give 5 challenge ratings because of lack of feedback:"I feel like I'm not making any progress; they are not moving at all ".

5.0.4 Force Meter Data (H2)

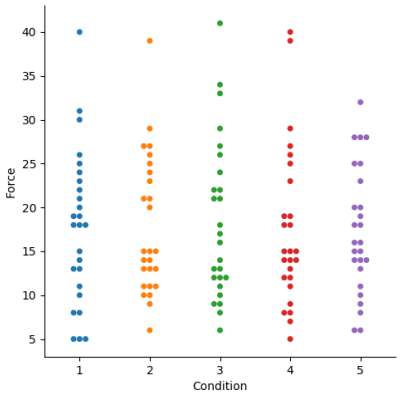
Despite positive qualitative feedback, there were inconsistencies in how strong participants pulled. We expected some variability because users interacted with the system in unforeseen ways. These quantitative results, however, do not appear to support H2.
In the above figures we show an overview of all the data measured from the force meter, namely the maximum force pulled for each trial (kg). For each participant, we normalize this data in two ways, N_{first} and N_{max}^{min}. N_{first} shows the relative force to the first trial. For N_{max}^{min}, we normalize values between 0 and 1 with min-max normalization.
The force meter measurements do not seem to support our hypothesis, that the stronger avatar looks the harder people would pull (H2). The differences between trials are not very significant and the error bars vary. The data shows participants pulled slightly more in the third condition (average). In the tables below we present mean and standard deviation per trial and condition. For all force data we have: $M = 17.76153,\ SD = 8.29289$. From the tables below, there appears to be more variation by condition than by trial. Also, participants seemed to pull the strongest in the second trial and for condition 3 (average-looking).

Cond	Mean	SD
1	17.73076	8.70658
2	17.76923	7.88064
3	18.461538	9.06523
4	17.5	9.06090
5	17.34615	7.20523

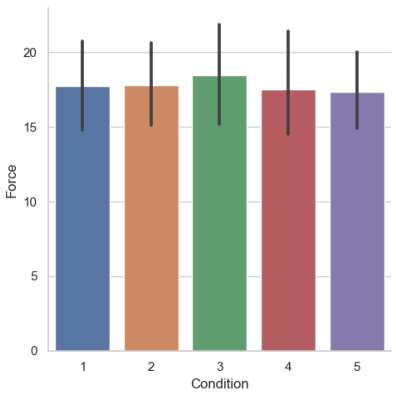
Table 1: Mean force and standard deviation by condition.

Trial	Mean	SD
1	17.69230	8.75794
2	18.65384	8.42350
3	17.88461	8.19427
4	17.19230	8.37147
5	17.38461	8.28529

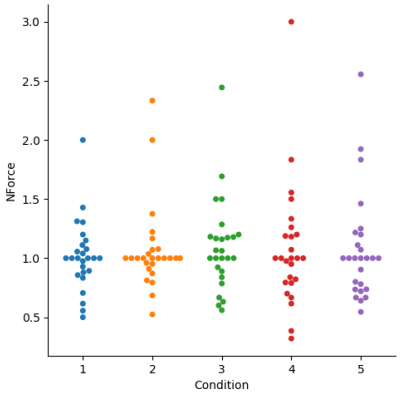
Table 2: Mean force and standard deviation by trial.



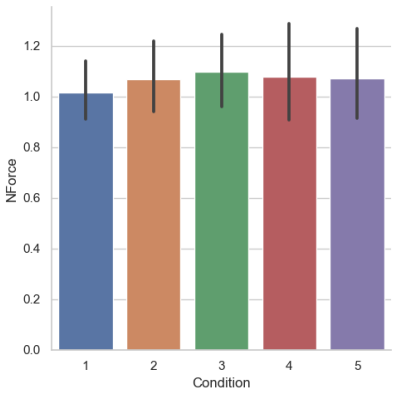
(a) All force values by condition.



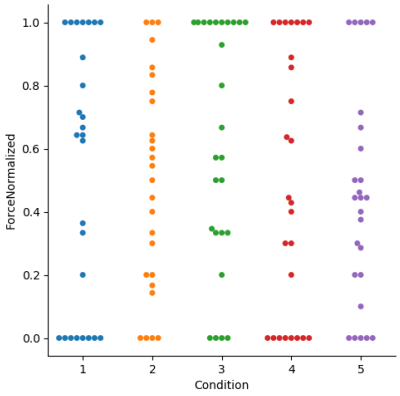
(b) Mean force by condition



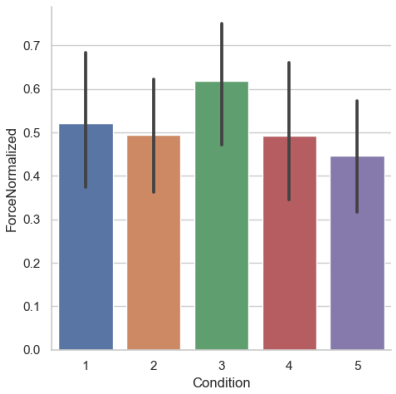
(c) All N_{first} by condition.



(d) Mean of N_{first} by condition.

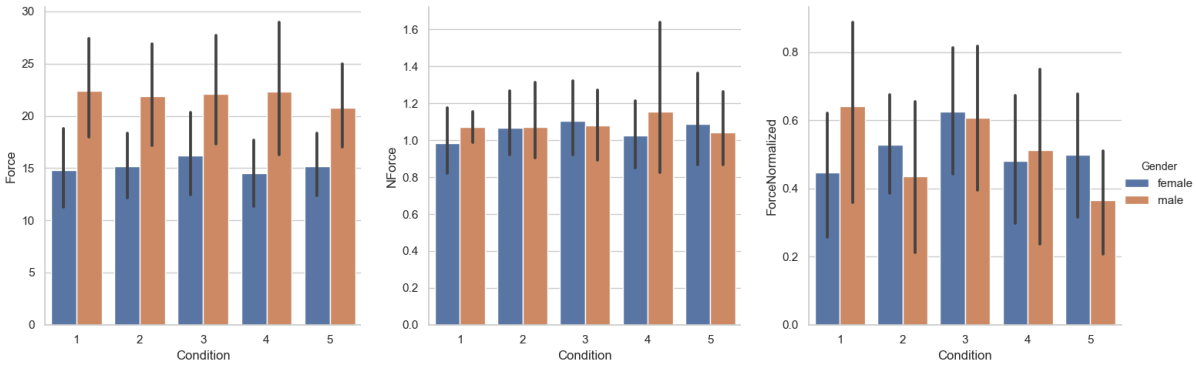


(e) All N_{max}^{min} by condition.



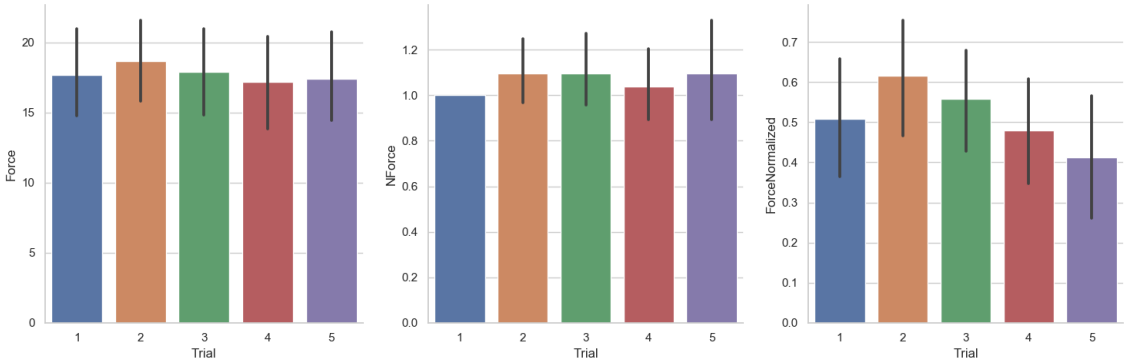
(f) Mean of N_{max}^{min} by condition.

Figure 13: All force (kg) values by condition. Error bars show 95% confidence interval.

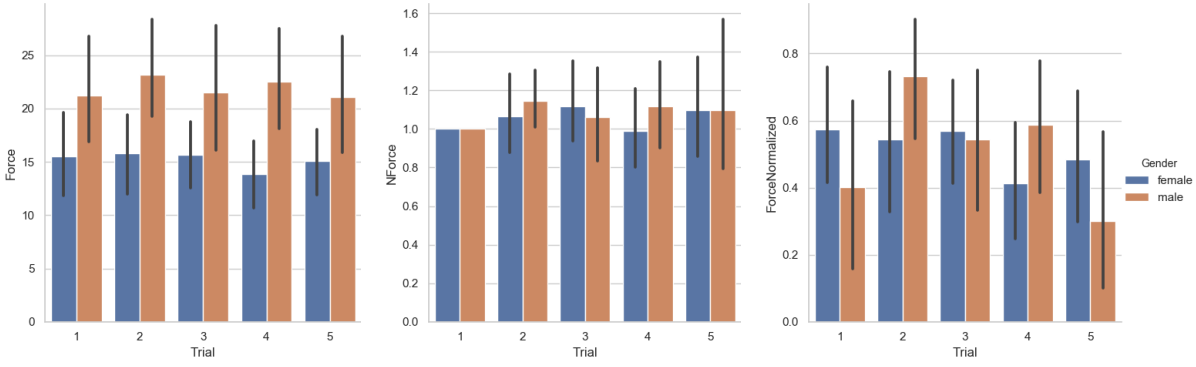


(a) Mean force. (b) Mean N_{first} force. (c) Mean N_{max}^{min} force.

Figure 14: Mean force, N_{first} , N_{max}^{min} by condition and by gender. Lines on bars denote confidence intervals.



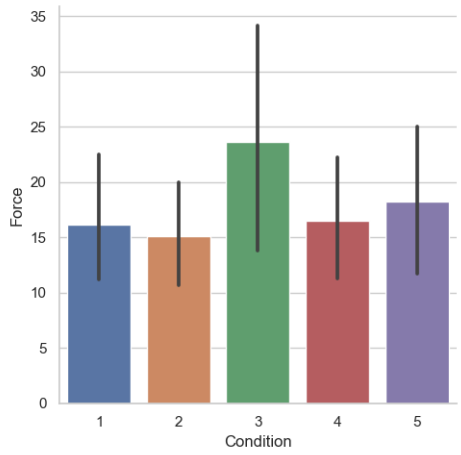
(a) Mean force. (b) Mean N_{first} force. (c) Mean N_{max}^{min} force.



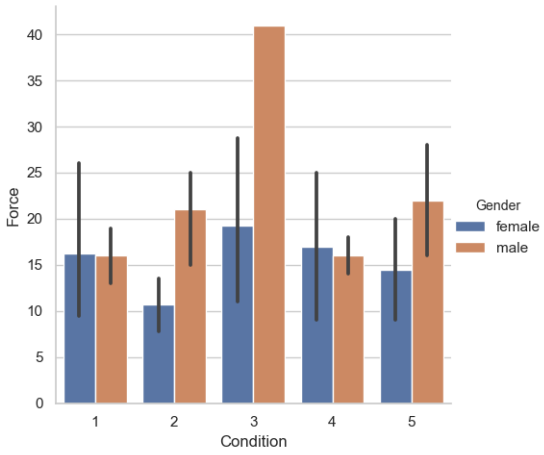
(d) Mean force. (e) Mean N_{first} force. (f) Mean N_{max}^{min} force.

Figure 15: Mean force, N_{first}, N_{max}^{min} by trial, and by gender in the second row. Error bars show 95% confidence interval.

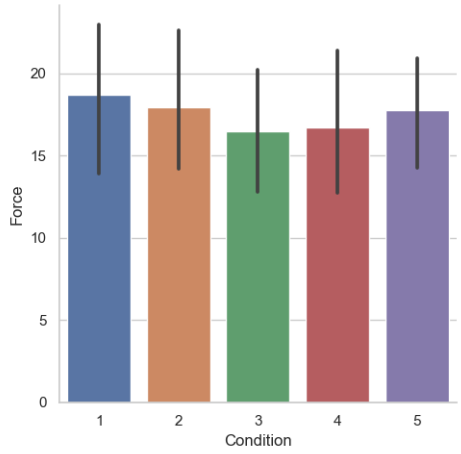
From figures displayed in figure 14 we can see males pulled more than females. Additionally, the error bars seem to be overall bigger for males than females. Error bars are biggest for N_{max}^{min} due to the normalization procedure, as minimum values are considered 0 and maximum values 1. Male participants seem to have the biggest error bars in condition 4.
In figure 15 we display our results by trial. Both males and females seemed to have pulled harder in the second trial. We look at the distributions of avatars per trial in order to determine if the effect is due to order or representation. Figure 11 shows that in the third trial there were the most avatars in condition 3. The high pulling values would, therefore, more likely occur because of condition than trial order, as participants seemed to pull hardest for the average condition.
While an effect of ordering seems to be less strong than appearance, we observed some design issues for the first trial. Participants reacted in various ways due to the novelty of the experience or the setup. As such, we isolate the first trial to look for discernible differences.
We examine our results considering only the first trial in figures 16a and 16b. In figures 16c and 16d we look at our results excepting the first trial. We notice there are 2 trends between these force pulls. In the first trial, females' pull on the rope increases until the third condition, then decreases. They pull the least in condition 2, which was the avatar rated most strong and intimidating by women participants (see 12a). There is only one data point for males in the average condition, which seems to be an outlier of very high force. It would explain the high value for condition 3 for all participants in figure 16a. Males' pull increases again for the last condition. When we consider trials 2,3,4 and 5 there is also an increase in pull for women in condition 5. Overall it looks like the first trial has an uptick in the force for the average condition, something not present when considering only the last 3 trials.



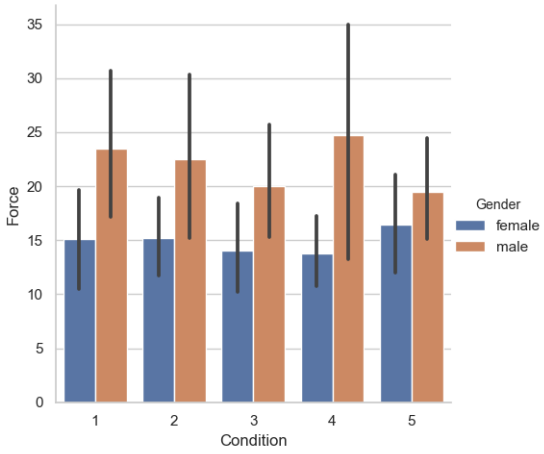
(a) Mean force in 1st trial.



(b) Mean force in first trial gendered.



(c) Mean force in all trials except first one.



(d) Mean force in all trials except the first one gendered.

Figure 16: Mean force by condition and by gender, taken for the first and first 3 trials. Error bars show 95% confidence interval.

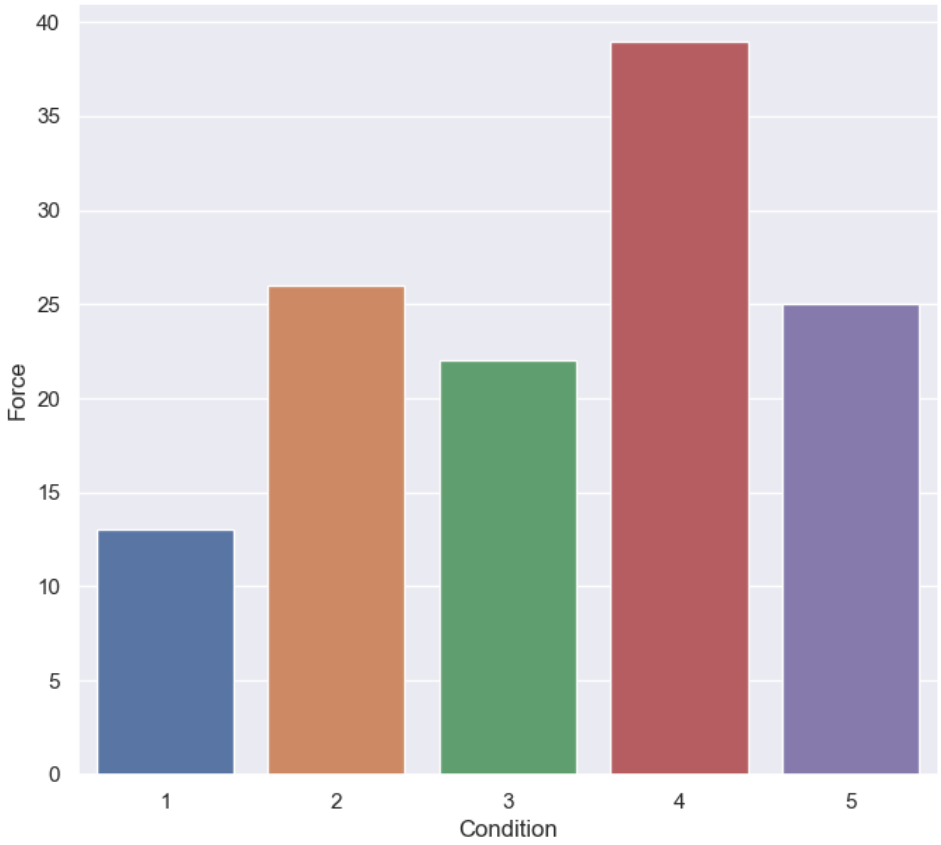


Figure 17: Force distribution of participant 4.

Participants did not appear to use force in accordance with our hypotheses (H2). However, our data shows some users varied their pull very much. On average, the difference was: $M = 8.61538, SD = 4.94995$. However, there are several outliers. The maximum pull difference was for participant 4 — 26 kg of force. Their force distribution is presented above in figure 17 With our hypothesis, we presumed there would be differences in pulling, however such large differences suggest participants were not simply reacting to their opponent. Rather, some other factors seem to have influenced their pull strategies. Qualitative feedback offers some insight into this behaviour. These large variations support our observations that participants tested the application. They behaved in an unexpected way in order to test their own assumptions about the opponents, not always respecting the instructions of the experimenter.

5.0.5 Perceived Pull and Challenge (H3,H4)

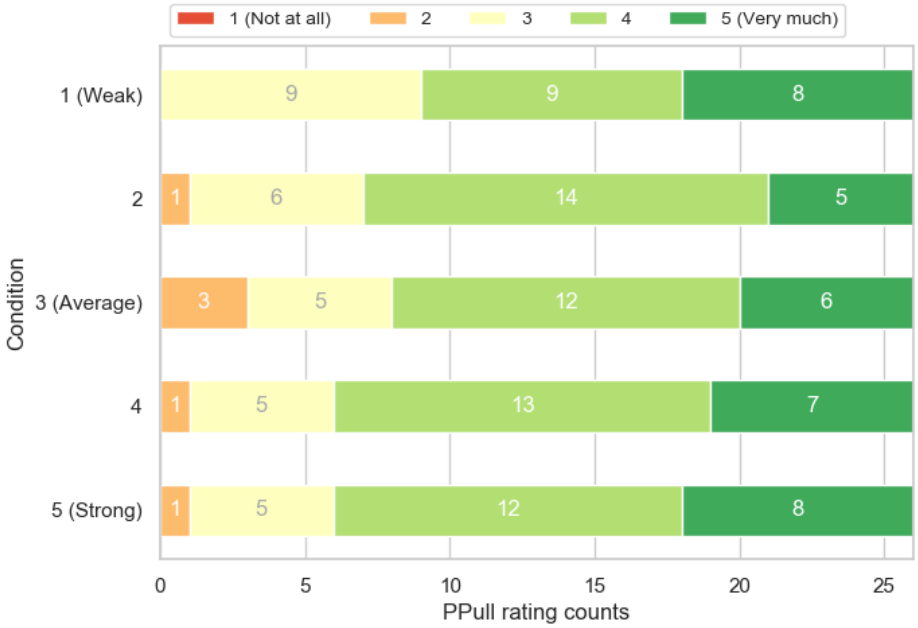


Figure 18: Count of perceived pull ratings by condition.

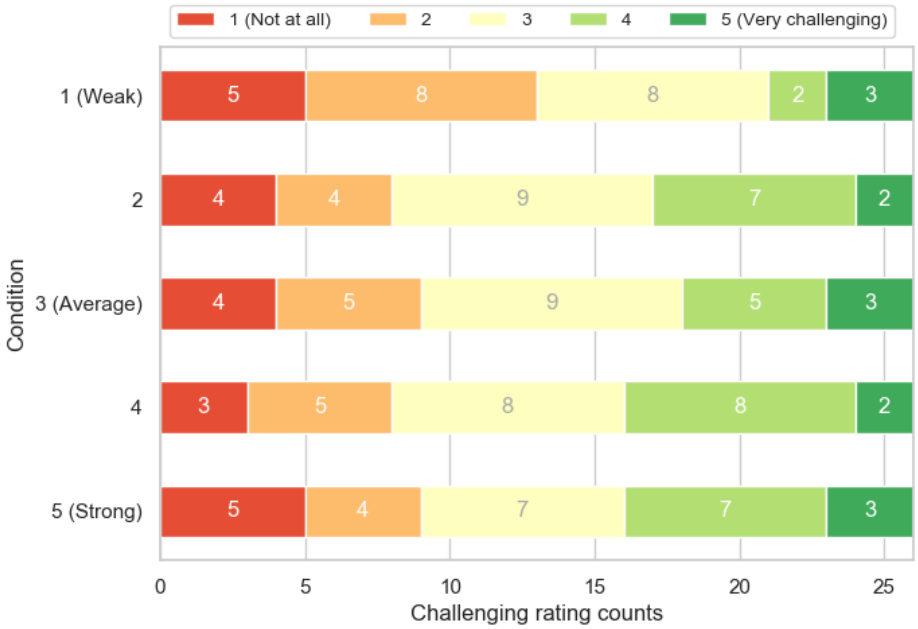
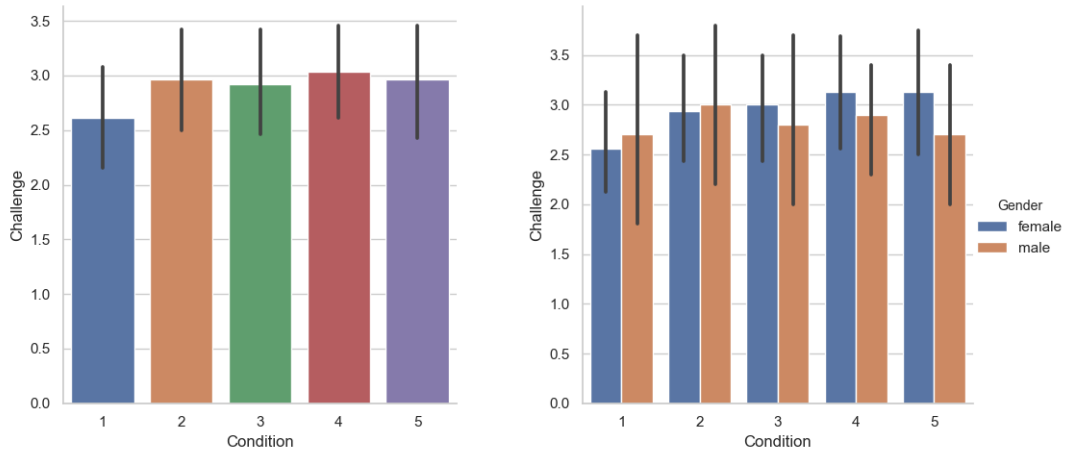


Figure 19: Count of challenge ratings by condition.

The majority of participants seemed to rate opponents with different challenges, which gives support to H1. Their reported values do not seem to coincide with our assumptions from H2, H3 and H4.
Above in figures 18 and 19 we present an overview for the ratings given to Q3 and Q4 between rope-pulls, namely challenge and perceived pull (see 3.2).

- **Q3**: How much did you pull the rope? Rating: 1 (*Not at all*), to 5 (*Very much*);
- **Q3**: How challenging was this round? Rating: 1 (*Not at all*), to 5 (*Very challenging*).

To see the total ratings for these metrics per participants, please refer to section 9.6, for perceived pull (33) and challenge (34).



(a) Mean challenge by condition.

(b) Mean challenge by condition gendered.

Figure 20: Mean challenge by condition, and by gender. Error bars show 95% confidence interval.

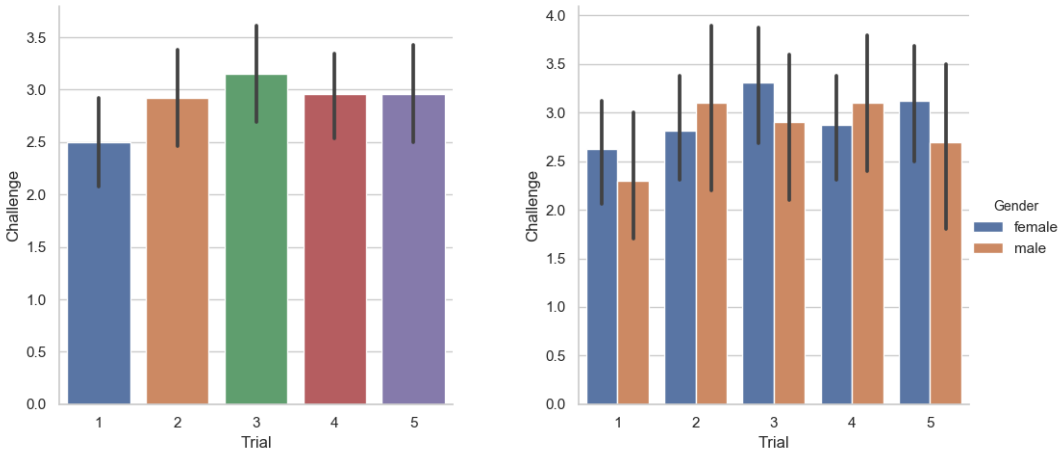
Most participants did give different ratings across trials, despite the rope having the same resistance ($M = 2.9,\ SD = 1.21265$), which appears to support **H1**. In tables 3 and 4 we display the means and standard deviation of challenge ratings by condition and trial respectively. Figure 20 shows mean challenge given by participants for each condition. In womens' case, it would appear to support our hypothesis, however women rated condition 2 as being strongest. Males had a downward trend when rating challenge. However, the confidence intervals for some condition indicate large variation. We can see in table 3 that condition 5 had the highest standard deviation. Participants also mentioned being unsure about the feedback of the rope. The variation in the challenge rating seem to reflect participants' uncertainty.

Cond	Mean	SD
1	2.6	1.2
2	2.9	1.1
3	2.9	1.2
4	3.0	1.14
5	2.9	1.3

Table 3: Mean challenge and standard deviation by condition.

Trial	Mean	SD
1	2.5	1.1
2	2.9	1.2
3	3.1	1.2
4	2.9	1.1
5	2.9	1.2

Table 4: Mean challenge and standard deviation by trial.



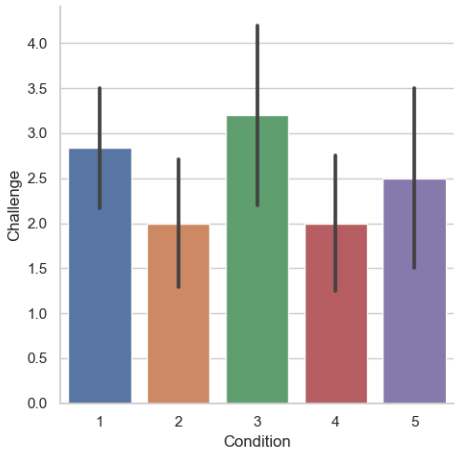
(a) Mean challenge by trial.

(b) Mean challenge by trial gendered.

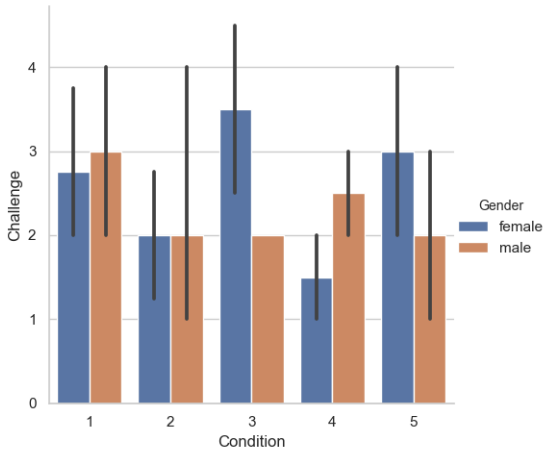
Figure 21: Mean challenge by trial, and by gender. Error bars show 95% confidence interval.

Figure 20 shows mean challenge ratings by trial. We can see the challenge increase until the second trial for males, and until the third trial for females. Results seem to vary more by trial which could support our observation that appearance had a stronger effect than ordering. We inspect the results for the first trial only in figure 22 below. These results seem to vary with the first condition being rated more challenging overall by males and females. The challenge ratings for the first trial seem to follow a similar trend as the force values. We note that in the first trial there were more avatars in condition 1 and 2 which could suggest an effect of over-representation.

We observed that participants did not use the same standard for the challenging rating. Furthermore, some changed their assumptions dynamically during the game, based on visual feedback or their internal state. We detail these observations in section 5.0.3. A limitation is that we did not gather the assumptions participants had about this rating, however we noted their voluntary comments.



(a) Mean challenge for 1st trial.



(b) Mean challenge for 1st trial gendered.

Figure 22: Mean challenge by condition, and by gender for first trial. Error bars show 95% confidence interval.

The majority of participants also reported pulling differently, despite competing in a game to pull the strongest ($M = 3.93846,\ SD = 0.82362$). In tables 5 and 6 we show the mean perceived pulls by condition and trial. Participants seemed to pull stronger in the last conditions, despite rating them less challenging. It appears they were also more consistent with their answers in this metric, as standard deviation for is less for perceived pulls.

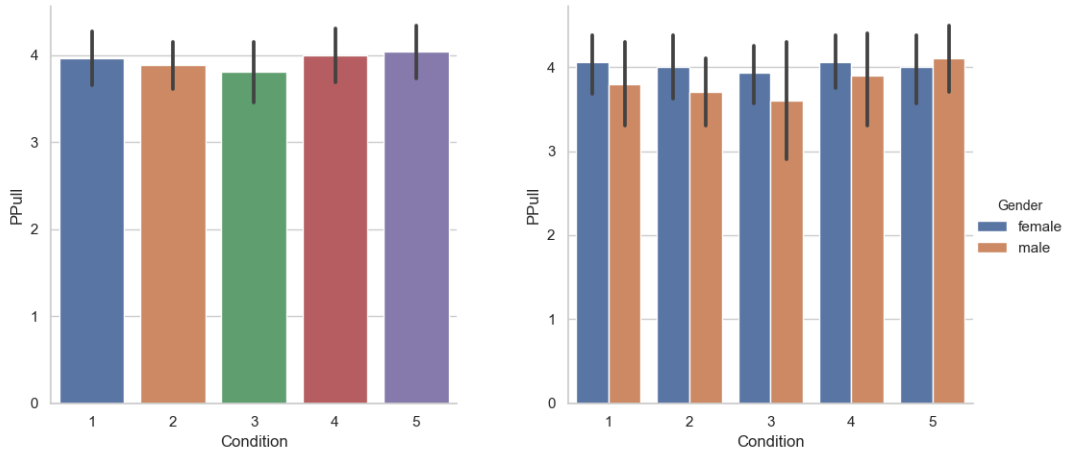
Cond	Mean	SD
1	3.9	0.8
2	3.81	0.7
3	3.8	0.9
4	4.0	0.8
5	4.0	0.8

Table 5: Mean perceived pull and standard deviation by condition.

Trial	Mean	SD
1	3.7	0.8
2	4.1	0.6
3	3.7	0.9
4	3.9	0.7
5	4.1	0.7

Table 6: Mean perceived pull and standard deviation by trial.

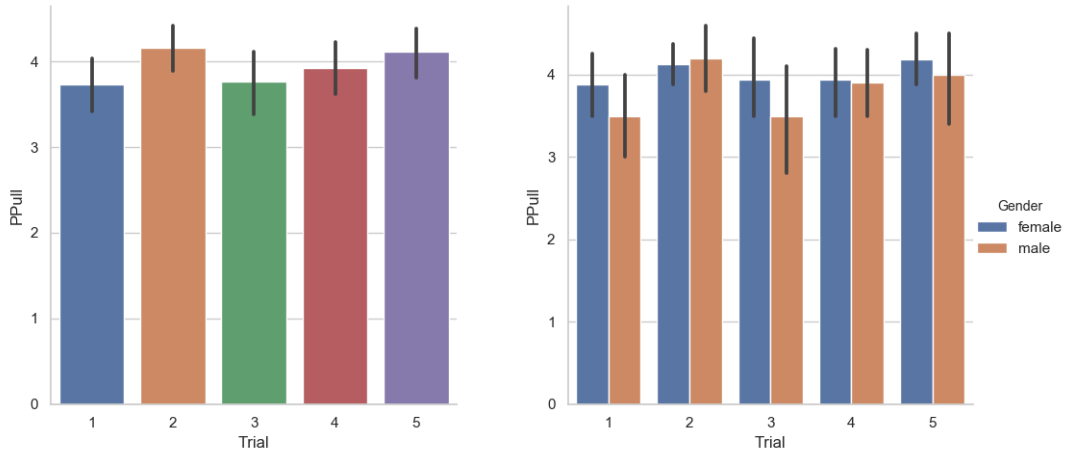
In figure 23, below, we show mean perceived pull values by condition and trial. By condition there seems to be a trend for the perceived pull. Namely that it seems to decrease until the average condition then increase again. This trend could support our hypothesis, namely that participants believed they pulled more for stronger-looking avatars.
The challenges and perceived pulls show two different trends. While challenges seem to peak at the average condition and look similar to the actual force, perceived pulls appear to decrease at condition 3.



(a) Mean perceived pull by condition. (b) Mean ppull by condition gendered.

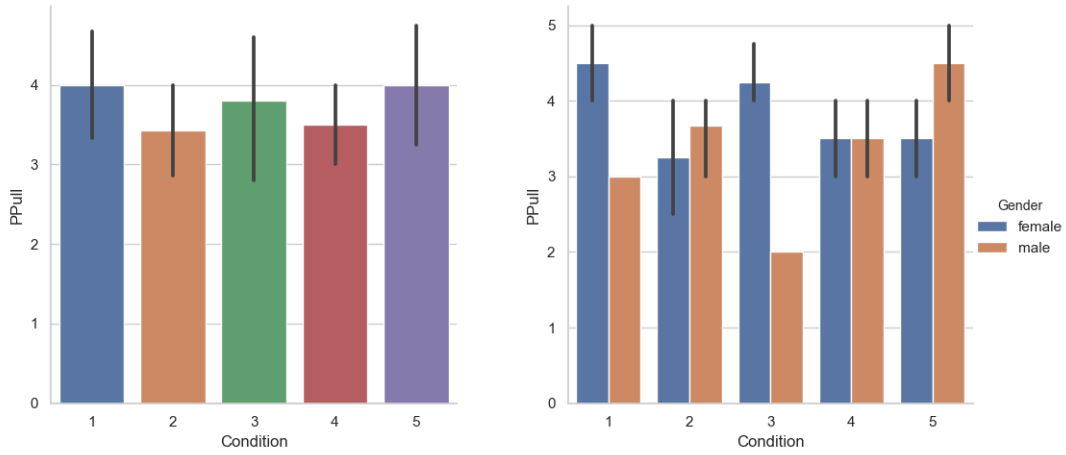
Figure 23: Mean perceived pull by condition, and by gender. Lines on bars denote confidence intervals.

In figure 24 we show the mean perceived pull by trials. Seems participants thought they pulled strongest in the second and last trials. Overall, they reported less perceived strength for the first trial.



(a) Mean perceived pull by trial. (b) Mean perceived pull by trial gendered.

Figure 24: Mean perceived pull by trial, and by gender. Error bars show 95% confidence interval.



(a) Mean perceived pull for 1st trial.

(b) Mean ppull for 1st trial gendered.

Figure 25: Mean perceived pull by condition, and by gender for first trial. Error bars show 95% confidence interval.

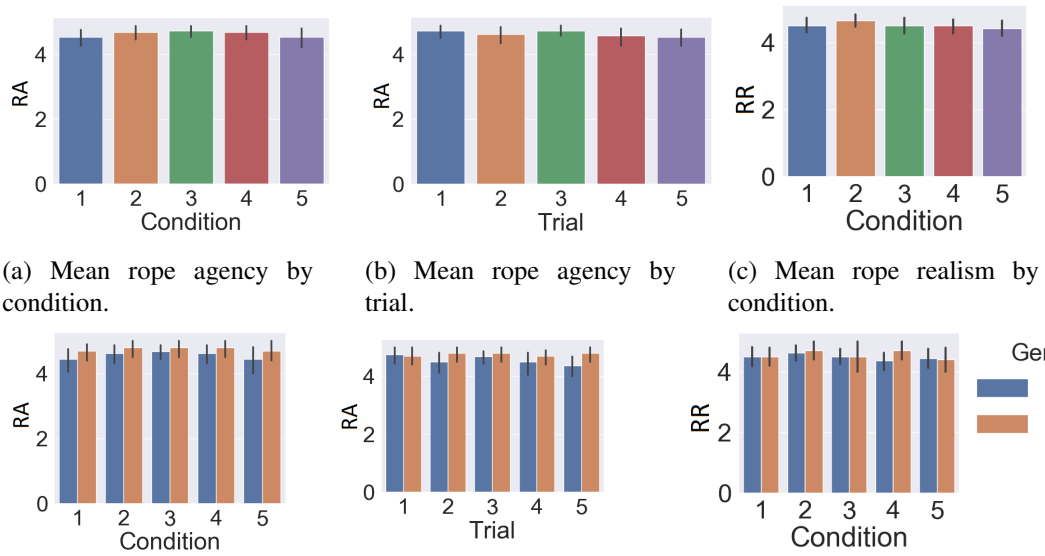
We also look at the perceived pulls only for the first trial in figure 25. Again, they seem to differ from the overall trend, but appear similar to the reported challenges for the first trial. However, there is only one data point for males in condition 1 and 3.

5.0.6 Rope Agency and Realism

Below we briefly present the results of the ratings for rope realism (Q1) and agency (Q2):

- **Q1**: I felt the virtual rope was realistic. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: It looked and felt like I was the one holding the rope. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*).

These questions were mostly meant to create the impression we were evaluating a game. However, we believe these results are noteworthy as they serve to confirm our qualitative feedback that participants were generally a *fan* of the rope. They rated rope agency highly ($M = 4.63846,\ SD = 0.59720$), despite noticing inaccuracies in their hands representation. The same applies to rope realism ($M = 4.51538,\ SD = 0.58713$), in figure 26. In general, participants found the rope to behave as a real rope. We speculate further on body ownership and agency in the next section.



(a) Mean rope agency by condition.

(b) Mean rope agency by trial.

(c) Mean rope realism by condition.

(d) Mean rope agency by condition gendered.

(e) Mean rope agency by trial gendered.

(f) Mean rope realism by condition gendered.

Figure 26: Mean rope agency and rope realism by condition, trial, and by gender. Error bars show 95% confidence interval.

5.0.7 Post-experimental Survey Results

In the following, error bars show standard deviation.

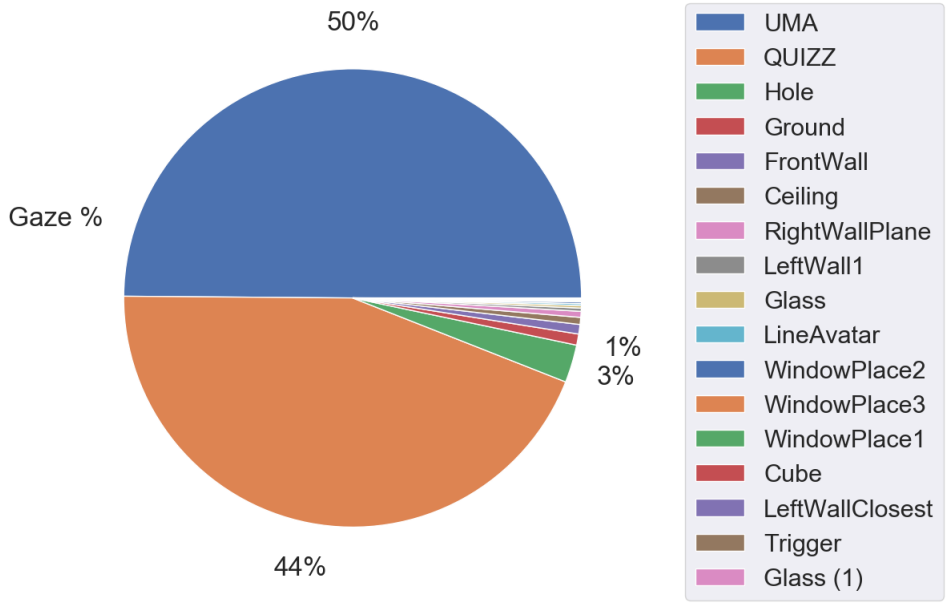
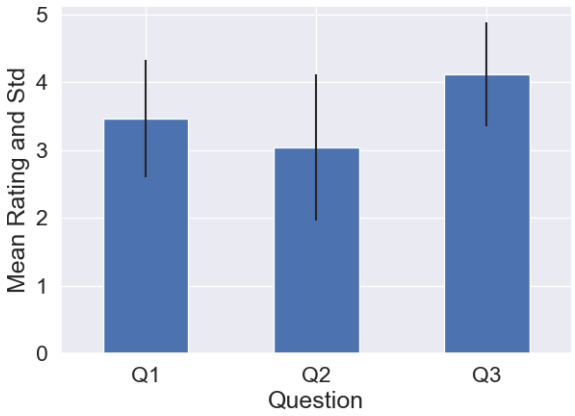
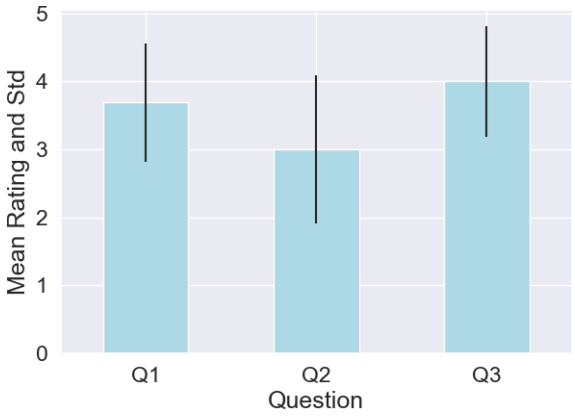


Figure 27: Percent participants gazed at objects during the whole experiment duration.

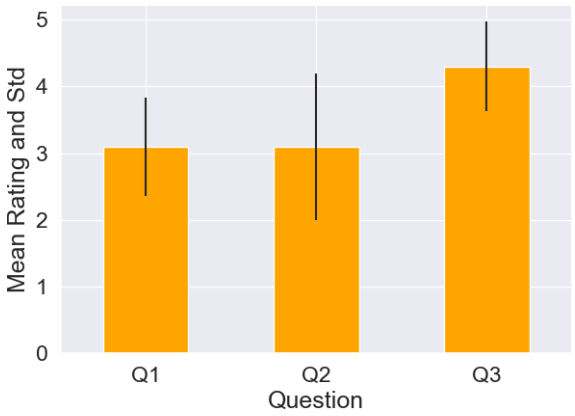
We did not have eye tracking, however we used the midpoint of the headset to estimate what objects participants were looking at. From figure 27, we can see that participants mostly gazed at their opponents. The second most gazed object is the quiz panel (*QUIZZ*), which is expected as participants were answering the questions written on it. The arms were not tagged with a separate ID due to programmatic constraints. However, the object behind the hands was logged. If participants were gazing at their hands, the recorded target of their gaze would most likely be the *Hole* or the *Ground*. Users seem to have paid little attention to the visual representation of their arms. This is a rough estimation, however it is supported by qualitative data. The ratings for rope realism and rope agency were generally very high despite reported inaccuracies in arm tracking.
Below we present results for co-presence, presence and arm ownership measures.



(a) Total mean ratings.



(b) Females mean ratings.



(c) Males mean ratings.

Figure 28: **Co-presence** ratings, by mean and gender. Error bars show standard deviation.

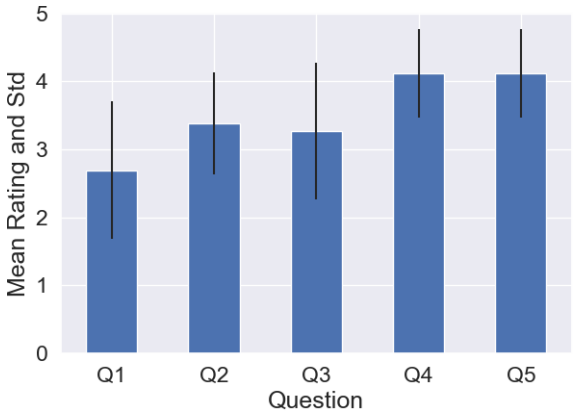
Q	Mean	SD	Median
1	3.4	0.8	4
2	3.0	1.0	3
3	4.1	0.7	4

Table 7: Mean co-presence, standard deviation and median by question.

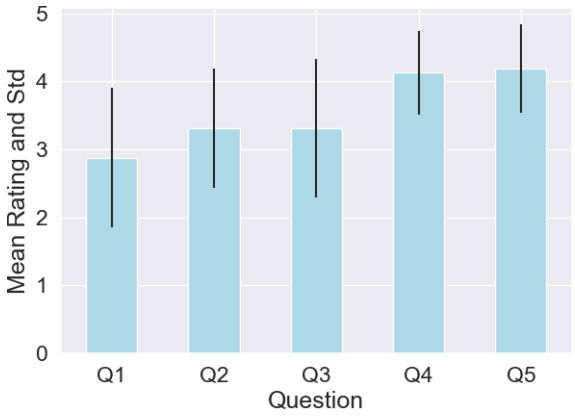
For co-presence, participants answered the following questions:

- **Q1**: My opponents were intensely involved in our interaction. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: To what extent did you feel able to assess your opponents' reactions?. Rating: 1 (*I was unable*), to 5 (*Their reactions were clear*);
- **Q3**: To what extent was this like you were in the same room with your opponents? Rating: 1 (*Did not feel in the same room*), to 5 (*Felt completely in the same room*).

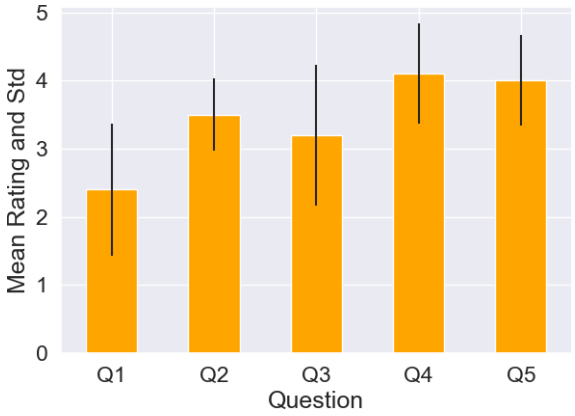
Despite feedback, co-presence mean ($M = 3.53842,\ \ SD = 0.99604$) does not differ significantly from the other categories. Most participants were commenting on the realism of their opponent, so we expected these ratings to be much lower. The lowest rated was question is Q2. This is expected, as several participants mentioned the opponents should be more expressive. Users also made comments about the avatars having the same animation. While they assumed there was only one mechanism driving the force pulls, they still expected the avatars to show differences in movement.



(a) Total mean ratings.



(b) Females mean ratings.



(c) Males mean ratings.

Figure 29: **Presence** ratings, by mean and gender. Error bars show standard deviation.

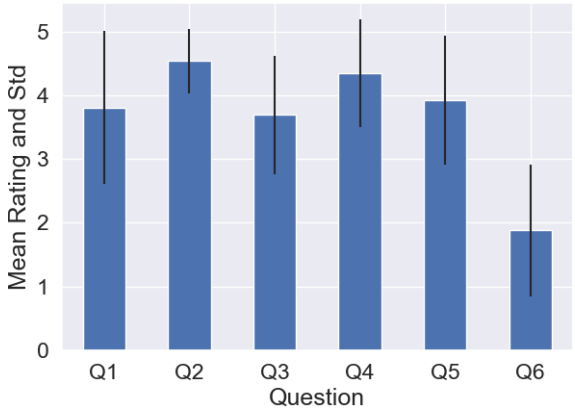
With respect to presence, participants answered the following:

- **Q1**: How aware were you of the real world surrounding while navigating in the virtual world? (i.e. sounds, room temperature, other people, etc.)? Rating: 1 (*Not aware at all*), to 5 (*Extremely aware*);
- **Q2**: How real did the virtual world seem to you? Rating: 1 (*Not real at all*) to 5 (*Completely real*);
- **Q3**: How much did your experience in the virtual environment seem consistent with your real-world experience ? Rating: 1 (*Not consistent*) to 5 (*Very consistent*);
- **Q4**: I felt present in the virtual space. Rating: 1 (*Fully disagree*) to 5 (*Fully agree*);
- **Q5**: In the computer generated world I had a sense of "being there". Rating: 1 (*Not at all*) to 5 (*Very much*).

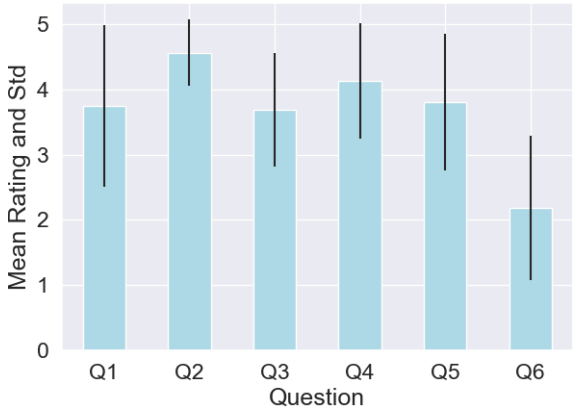
Q	Mean	SD	Median
1	2.6	1.0	3
2	3.3	0.7	3
3	3.2	1.0	3
4	4.1	0.6	4
5	4.1	0.6	4

Table 8: Mean presence, standard deviation and median by question.

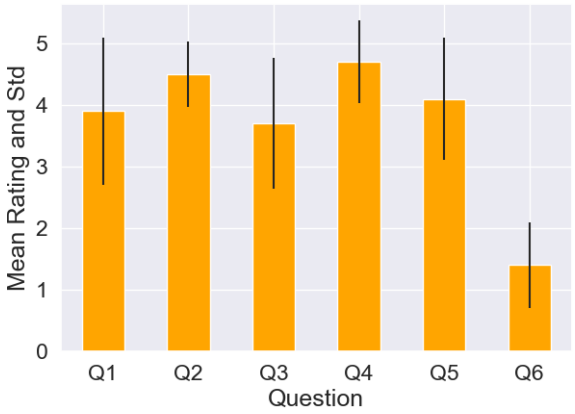
Generally, participants reported above average presence ratings. Males reported less awareness than females for Q1. Due to the design of the experiment, users had to maintain high awareness of their real world environment, as such the rating for Q1 is not as low as desirable. This is a limitation with respect to creating the feeling of presence. Participants were communicating with the experimenter and they were also hearing the countdown which was not provided in the speakers of the headset. Both Q1 and Q3 had higher standard deviation, with Q3 being the second lowest rated. Q3 refers to how consistent participants' experience was with real life. Users commented on the lack of realism of the avatars and how their actions seemed to have no repercussions for the virtual entities.



(a) Total mean ratings.



(b) Females mean ratings.



(c) Males mean ratings.

Figure 30: **Arm ownership** ratings, by mean and gender. Error bars show standard deviation.

Participants answered the following questions about arm ownership:

- **Q1**: I felt as if the virtual arm moved just like I wanted it to, as if it was obeying my will. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q2**: I expected the virtual arm to react in the same way as my own arm. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q3**: I felt that the interaction with the environment was realistic. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q4**: I felt like I controlled the virtual arm. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);

- **Q5**: I felt as if the virtual arm was part of my body. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*);
- **Q6**: I felt as if the virtual arm was someone else's. Rating: 1 (*Fully disagree*), to 5 (*Fully agree*).

Q	Mean	SD	Median
1	3.8	1.2	4
2	4.5	0.5	5
3	3.6	0.9	4
4	4.3	0.8	5
5	3.9	1.0	4
6	1.8	1.0	2

Table 9: Mean arm ownership, standard deviation and median by question.

Ownership ratings are generally high and support participants' feedback. For Q6, we verify if participants felt disembodied from their arms. They rated this feeling low, however females gave a higher rating than males. Among the other ratings there seem to be no significant gender differences. Q1 shows the most deviation. Only a few participants reported seeing inaccuracies with their arms, so we expect the variation to be due to this.

6 Discussion

Our results show that participants pulled differently from one opponent to another, which seems to support H1. Moreover, our qualitative feedback suggests they were expecting opponents to pull according to their perceived appearance. Contrary to their expectations, participants did not pull according to H2. Furthermore, results are inconclusive for H3 and H4. Participants appeared to be influenced by the realism of the experience in their behaviour. The visual feedback they received was inconsistent with the haptic feedback of the rope. That is, participants were unsure how much the rope was being pulled and tested the system. As such, the illusion of a force on the rope eventually appears to break. We discuss implications below with respect to cognitive processes underlying virtual reality illusions and realism.

The ratings for female avatars for the VR tug-of-war game were inconsistent with our results from the survey. There were several limitations when designing the female agents (detailed in section 4.2.1) which lead to females being perceived

overall as less strong and intimidating, even in the survey. One difference between the 2 studies is that participants playing tug-of-war could see the full body of the avatar in VR. On the other hand, participants filling in the survey only saw their thumbnails. Furthermore, due to the resolution of the headset, participants in VR saw the avatars with less clarity and texturing. One participant mentioned that she would have liked to see females having more hair. This aesthetic preference appears to have put off participants instead of making strength cues more salient.
Overall, from their feedback and our observations, participants seemed to use the system in unexpected ways. In one example, P23 acknowledges straying from experimenter instructions to test their assumption about the opponent. After the game, P23 mentions that when they saw a *strong guy* they *expected* them to pull harder, so they pulled less in order to feel this difference. When logging the results of the force meter, and during the game, we also noticed participants were pulling the rope in various ways. They were instructed to pull the rope and keep pulling until they see *Stop*, however many users tugged at the rope, pushed it back and forth instead of keeping it at maximum exertion.
Moreover, other participants seemed to be affected by the novelty of the experience. P10 was surprised that they actually had to pull the rope and realized this only after the trial began, when the experimenter remarked they should pull the rope. The first trial had other design problems. Three participants used their legs initially and pulled too hard, moving the equipment. They realized they could not pull so hard, and restricted themselves in the following trials. Other participants also displayed this restraining behaviour and reluctance to pull hard. Overall, users showed varying degrees of VR illusions. These individual differences have been observed in many VR experiments [20]. Some participants showed a strong inclination to believe the rope is being pulled, while others always maintained that the rope had the same resistance. Gonzalez and Lanier posit that providing a higher cognitive load would allow these illusions to take place more seamlessly. However, our findings suggest that, if cognitive load is too high, users may be too focused on their internal cognitive processes to take in their environment and generate VR illusions.
From qualitative feedback we also noticed that, as participants carry on with the rope-pulling, the illusion that the rope is being pulled eventually breaks. This suggests there might be a trend of lower force across trials. Figure 15 shows some indication of a downward trend for mean force and N_{max}^{min}. Another explanation could be that participants simply felt tired and pulled less. A correlation between performance and realism should be explored in future research. Keeping the number of trials at a minimum to avoid fatigue should be considered a priority.
Nass et al. put forward the social response theory known as *Computers Are Social Actors* (CASA). Their observations have shown that people treat computers as social actors [31]. Essentially, it means that people conform to social rules

throughout their interactions and behave towards a computer as they would towards another human. Furthermore, people seem to mindlessly apply these social rules and other expectations in this context. Some examples are: attributing social categories as gender stereotypes, increased reciprocity in response to assistance, self disclosure reciprocity. We can explain our participants' expectations in terms of the CASA paradigm. People expect strong people to pull harder in real life, so this expectation was projected onto virtual humans and its associated force *mechanism*. It seems clear from our feedback that people assumed there was a machine, and only one, which provided the force and the virtual humans were meant to be representations of this force.

Social comparison theory posits people judge their qualities with respect to the perceived attributes of others. Peña and colleagues frame their results for exergames performance in terms of this theory and note that weight did not solely have a significant effect, but rather the observed differences were a result of an interplay between avatars' and agents appearance' [38]. Some qualitative feedback suggests participants were comparing themselves to the opponents, especially relative to size. Thus, assuming they would pull less with less stronger looking avatars would be in conformity with this theory. We considered that participants could be intimidated by the strong avatars and thus, pull less due to inhibition. From qualitative feedback, avatars did not seem to generate enough realism for participants to react in this manner. Participants remarked being spurred into action by strong appearances, as opposed to being put off.

Comparatively, participants' avatars were bigger than the weaker opponents. This could explain why users tended to pull stronger in the weakest conditions. By the self-perception paradigm, users would see themselves stronger and, therefore, perform stronger. An issue with adopting a Protean perspective for this research is that participants did not see themselves in a virtual mirror before the game, as with usual Proteus Effect studies [3]. Furthermore, they did not appear to look at their arms nearly as much as they looked at their opponents, as shown in figure 27. However, we postulate that the novelty of the experience and ambiguity determined participants to act in a way that is not natural. After all, playing tug-of-war with a virtual person in a virtual environment is not a natural, commonplace experience. Furthermore, participants also mentioned restraining themselves in various ways, by keeping their legs still or being careful not to pull too hard. These actions might determine high cognitive load preventing a natural social response. Despite this, the general consensus among researchers seems to be that VR can generate realistic reactions and is ideal to study psychological and social behavior without endangering participants [16].

Another reason why we expected participants to pull harder was due to the priming effect of avatars' appearances [37]. Aggressive cues were meant to activate peoples' aggressive behavior and enable them to pull stronger. Considering the

interplay between these psychological effects, we expected a similar outcome for our experiment, that people would pull stronger for strong avatars. Our results, however, are inconclusive. Fundamentally, they capture the variability of human behaviour .

The large force variations make our measurements unreliable to support hypothesis H2. This is especially true for some participants with a very big difference in force. If users have different motivations for pulling stronger than simply reacting to their opponent, we cannot make a reliable case for our research aims. Moreover, testing the application suggests that participants did not follow the instructions provided by the experimenter. We do not know what effect this could have on their perceived pull value, but we speculate participants could have overcompensated in cases in which they pulled much less. We believe there is a reasonable doubt to be placed over the measures for H3 and H4. The trends they show, however, lead us to believe that H3 and H4 should not be dismissed. So far, perceived pulls seem to conform the most with our hypothesis.

We cannot and should not prevent participants from using our systems in playful and unexpected ways. This behaviour is documented for other new technologies such as conversational agents [24]. For a question-answering system, Liao et al. document similar playful and curiosity-driven behaviour. Approximately 85% of users asked the system questions outside the scope of its advertised purpose [2]. For games or VR experiences meant to give rise to VR illusions, we suggest extensive piloting in order to establish users' expectations. This would allow for more realistic interactions with the system and reduce the number of breaks in presence or plausibility.

Our observations that the illusion seemed to break as the game went on would lead us to believe participants behaved more realistically in the first few trials. For this reason we look at results only for the first trials in figures 22,25,16. Overall, measurements do not seem to show a consistent trend. However, isolating results only for the first trial seems to paint a more consistent picture of our data.

As values for perceived pull and challenge appear to increase for the last conditions, it seems less likely participants would be intimidated by the opponents into pulling less. We also suggest, based on our feedback, that the behaviour of the agents was not realistic enough to determine such a response. Participants noticed their opponents did not react in sync with their pulling.

To produce VR illusions, it seems essential for users to be at the center of events and observe synchronized visual and motor cues [47]. We designed an experience meant to provide visual and haptic synchronization. That is, when users pull the rope, they feel resistance and see the avatars tug at the rope. However, we did not account for several *sensorimotor contingencies*. We adopt this terminology to express the actions of our participants:

- They pulled the rope and expected the avatars to look like they are being pulled back, a result of their motor actions;
- They let go of the rope and expected to still feel resistance, in congruence with the visual feedback.

Most comments relating to the challenge ratings were centered around the fact that opponents were not reacting to users. That is, their sense of agency was being called into question. We believe that failure to provide these affordances lead to the breakdown of our illusion. Our experiment suggests that sustaining realism strongly depends on credibility and conformity to expectations [44]. Indeed, breaching a fundamental assumption about the affordance of an object could place doubt on the entire context. If opponents do not pull back, then users may question the competitive nature of the game itself. This would place doubt on the validity of all our qualitative measures.

In conformity with the CASA paradigm [31], participants were expecting their opponents to be socially different. Users were not expecting the avatars to behave the same since they had different appearances. We assumed having differently sized opponents would determine observers to perceive their movements with a different magnitude. For this reason we used the same animation for all opponents. However, some participants noticed the movements were the same. We speculate the resolution of the headset also affected how opponents were perceived. The avatars had various facial expressions when they pulled, though it seemed a few participants did not notice this.

Participants suggested various ways of increasing behaviour realism in their opponents, from giving them voices to enhancing their facial expressions and varying their movements. These changes should be considered carefully and we suggest extensive piloting when high behaviour realism is desired. Adding more cues also increases the likelihood of errors and inconsistent feedback. Increasing behaviour realism in virtual humans could give rise to uncanniness and provoke aversion [12, 51]. Studies examining this phenomenon suggest that inconsistent visual or auditory feedback can push human-like behaviour in the uncanny valley [43]. The illusion of human appearance seems to be very similar with illusions bred within virtual reality. Synchronization and congruence are key. Furthermore, allowing for these behaviour variations would make empirical correlations of independent variables difficult. The effects of movement and appearance would be hard to decouple.

Gonzalez and Lanier further detail the cognitive perceptual mechanism underlying VR illusions [20]. They mention that the minimal requirement to produce such illusions without breaks in presence is to combine continuous visual feedback with sensorimotor contingencies in a synchronous manner. It seems congruent information is more likely to be accepted by the brain as being *real* than ambiguous

information, and VR motion sickness is an example of this. The observations gathered from our experiment seem support this congruent feedback integration framework. We observed that these illusions can occur in an ambiguous setting as well. We echo the authors' call to more research investigating VR illusions. The processes that enable them could deepen our understanding of the human mind and allow designers to make ethical considerations with respect to their products. Many participants mentioned that holding the rope helped with realism. We believe the illusion of holding the rope was strongly supported by the haptic feedback. Participants did not generally look at their hands, and we speculate this was because of the feedback. It is surprising participants did not seem to rely on their visual senses as much, when holding the rope. Dominance of the visual sense is a common human property [40], however we believe that haptic feedback together with a social distractors kept people's gaze away from their body. Gonzalez and Lanier posit that sensory saturation will mitigate awareness of illusions and we believe haptic feedback contributed to higher body ownership in our case [20]. Participants appeared driven by curiosity and tugged the rope and moved it in unusual ways. We believe physical objects can contribute to higher presence and engagement in VR, if visual and tactile feedback are integrated in a congruent manner. Researchers have already begun exploring the advantages of multi-sensory feedback for physical objects for VR ([50]). We consider they have great potential in serving as distractors and mitigating visual inconsistencies.

7 Limitations

While our study showed great potential in supporting H1, we had many technical and functional limitations.

First of all the spring we used to provide resistance absorbed some amount of force and the data we gathered is not the actual pull. Changes in the physical structure of the spring could also have altered our measurements between participants. For example. one of the participants stretched the spring so much that it had to be replaced. However, we believe that, if the spring had not been integrated, some participants would have pulled over the maximum measurable force of the meter. Ultimately, this is a constraint of the measuring instrument.

Furthermore, the force data we gathered does not reflect the whole story of how participants pulled. We measured the maximum force, however other measures such as the longest value displayed on the monitor or the average value during a trial might reflect participants' performance better. Other suggested measuring techniques would be to take intervals at the end of the countdown and at the start of the countdown to see how fast participants got to the maximum value of that trial. Unfortunately, as participants did not continuously pull the rope, our finding

suggest fixed interval measures would be unreliable.
Another limitation is that he rope was not tied to a fixed object and, as such, participants were asked to keep their legs still when pulling. Restraining participants in this manner may have induced higher cognitive load and prevented them from acting realistically. While we had extensive piloting to verify the setup, a few participants still showed great force and pulled the force meter too much. These situations affected their subsequent pulls as they noticed they had to restrain themselves.

8 Conclusion and Future Work

We hypothesized that seeing stronger opponents would determine people to pull harder. As P13 mentioned, it seemed like a *a natural response*. Perhaps more surprising than supporting this hypothesis, is that our experiment showed inconclusive results and great individual variation. Our main findings suggest that most participants perceived differences in the force acting on the rope. This is reflected by the fact that they reported different challenges across trials, giving support to H1. Furthermore, we also observed breaks in the illusion, when participants' expectations were not met and, instead, the system provided incongruent motor and visual feedback.
The virtual human opponents seemed to be the focal point of the interaction. Participants had expectations that were not met about the interaction with the avatar, which seemed to place doubts about the realism of the whole system in their minds. For example, they expected the opponent to react to their feedback. Opponents were visually shown to push and pull with the beginning and end of the countdown. We did not provide contingencies, however, if participants strayed from the standard flow of the interaction. When they did, they found their expectations were not met. In order to sustain realism, we believe it is useful to provide as few boundaries for the interaction as possible. This means that designers should provide as many contingencies as they can in their virtual experiences. Assuming that users will follow a standard path does not seem feasible.
When users' expectations are not met, breaks in realism and ownership appear to occur. This eventually appears to decay the illusions generated in VR. These illusions that the virtual world behaves like the real world are meant to elicit realistic behaviour from users. Additionally, we believe sustaining these illusions is also required in order to maintain high engagement and keep users motivated to use the system. High behaviour and physical realism might not be necessary, however it seems to provide limits in which these kinds of illusions can occur [44]. While we provide some findings with respect to decays of illusion, more research is necessary to examine the underlying cognitive mechanisms and their effect on

the overall user experience. We suggest that this process can be examined with controlled experiments where breaks are induced by the experimenter.

Embodiment is one of the many illusions generated by the virtual environment. In the present work, we do not manipulate the appearance of the users' avatar. However, we acknowledge the significant effect of embodiment on users' perceptions and postulate that a physical transformation of the self, coupled with an appropriate context, would generate a better illusion. Providing a contrast between the opponent and the agent could highlight various stimuli, such as making the opponent's strength more obvious. However, the interplay between the observer's appearance and the virtual representation of others is not fully understood. Decoupling the two remains a challenge for future researchers.

Researchers have shown that body ownership occurs with dramatically different appearances, with larger or even multiple body parts, dubbed as *Homuncular flexibility* [33, 55]. Evidently, virtual reality has demonstrated its ability to allow human perception to transcend its normal bounds. However, we believe more research is required to explore how far VR can go in altering lived experiences. While most research has focused on internal perceptions, we have shown evidence that virtuality is capable of altering people's beliefs about the real world. As VR technology becomes more advanced, designers need to know the extent of the kind of experience they want to create.

Additionally, we observed that physical objects which provide continuous haptic feedback might serve as distractors and mitigate visual inaccuracies. However, we believe that affording realistic interaction with these objects is key to sustaining presence and engagement. Our findings suggest that haptic feedback could be used to induce stronger ownership illusions. This is especially desirable when visual cues are lacking. More research is required to support our observations.

Virtual reality's immersive capabilities have been successfully applied in fields such as entertainment, military training and phobia management in order to elicit real-world human experiences. Bowman and McMahan examine some of these successful VR ventures and suggest their aspired fidelity to real life is what sets them apart [11]. In the long term, the goal of many VR researchers and enthusiasts seems to be engaging as many of the human senses as possible to replicate real-life sensory experience with high accuracy. Ultimately, our findings inform the design of realistic VR applications. As mixed reality becomes more and more popular, the seamless integration of haptic feedback appears to be a necessity. Therefore, generating physical perceptual differences which comply with visual feedback.

9 Appendix

9.1 Additional Quantitative Data

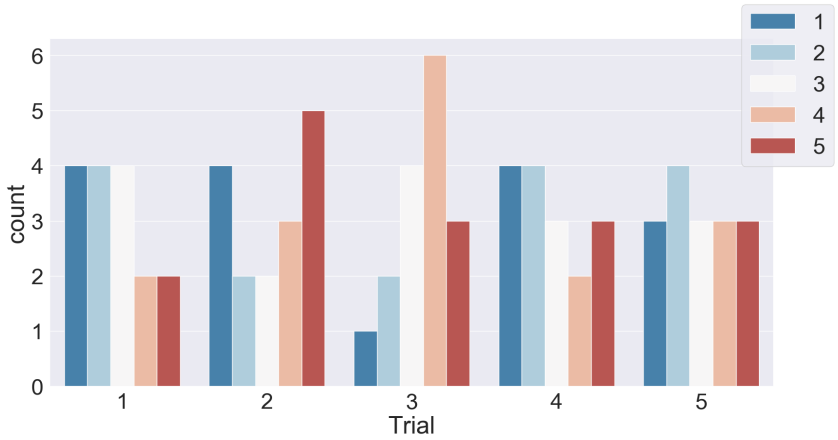


Figure 31: Number of conditions per trial for females.

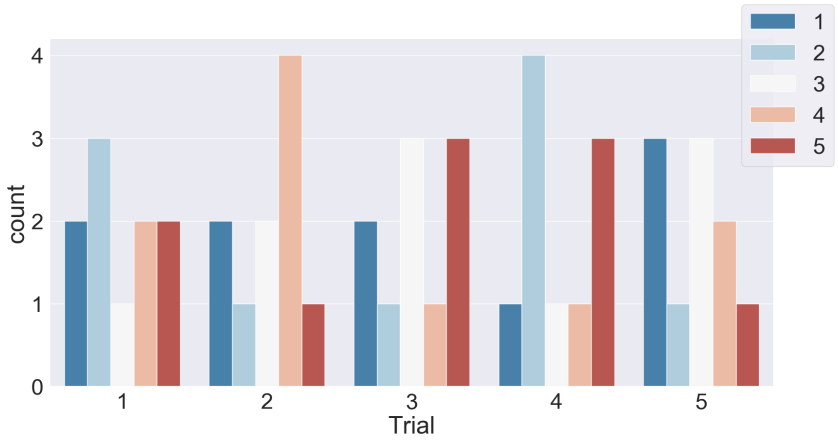


Figure 32: Number of conditions per trial for males.

9.2 Instruction Sheet

Instruction sheet for experimenter for VR games study

1. Describe the games to users at arrival
We made 2 virtual reality games - a virtual reality tug-of-war game and a whack-a-mole game! We are interested in providing users with the best gaming VR

experience and we need your feedback for the games we designed.

Tug of war, if you have not played it, is a game in which you pull a rope with an opponent and the one that pulls stronger wins! For whack a mole you see a mole come up and you have to slap it with your right hand and hold your hand straight like this [show participant].

First, for each game we must calibrate these gloves for your hands and then we can start playing. There will be a screen in front of you and you will have to follow some hand movements. I will walk you through the instructions.

We first play tug of war and then 4 minutes of whack a mole! After each game, I will kindly ask you to fill in a short survey about your experience and other design aspects. Before you leave, I will ask you for any feedback or suggested improvements you might have about each game. For full disclosure, our conversations during the experiment will be recorded. All data I record will be transcribed and anonymized. Only I have access to the code.

But before you I continue the instructions could you please have a look at these two consent forms. There is a place for your name and signature at the end. I will be here and you can ask me any questions while you read.

2. Tug-of-War Instructions
For tug of war, your task is to play the game and compete at pulling the rope! You will pull the rope 5 times and you can use both hands for that. You must put your right hand in front of the left hand on the rope and hold it like this [show]. I will give you the rope in your hands once we finish calibrating the gloves. Before I place the gloves on your hands could yoou please fill in this page from the survey with your information [Show participant survey on laptop page and tell them to click next after inputting their data. The next page will not be filled in. Tell them you will shortly explain what that is. After they fill in the questions start putting the gloves on their hands and continue explaining.]

In the game you will see your opponent and a countdown will begin. When you see START you should start pulling the rope and keep pulling until you see stop, When you see stop you should stop pulling the rope and the round will be over.

After each rope-pull, there will be a set of questions in VR on a panel and I will kindly ask you to read the questions out loud and tell me your answer. I will record your answers in this survey on the laptop, there [show].

You have to remember 2 things:

1. Try pulling the rope without moving your feet.
2. Keep your hands on the rope at all times, even when you are reading questions between rope-pulls.

3. Setup and runtime instructions

At all times:

1. Keep trackers of gloves charged
2. Keep batteries in charging port
3. Gloves demagnetized daily/several times per day
4. Always let participants know what you want to do. (eg. I will turn off the game now).
5. Always let participants know if you want to move them or adjust any parts of the VR setup such as the headset or gloves, especially if it involves touching them.

Before participant arrival:

1. Wipe headset
2. Close windows in experiment room
3. Check all devices are tracked
4. Put laptop speakers on the correct output
5. Start survey
6. Adds participant ID and click next
7. Replace batteries on gloves if needed
8. Demagnetize gloves if needed
9. Turn on trackers of gloves
10. Turn on gloves
11. Make sure trackers are synced with steam (they have a green light)
12. Turn on both unity projects
13. Start force meter
14. Turn on camera
15. Turn on OBS

16. Make sure Windows picks the correct webcam, the one over the force meter
17. Start Unity and open both games
18. Place Camera and Unity Game view side by side to capture them with OBS.

At start experiment

1. Greet participant.
2. Explain intro story
3. Give participant both consent forms to sign
4. Participant is invited to complete data with gender and age
5. Do you have any questions? You can ask me anything throughout the experiment.
6. Give participants gloves
7. Give participant headset protector
8. Give participants headset
9. Do glove calibration
10. Tell participants you will give them the rope in their hands now
11. Give participant rope in hands
12. Tell them to grab the rope with right hand in front of left
13. Tell participant First I will show you the VR setup. I would like you to look around for a minute, tell me if everything is clear, if you can see your hands, the rope, and if the words on the panel are clear.
14. Start PreTrial scene.
15. Stop scene.
16. START FORCE METER (if it turned off)
17. START RECORDING OBS
18. START RECORDING CAMERA
19. START GAME in Experiment scene
20. Tell participant: *Now I will start the game and you will be facing your opponents. Remember to start pulling when you see START and stop pulling when you see stop. Don't forget, always keep your hands on the rope.*
21. Remind participants of constraints at trial 3
22. At 4th rope pull, tell participant there are 2 more rounds.

23. For question-answering, tell participant whenever you are ready to take their answers and after each answer tell them some word of acknowledgement that you got their answer (eg. ok, Alright). Refrain from using overly positive adjectives like great, to prevent encouraging them to give positive answers.

4. End instructions
Note that the experiment instructions for the whack-a-mole game are nor presented as they are out of the scope for this research. Participants always played tug-of-war first, completed its related survey, then played whack-a-mole and completed its related survey.

Take a pen, paper and write participant feedback during the interview.

1. Thank the participant.
2. I would like to ask you now if you have any suggestions to improve the first game, the tug-of-war game? What was your impression?
3. What about the whack-a-mole game? Do you have any thoughts about that game?
4. Thank you for the feedback. One last question before you go. Could explain to me, in your own words what this experiment was about? Based on what you saw and what I explained.

At the end of each day do the logging:

1. Log participant feedback in document on Google Drive.
2. Move logs in experiment folder for both games and commit to Git.
3. Rename recorded videos with participant id and upload them to Google Drive.

9.3 Consent form

Informed Consent Form for Volunteer Participants This informed consent form is for volunteer participants to take part in an experiment about playing tug of war in virtual reality with different players. This research is done by a student following a master's degree in Computer Science at the University of Copenhagen, Denmark as part of their Master Thesis. This research is under the supervision of professors Henning Pohl and Kasper Hornbæk from the University of Copenhagen.

This Informed Consent Form has two parts:

- Information Sheet (to share information about the research with you, the volunteer participant)
- Certificate of Consent (for giving your consent if you agree to take part)

PART I: Information Sheet Introduction I am Andreea-Anamaria Muresan, a student at the University of Copenhagen, and I am doing a master thesis about interactions in virtual reality.

I would like to invite you to be part of this research. In this document, you can find detailed information about the project. Before you decide, you can talk to anyone you feel comfortable with about the research. If there are words that you do not understand please ask me about them and I will take time to explain. If you have questions later, you can ask me or even pose your query to the project supervisors.

Purpose of the research and type of manipulation The purpose of this research is to evaluate a rope-pulling game in virtual reality. In this evaluation, we look at the quality of the virtual rope, the design of the players and measure body ownership and presence in the virtual world.

Participant selection You have been invited to this study because you satisfy the following requirements: you are representative of a group of users who may use virtual reality.

Voluntary Participation
Your participation in this research is entirely voluntary. It is your choice whether to participate or not. You can change your mind later and stop participating even if you agreed earlier.

Procedures and Protocol
Throughout this experiment you will be asked to:

1. Play tug-of-war,
2. Complete a verbal questionnaire after each rope-pull,
3. Fill in a survey after finishing the game,
4. Have a short chat before you leave and tell us your opinion about the game and how we can improve it. Our conversations will be recorded and anonymized.

At the beginning of each round, you will hear and see a countdown. You must start pulling when you see **Start** and you must stop pulling once you see **Stop**.
Please try to:

1. Grab the rope with your right hand in front of the left hand.
2. Keep your hands on the rope at all times.
3. Do not move your hands on the rope once you grab it.
4. Try pulling the rope with your upper body and arms. Try keeping your feet in the same starting position.

Duration
The experiment is expected to last between 20 and 30 minutes.

Risks and Benefits
We anticipate no risks from participating in the experiment. If you participate in the experiment you will help the student conducting the experiment complete their Master Thesis successfully as well as gain experience running, designing experiments and interacting with participants.

Reimbursements You will not be given any money or gifts to take part in this research.

Confidentiality
The data we collect from you will be made anonymous; your name will be replaced with a number known only to the person running and designing the experiment.

Sharing the Results
If you are interested, you may leave us your email or contact us at a later stage to learn about the outcomes of the study. No confidential information will be shared.

Right to Refuse or Withdraw
You do not have to take part in this research if you do not wish to do so. You may also stop participating in the research at any time you choose. It is your choice and all your rights will be respected.

Who to Contact
If you have any questions about the study at a later stage, please contact the student investigator at the following email addresses: Andreea-Anamaria Muresan: zph748@alumni.ku.dk
You can ask me any questions about any part of the research study if you wish. Do you have any questions so far?

PART II: Certificate of Consent

I have read the foregoing information. I have had the opportunity to ask questions about it and any questions that I asked have been answered to my satisfaction. Below I mark that I consent voluntarily to participate as a participant in this research.
Print Name of Participant:
Signature of Participant:
Date (Day/Month/Year):

Statement by the researcher/person taking consent
I confirm that the participant was given an opportunity to ask questions about the study, and all the questions asked by the participant have been answered correctly and to the best of my ability. I confirm that the individual has not been coerced into giving consent, and the consent has been given freely and voluntarily.
Print Name of person taking the consent:
Signature of person taking the consent:
Date (Day/Month/Year):

9.4 Coding Tables

ID	Name	Condition
1	F5_C3	Weak
2	F4_C1	Strong
3	F1_C2	Average
4	F6_C1	Strong
5	F1_C1	Strong
6	F3_C1	Strong
7	F5_C1	Strong
8	F5_C2	Average
9	F1_C3	Weak
10	F4_C3	Weak
11	F2_C3	Weak
12	F6_C2	Average
13	F6_C3	Weak
14	F4_C2	Average
15	F3_C3	Weak
16	F2_C2	Average
17	F2_C1	Strong
18	F3_C2	Average

Table 10: Female avatar designs coding.

ID	Name	Condition
1	M6_C3	Weak
2	M5_C1	Strong
3	M1_C2	Average
4	M7_C1	Strong
5	M1_C1	Strong
6	M3_C1	Strong
7	M6_C1	Strong
8	M6_C2	Average
9	M1_C3	Weak
10	M5_C3	Weak
11	M2_C3	Weak
12	M7_C2	Average
13	M7_C3	Weak
14	M5_C2	Average
15	M3_C3	Weak
16	M2_C2	Average
17	M2_C1	Strong
18	M3_C2	Average

Table 11: Male avatar designs coding.

9.5 Survey Mean Ratings

ID	Attractive	Intelligent	Intimidating	Strong	Weighted	UMA
10	1.93333	2.66667	1.66667	1.46667	1.56667	f4 c3
15	1.73334	2.46667	1.73334	1.53334	1.63334	
14	1.93334	2.8	2	1.4666	1.73333	
8	2.46667	2.8	1.8	1.86667	1.83333	
11	2.33333	3	1.6	2.06667	1.83333	
16	2.86667	3	1.73333	2.06667	1.9	f2 c2
9	1.8	2.6	2.2	1.66667	1.93333	
13	2.06667	2.66667	1.93333	1.93333	1.93333	
1	2	2.33333	1.86667	2.06667	1.96667	
18	2.26667	2.6	2.2	2	2.1	f3 c2
12	2.4	2.66667	2.06667	2.26667	2.16667	
7	1.8	2.13333	2.13333	2.33333	2.23333	
3	2	2.8	2.66667	2.66667	2.66667	
5	1.53333	2.2	2.73333	3	2.86667	f1 c1
17	1.6	2.2	2.86667	3.06667	2.96667	
4	1.73333	2.33333	3	3.26667	3.13333	
6	1.66667	2.4	2.73333	3.53333	3.13333	
2	1.46667	2.2	3.4	3.46667	3.43333	f4 c1

Table 12: Female mean ratings.

ID	Attractive	Intelligent	Intimidating	Strong	Weighted	UMA
13	2.625	3.0625	1.25	1.25	1.25	m7 c3
8	2.5	2.6875	1.5625	1.875	1.71875	
1	2.1875	2.4375	2.125	1.375	1.75	
9	1.25	1.625	2.125	1.75	1.9375	
12	3.5	3.3125	1.5625	2.375	1.96875	m7 c2
11	1.8125	2.3125	2	2.25	2.125	
15	1.8125	2.1875	2.125	2.5625	2.34375	
10	2	2.375	2.3125	2.4375	2.375	
7	1.75	2.8125	2.3125	2.75	2.53125	m6 c1
4	2.375	2.5625	2.3125	3.3125	2.8125	
16	2.0625	2.25	2.375	3.3125	2.84375	
18	2.625	2.625	2.5625	3.5625	3.0625	
14	2.3125	2.5	2.75	3.5625	3.15625	
3	1.375	1.9375	3.5	3.0625	3.28125	m1 c2
6	2.5625	2.5625	3.0625	4.1875	3.625	
5	1.5625	2.3125	3.5625	4.0625	3.8125	
17	2.25	2.125	3.625	4.6875	4.15625	
2	1.9375	2	3.8125	4.75	4.28125	m5 c1

Table 13: Male mean ratings.

9.6 All Challenge and PPull Ratings

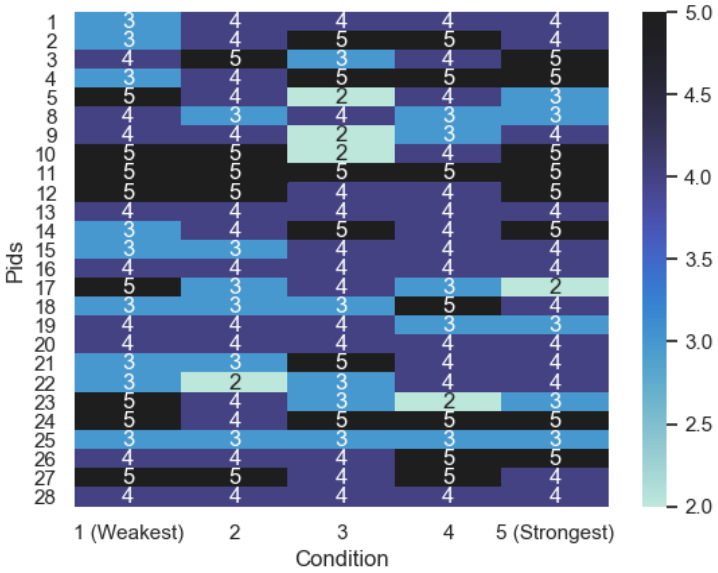


Figure 33: Perceived pull ratings by condition per user coded as user Id.

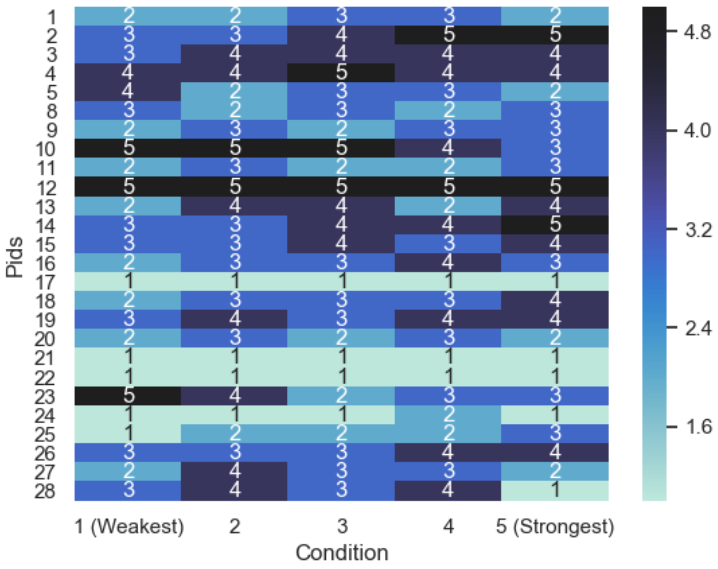


Figure 34: Challenge ratings by condition per user coded as user Id.

9.7 VR Avatars

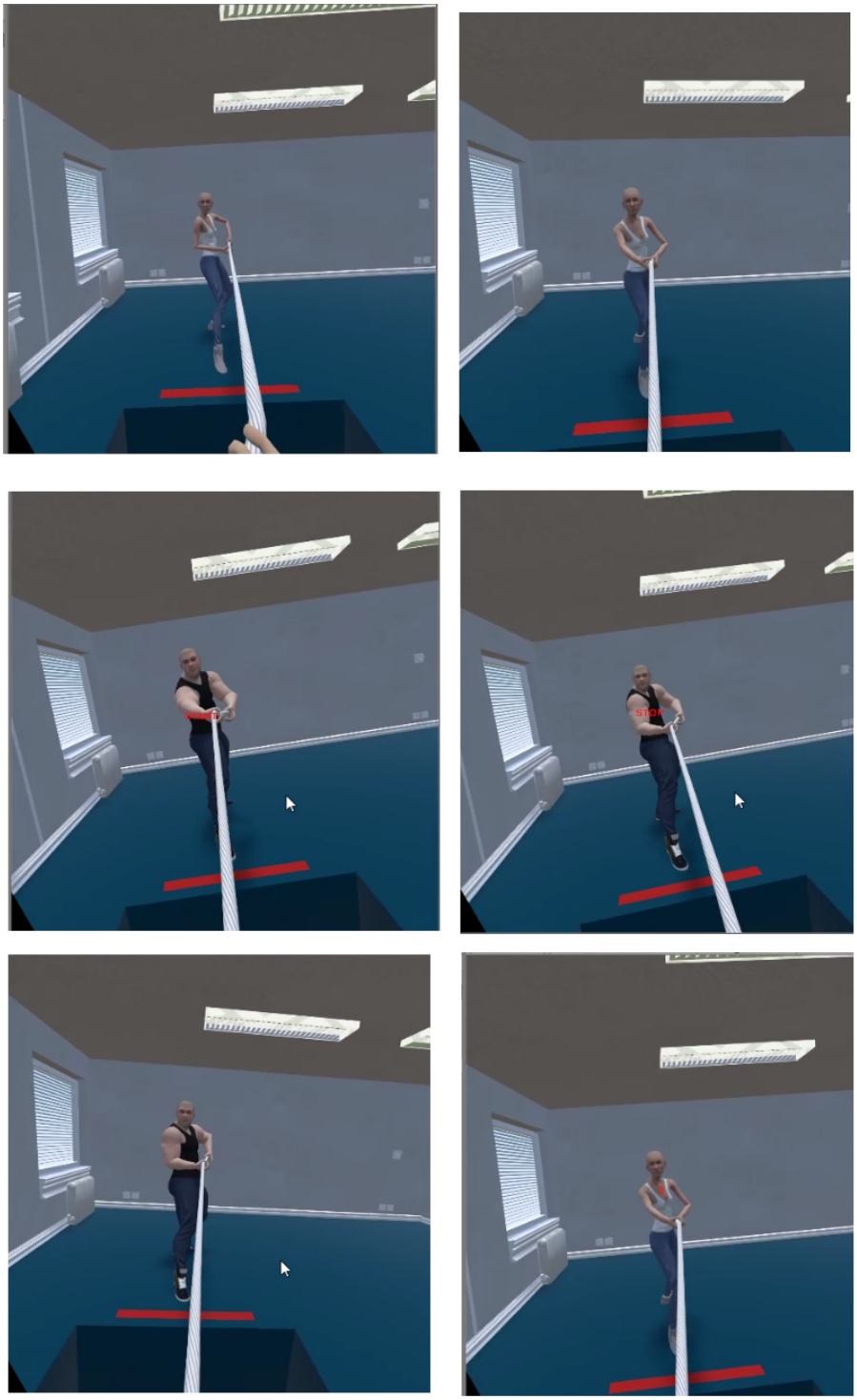


Figure 35: Screenshots of participant's view of the in-game VR avatars in different states of the game.

9.8 User Study Thumbnails and Ratings

9.8.1 Females

Figure 36: Female in condition 1 (weakest) thumbnail and ratings.

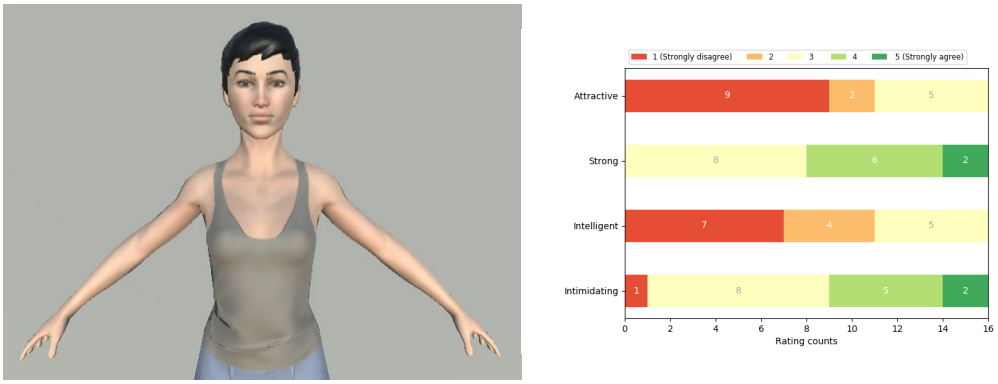


Figure 37: Female in condition 2 (low-average) thumbnail and ratings.

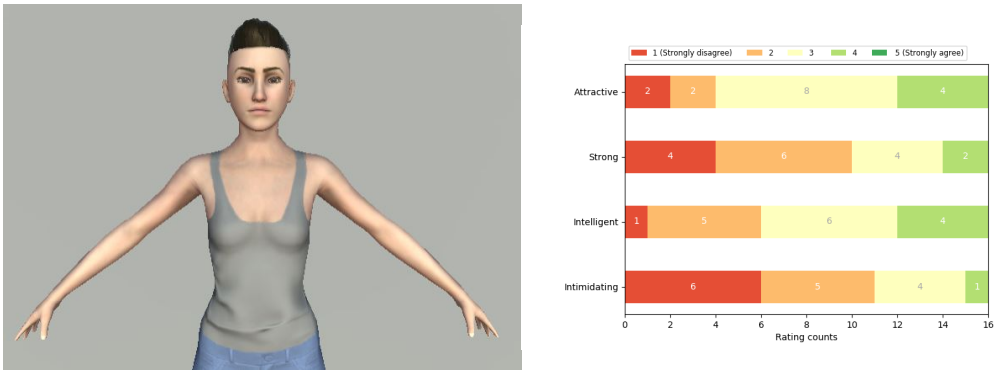


Figure 38: Female in condition 3 (average) thumbnail and ratings.

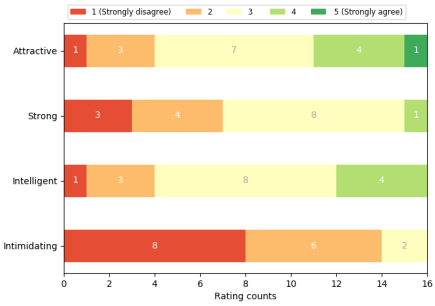


Figure 39: Female in condition 4 (high-average) thumbnail and ratings.

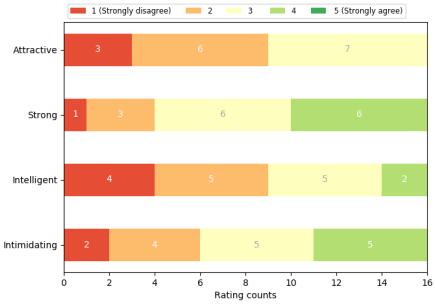


Figure 40: Female in condition 5 (strong) thumbnail and ratings.

9.8.2 Males

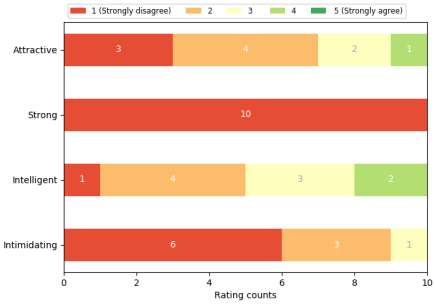


Figure 41: Male in condition 1 (weakest) thumbnail and ratings.

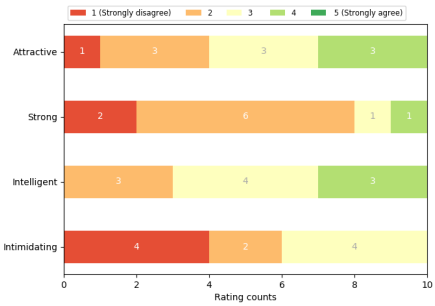


Figure 42: Male in condition 2 (low-average) thumbnail and ratings.

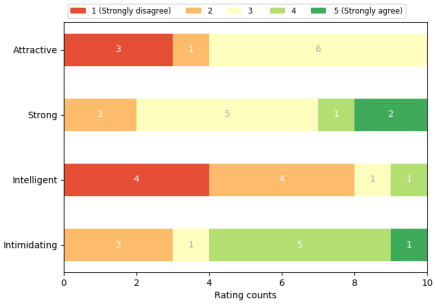


Figure 43: Male in condition 4 (average) thumbnail and ratings.

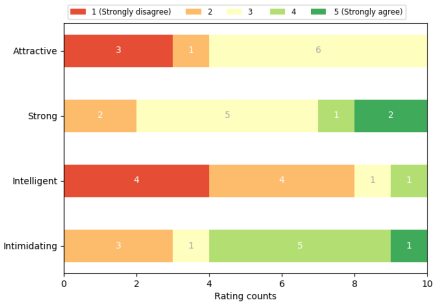


Figure 44: Male in condition 4 (high-average) thumbnail and ratings.

Figure 45: Male in condition 5 (strong) thumbnail and ratings.

9.9 Survey Thumbnails and Ratings

9.9.1 Females

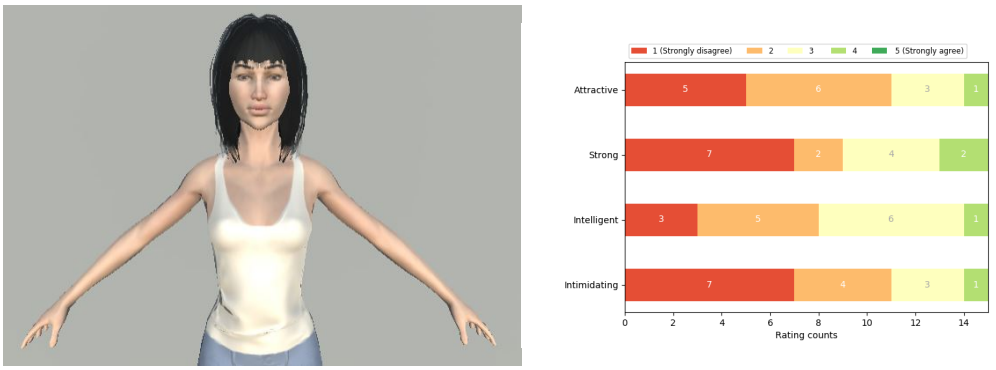


Figure 46: Female 1 thumbnail and ratings.

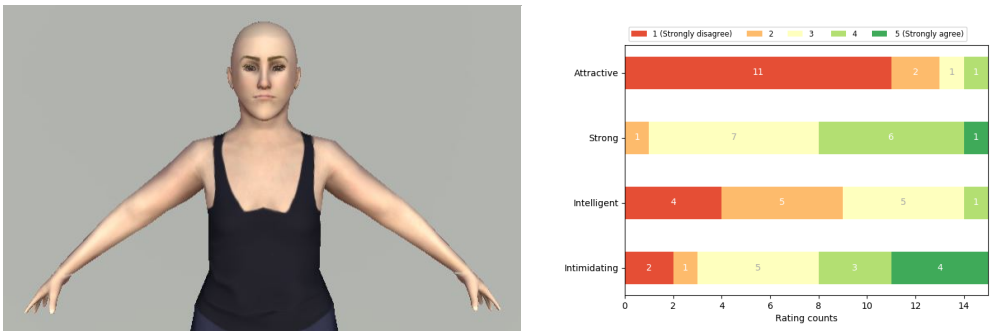


Figure 47: Female 2 thumbnail and ratings.

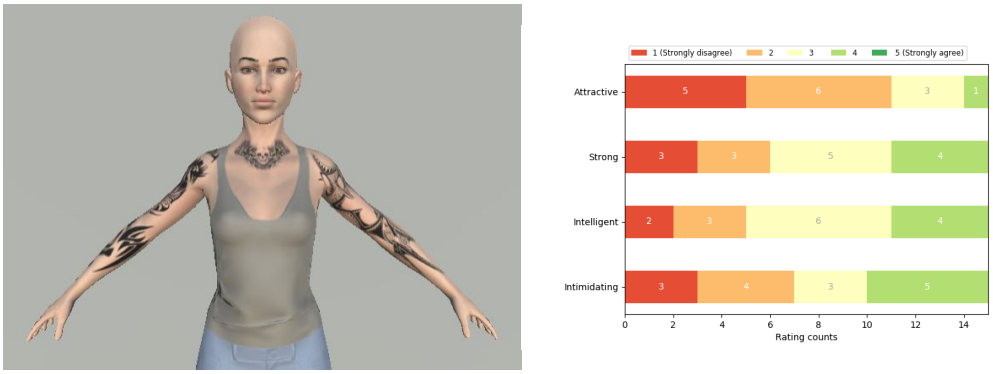


Figure 48: Female 3 thumbnail and ratings.

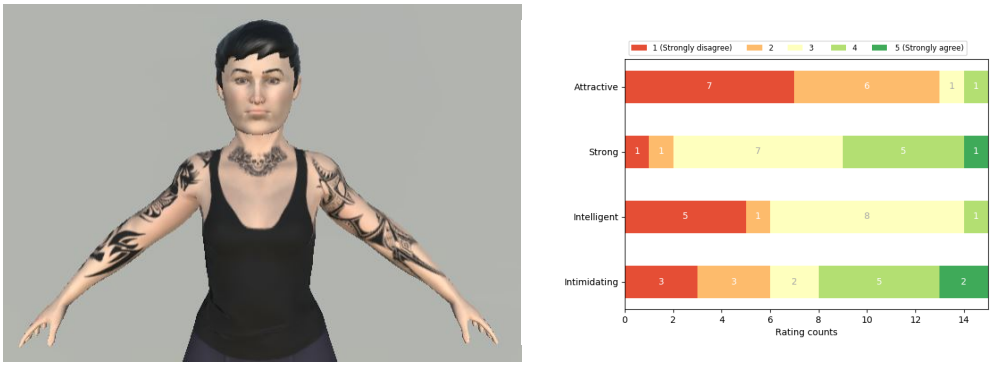


Figure 49: Female 4 thumbnail and ratings.

Figure 50: Female 5 thumbnail and ratings.

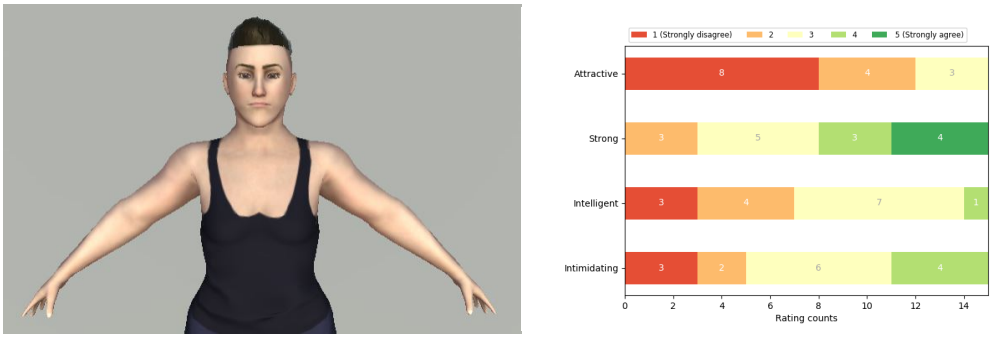


Figure 51: Female 6 thumbnail and ratings.

Figure 52: Female 7 thumbnail and ratings.

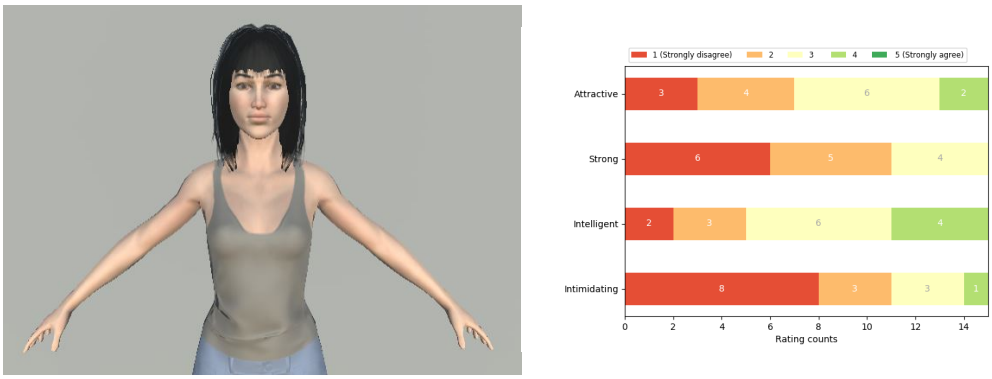


Figure 53: Female 8 thumbnail and ratings.

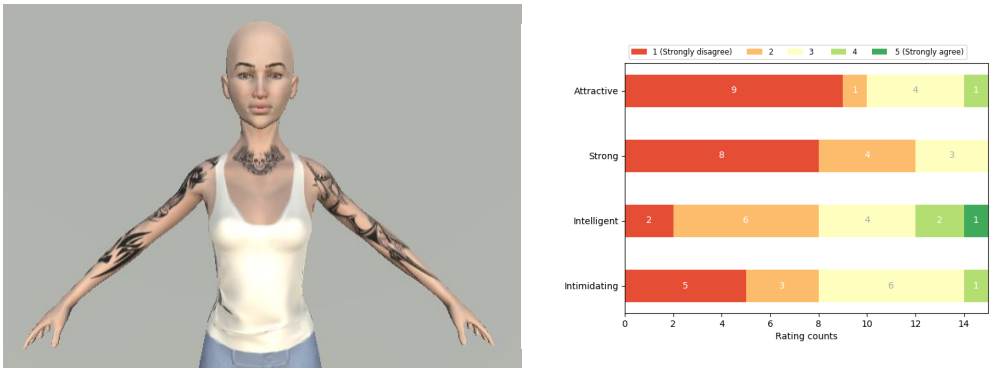


Figure 54: Female 9 thumbnail and ratings.

Figure 55: Female 10 thumbnail and ratings.

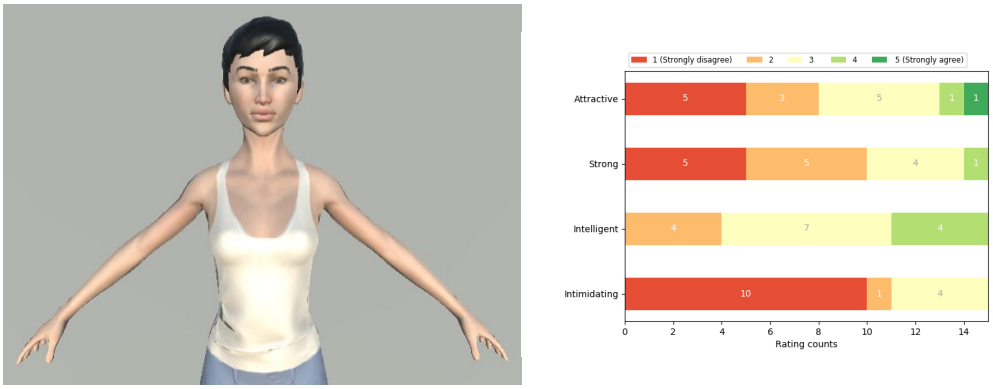


Figure 56: Female 11 thumbnail and ratings.

Figure 57: Female 12 thumbnail and ratings.

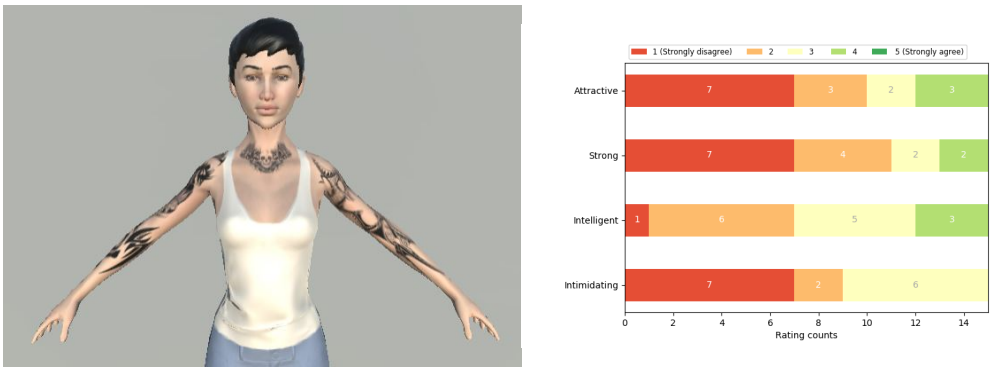


Figure 58: Female 13 thumbnail and ratings.

Figure 59: Female 14 thumbnail and ratings.

Figure 60: Female 15 thumbnail and ratings.

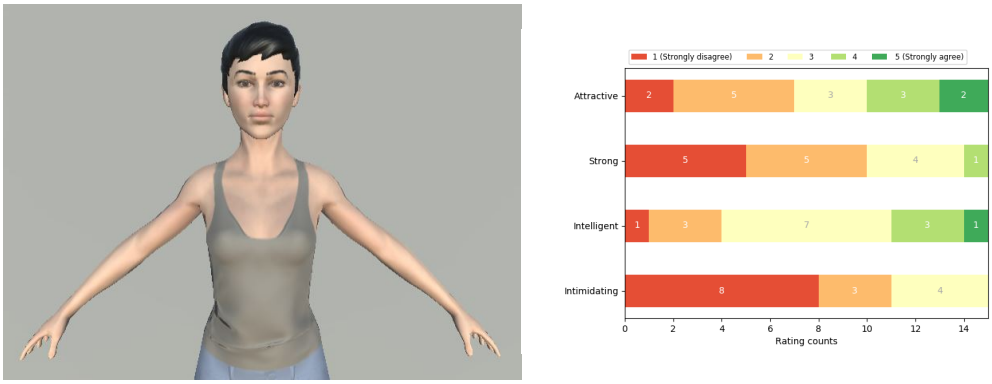


Figure 61: Female 16 thumbnail and ratings.

Figure 62: Female 17 thumbnail and ratings.

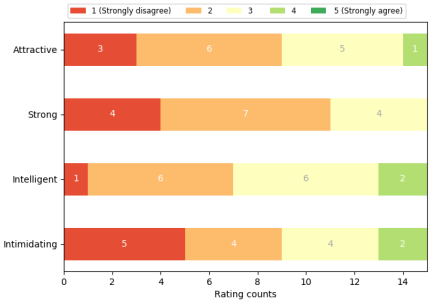


Figure 63: Female 18 thumbnail and ratings.

9.9.2 Males

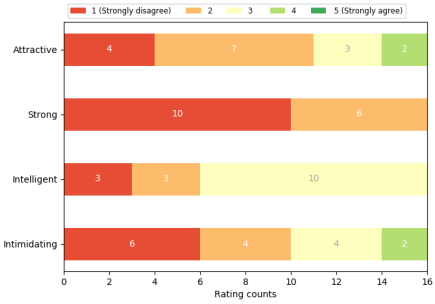


Figure 64: Male 1 thumbnail and ratings.

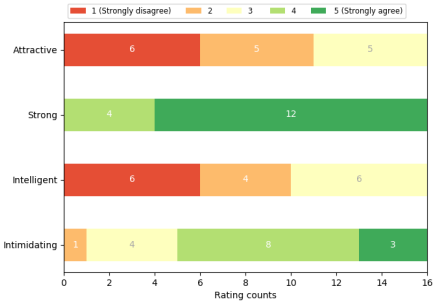


Figure 65: Male 2 thumbnail and ratings.

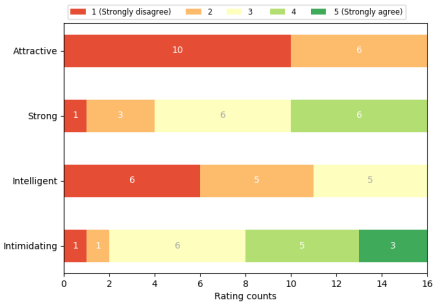


Figure 66: Male 3 thumbnail and ratings.

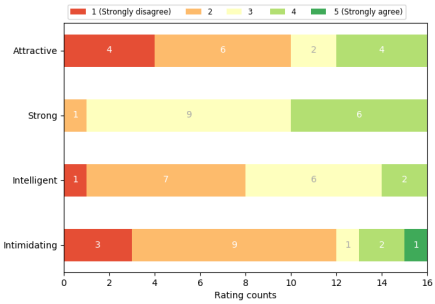


Figure 67: Male 4 thumbnail and ratings.

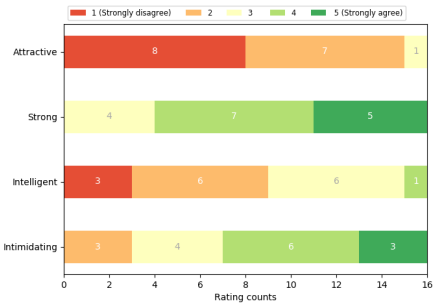


Figure 68: Male 5 thumbnail and ratings.

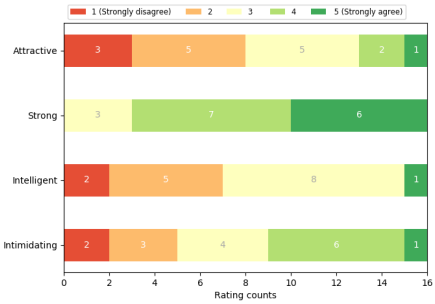


Figure 69: Male 6 thumbnail and ratings.

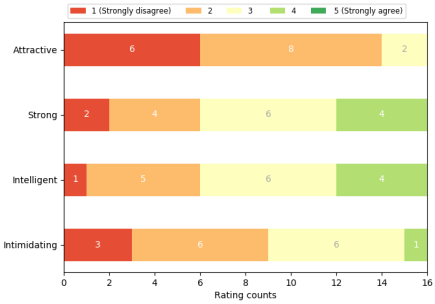


Figure 70: Male 7 thumbnail and ratings.

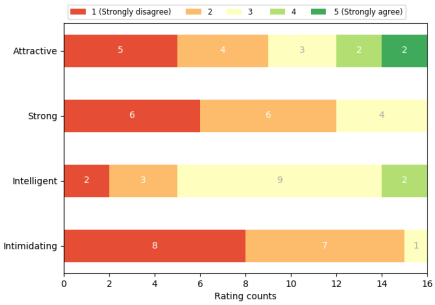


Figure 71: Male 8 thumbnail and ratings.

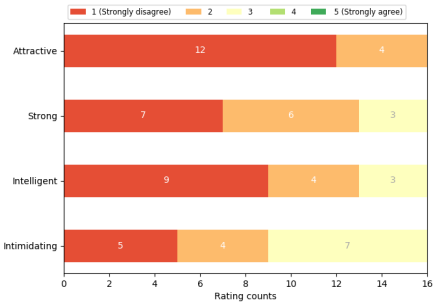


Figure 72: Male 9 thumbnail and ratings.

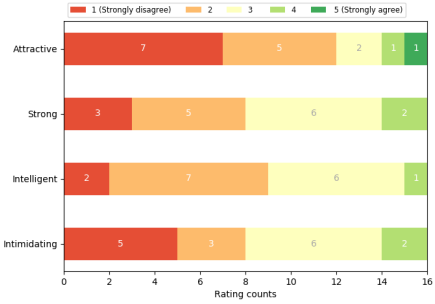


Figure 73: Male 10 thumbnail and ratings.

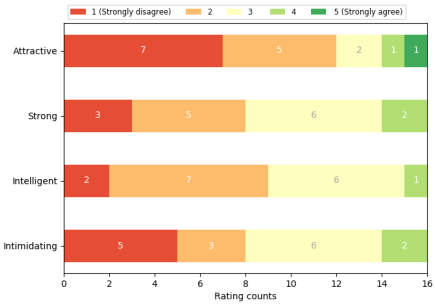


Figure 74: Male 10 thumbnail and ratings.

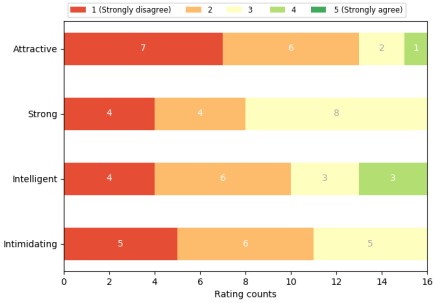


Figure 75: Male 11 thumbnail and ratings.

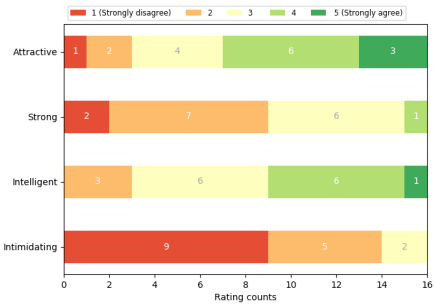


Figure 76: Male 12 thumbnail and ratings.

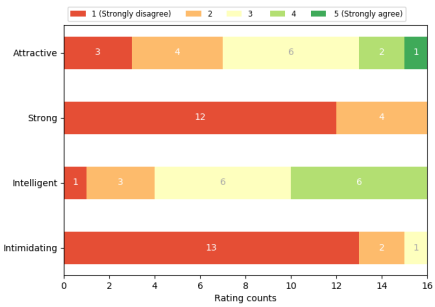


Figure 77: Male 13 thumbnail and ratings.

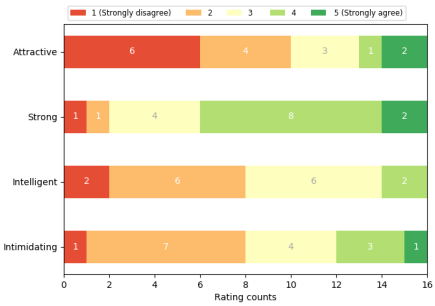


Figure 78: Male 14 thumbnail and ratings.

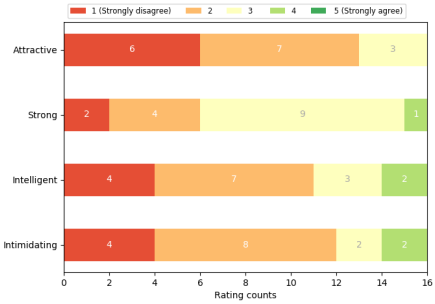


Figure 79: Male 15 thumbnail and ratings.

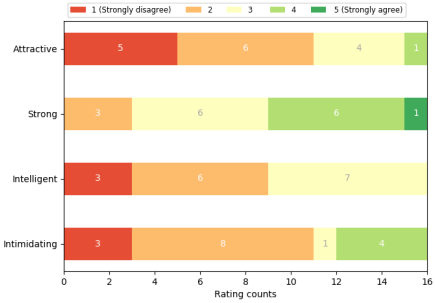


Figure 80: Male 16 thumbnail and ratings.

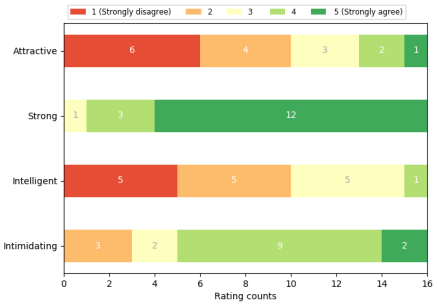


Figure 81: Male 17 thumbnail and ratings.

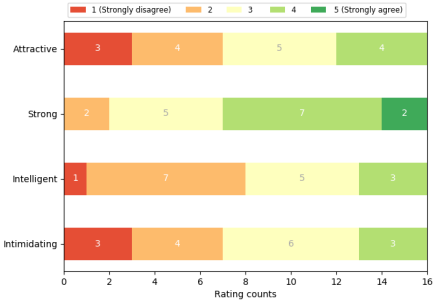


Figure 82: Male 18 thumbnail and ratings.

9.10 Recruiting Poster

10 References

[1] Ferran Argelaguet, Ludovic Hoyet, Michaël Trico, and Anatole Lécuyer. The role of interaction in virtual embodiment: Effects of the virtual hand representation. In *2016 IEEE Virtual Reality (VR)*, pages 3–10. IEEE, 2016.

[2] Zahra Ashktorab, Mohit Jain, Q Vera Liao, and Justin D Weisz. Resilient chatbots: Repair strategy preferences for conversational breakdowns. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, page 254. ACM, 2019.

[3] Jeremy N Bailenson and Andrew C Beall. Transformed social interaction: Exploring the digital plasticity of avatars. In *Avatars at work and play*, pages 1–16. Springer, 2006.

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

[5] Jakki O Bailey, Jeremy N Bailenson, and Daniel Casasanto. When does virtual embodiment change our minds? *Presence: Teleoperators and Virtual Environments*, 25(3):222–233, 2016.

[6] Domna Banakou, Parasuram D Hanumanthu, and Mel Slater. Virtual embodiment of white people in a black virtual body leads to a sustained reduction in their implicit racial bias. *Frontiers in human neuroscience*, 10:601, 2016.

[7] Daryl J Bem. Self-perception theory. In *Advances in experimental social psychology*, volume 6, pages 1–62. Elsevier, 1972.

[8] Jim Blascovich. A theoretical model of social influence for increasing the utility of collaborative virtual environments. In *Proceedings of the 4th international conference on Collaborative virtual environments*, pages 25–30. ACM, 2002.

[9] Michael H Bond. Effect of an impression set on subsequent behavior. *Journal of Personality and Social Psychology*, 24(3):301, 1972.

[10] Matthew Botvinick and Jonathan Cohen. Rubber hands 'feel'touch that eyes see. *Nature*, 391(6669):756, 1998.

[11] Doug A Bowman and Ryan P McMahan. Virtual reality: how much immersion is enough? *Computer*, 40(7):36–43, 2007.

[12] Harry Brenton, Marco Gillies, Daniel Ballin, and David Chatting. The uncanny valley: does it exist and is it related to presence. *Presence connect*, 2005.

[13] Chris Christou and Despina Michael. Aliens versus humans: Do avatars make a difference in how we play the game? In *2014 6th International Conference on Games and Virtual Worlds for Serious Applications (VS-GAMES)*, pages 1–7. IEEE, 2014.

[14] Ron Dotsch and Daniël HJ Wigboldus. Virtual prejudice. *Journal of experimental social psychology*, 44(4):1194–1198, 2008.

[15] Jesse Fox, Sun Joo Ahn, Joris H Janssen, Leo Yeykelis, Kathryn Y Segovia, and Jeremy N Bailenson. Avatars versus agents: a meta-analysis quantifying the effect of agency on social influence. *Human–Computer Interaction*, 30(5):401–432, 2015.

[16] Jesse Fox and Jeremy N Bailenson. Virtual self-modeling: The effects of vicarious reinforcement and identification on exercise behaviors. *Media Psychology*, 12(1):1–25, 2009.

[17] Jesse Fox, Jeremy N Bailenson, and Liz Tricase. The embodiment of sexualized virtual selves: The proteus effect and experiences of self-objectification via avatars. *Computers in Human Behavior*, 29(3):930–938, 2013.

[18] Azucena Garcia-Palacios, Hunter Hoffman, Albert Carlin, TA Furness Iii, and Cristina Botella. Virtual reality in the treatment of spider phobia: a controlled study. *Behaviour research and therapy*, 40(9):983–993, 2002.

[19] Tom Geller. Overcoming the uncanny valley. *IEEE computer graphics and applications*, 28(4):11–17, 2008.

[20] Mar Gonzalez-Franco and Jaron Lanier. Model of illusions and virtual reality. *Frontiers in psychology*, 8:1125, 2017.

[21] Victoria Groom, Jeremy N Bailenson, and Clifford Nass. The influence of racial embodiment on racial bias in immersive virtual environments. *Social Influence*, 4(3):231–248, 2009.

[22] Rosanna E Guadagno, Jim Blascovich, Jeremy N Bailenson, and Cade McCall. Virtual humans and persuasion: The effects of agency and behavioral realism. *Media Psychology*, 10(1):1–22, 2007.

[23] Wijnand A IJsselsteijn, Yvonne A W de Kort, and Antal Haans. Is this my hand i see before me? the rubber hand illusion in reality, virtual reality, and mixed reality. *Presence: Teleoperators and Virtual Environments*, 15(4):455–464, 2006.

[24] Mohit Jain, Pratyush Kumar, Ramachandra Kota, and Shwetak N Patel. Evaluating and informing the design of chatbots. *Proceedings of the 2018 Designing Interactive Systems Conference*, pages 895–906, 2018.

[25] Konstantina Kilteni, Raphaela Groten, and Mel Slater. The sense of embodiment in virtual reality. *Presence: Teleoperators and Virtual Environments*, 21(4):373–387, 2012.

[26] Oswald D Kothgassner, Mirjam Griesinger, Kathrin Kettner, Katja Wayan, Sabine Völkl-Kernstock, Helmut Hlavacs, Leon Beutl, and Anna Felnhofer. Real-life prosocial behavior decreases after being socially excluded by avatars, not agents. *Computers in human behavior*, 70:261–269, 2017.

[27] Carmen E Lefevre, Gary J Lewis, David I Perrett, and Lars Penke. Telling facial metrics: facial width is associated with testosterone levels in men. *Evolution and Human Behavior*, 34(4):273–279, 2013.

[28] Lorraine Lin and Sophie Jörg. Need a hand?: how appearance affects the virtual hand illusion. In *Proceedings of the ACM Symposium on Applied Perception*, pages 69–76. ACM, 2016.

[29] Michael Meehan, Brent Insko, Mary Whitton, and Frederick P Brooks Jr. Physiological measures of presence in stressful virtual environments. In *Acm transactions on graphics (tog)*, volume 21, pages 645–652. ACM, 2002.

[30] Andreas Mühlberger, Matthias J Wieser, Ramona Kenntner-Mabiala, Paul Pauli, and Brenda K Wiederhold. Pain modulation during drives through cold and hot virtual environments. *CyberPsychology & Behavior*, 10(4):516–522, 2007.

[31] Clifford Nass, Jonathan Steuer, and Ellen R Tauber. Computers are social actors. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, pages 72–78. ACM, 1994.

[32] Richard E Nisbett and Timothy D Wilson. The halo effect: evidence for unconscious alteration of judgments. *Journal of personality and social psychology*, 35(4):250, 1977.

[33] Jean-Marie Normand, Elias Giannopoulos, Bernhard Spanlang, and Mel Slater. Multisensory stimulation can induce an illusion of larger belly size in immersive virtual reality. *PloS one*, 6(1):e16128, 2011.

[34] Kristine L Nowak and Frank Biocca. The effect of the agency and anthropomorphism on users' sense of telepresence, copresence, and social presence in virtual environments. *Presence: Teleoperators & Virtual Environments*, 12(5):481–494, 2003.

[35] Xueni Pan and Mel Slater. A preliminary study of shy males interacting with a virtual female. In *Presence: The 10th Annual International Workshop on Presence*. Citeseer, 2007.

[36] Thomas D Parsons and Albert A Rizzo. Affective outcomes of virtual reality exposure therapy for anxiety and specific phobias: A meta-analysis. *Journal of behavior therapy and experimental psychiatry*, 39(3):250–261, 2008.

[37] Jorge Peña, Jeffrey T Hancock, and Nicholas A Merola. The priming effects of avatars in virtual settings. *Communication Research*, 36(6):838–856, 2009.

[38] Jorge Peña, Subuhi Khan, and Cassandra Alexopoulos. I am what i see: How avatar and opponent agent body size affects physical activity among

men playing exergames. *Journal of Computer-Mediated Communication*, 21(3):195–209, 2016.

[39] Jorge Peña and Eunice Kim. Increasing exergame physical activity through self and opponent avatar appearance. *Computers in Human Behavior*, 41:262–267, 2014.

[40] Michael I Posner, Mary J Nissen, and Raymond M Klein. Visual dominance: an information-processing account of its origins and significance. *Psychological review*, 83(2):157, 1976.

[41] Marieke Rohde, Massimiliano Di Luca, and Marc O Ernst. The rubber hand illusion: feeling of ownership and proprioceptive drift do not go hand in hand. *PloS one*, 6(6):e21659, 2011.

[42] Aaron Sell, Leda Cosmides, John Tooby, Daniel Sznycer, Christopher von Rueden, and Michael Gurven. Human adaptations for the visual assessment of strength and fighting ability from the body and face. *Proceedings of the Royal Society B: Biological Sciences*, 276(1656):575–584, 2008.

[43] Jun'ichiro Seyama and Ruth S Nagayama. The uncanny valley: Effect of realism on the impression of artificial human faces. *Presence: Teleoperators and virtual environments*, 16(4):337–351, 2007.

[44] Mel 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.

[45] Mel Slater, Angus Antley, Adam Davison, David Swapp, Christoph Guger, Chris Barker, Nancy Pistrang, and Maria V Sanchez-Vives. A virtual reprise of the stanley milgram obedience experiments. *PloS one*, 1(1):e39, 2006.

[46] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria V Sanchez-Vives. Towards a digital body: the virtual arm illusion. *Frontiers in human neuroscience*, 2:6, 2008.

[47] Mel Slater, Daniel Pérez Marcos, Henrik Ehrsson, and Maria V Sanchez-Vives. Inducing illusory ownership of a virtual body. *Frontiers in neuroscience*, 3:29, 2009.

[48] Mel Slater and Maria V Sanchez-Vives. Transcending the self in immersive virtual reality. *Computer*, 47(7):24–30, 2014.

[49] Mel Slater, Martin Usoh, and Anthony Steed. Taking steps: the influence of a walking technique on presence in virtual reality. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 2(3):201–219, 1995.

[50] Misha Sra and Chris Schmandt. Metaspace ii: Object and full-body tracking for interaction and navigation in social vr. *arXiv preprint arXiv:1512.02922*, 2015.

[51] Jan-Philipp Stein, Benny Liebold, and Peter Ohler. Stay back, clever thing! linking situational control and human uniqueness concerns to the aversion against autonomous technology. *Computers in Human Behavior*, 95:73–82, 2019.

[52] Brandon Van Der Heide, Erin M Schumaker, Ashley M Peterson, and Elizabeth B Jones. The proteus effect in dyadic communication: Examining the effect of avatar appearance in computer-mediated dyadic interaction. *Communication Research*, 40(6):838–860, 2013.

[53] Sonja Windhager, Katrin Schaefer, and Bernhard Fink. Geometric morphometrics of male facial shape in relation to physical strength and perceived attractiveness, dominance, and masculinity. *American Journal of Human Biology*, 23(6):805–814, 2011.

[54] Silke Wohlrab, Bernhard Fink, Peter M Kappeler, and Gayle Brewer. Perception of human body modification. *Personality and Individual Differences*, 46(2):202–206, 2009.

[55] Andrea Stevenson Won, Jeremy Bailenson, Jimmy Lee, and Jaron Lanier. Homuncular flexibility in virtual reality. *Journal of Computer-Mediated Communication*, 20(3):241–259, 2015.

[56] Nick Yee and Jeremy Bailenson. The proteus effect: The effect of transformed self-representation on behavior. *Human communication research*, 33(3):271–290, 2007.

[57] Nick Yee and Jeremy N Bailenson. The difference between being and seeing: The relative contribution of self-perception and priming to behavioral changes via digital self-representation. *Media Psychology*, 12(2):195–209, 2009.

[58] Nick Yee, Jeremy N Bailenson, and Nicolas Ducheneaut. The proteus effect: Implications of transformed digital self-representation on online and offline behavior. *Communication Research*, 36(2):285–312, 2009.

[59] Shanyang Zhao. Toward a taxonomy of copresence. *Presence: Teleoperators & Virtual Environments*, 12(5):445–455, 2003.

[60] Philip G Zimbardo. The human choice: Individuation, reason, and order versus deindividuation, impulse, and chaos. In *Nebraska symposium on motivation*. University of Nebraska press, 1969.