Ecological Validity and the Evaluation of Avatar Facial Animation Noise

Marta Wilczkowiak, Ken Jakubzak, James Clemoes, Cornelia Treptow, Michaela Porubanova, Kerry Read, Daniel McDuff, Marina Kuznetsova, Sean Rintel, Mar Gonzalez-Franco
Microsoft

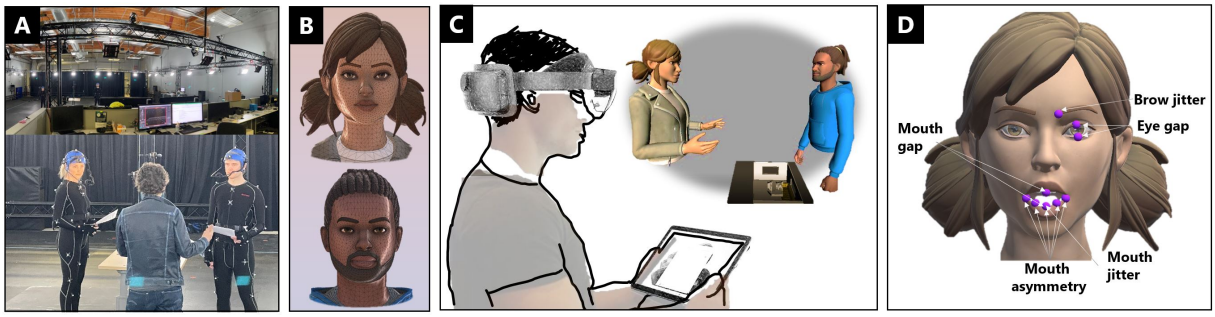

Figure 1: **A**. Mo-cap of a social interaction scenario for the experiment. **B**. The mo-cap was remapped to high-end stylized avatars. **C**. Participants experienced the social interaction wearing a HoloLens2 and responded to questions on a tablet app. **D**. Facial features of the avatars were degraded with different types and severity of animation noises.

Abstract

Facial animation noise levels affect the acceptance of avatars in communication systems. However, there is no standard for evaluation, especially with regard to ecological validity. We investigate low and high ecological validity on two within-subjects experiments conducted in Augmented Reality on a Hololens2. We simulated facial-expression noise introduced on stylized cartoon avatars, and found that in the high ecological validity experiment, subjects were less sensitive to noise parameters, but their judgement was more influenced by empathy scores and gender biases. This highlights the importance of considering both technical parameters and user experience when designing communication systems. We make some general recommendations for evaluating issues of avatar acceptance given the trade-offs between the approaches, and propose the 'Triple-C' factors of Context, Culture and Character as an important set of ecological factors to consider.

Index Terms: Human-centered computing—Empirical studies in HCI; Human-centered computing—HCI design and evaluation methods

1 Introduction

Avatars are essential in Virtual Reality (VR) and Augmented Reality (AR), and are now even available in traditional 2D video meetings [25, 35, 41, 43, 65]. A key aspect to the advancement of avatars and their transfer to products is the design of evaluation schemes that take into account users' behaviours and perception of their own and others' avatars [36].

Many confounding factors can alter users' perceptions of avatars. Parameters assumed to be universal, such as animation noise thresholds, seem to change from study to study, reducing comparability, repeatability, and extrapolation of work, ultimately serving to hinder progress within the field. The struggle is one of ecological validity, which is the extent to which an experiment reproduces the situation and environment in which the technology will be used, and hence the extent to which the research findings generalize. This not only hampers the advance of the pipeline tools, but also ultimately has ethical considerations to the final impact these digital representations of people will have on other people and themselves.

In this paper, we aim to better understand the acceptability of facial animation noise on stylized cartoon avatars. We explore different types and levels of degradation, including lip and brow jitter, full closure of the eyes and mouth, facial asymmetries, and drops in frame rate. We present the types of noise in isolation and also evaluate them in combination, exploring different forms of ecological validity: i.e. presenting the avatars with the noise in isolation to participants or as part of a full conversation. We report on two studies conducted in AR on a Hololens2 which build on top of a pre-recorded, high-quality mo-cap setting. Our research questions are:

- Which type of facial animation noise has the greatest impact on avatar acceptance?
- What is the threshold at which facial animation noise negatively impacts communication experience?
- How does ecological validity affect this threshold?

2 Related Work

A particular aspect of avatar evaluation that is very prone to being affected by ecological validity is communication, because the combination of verbal and nonverbal cues are salient. Nonverbal behaviours need to be transferred to avatars to enable comfortable and successful communication, especially facial expressions [9].

2.1 Creation of avatars

Creating believable facial animations has been the topic of varied research approaches. Physically-based simulations can produce realistic animations as long as the underlying model and simulation framework are faithful to the anatomy of the face and physics of facial tissue [50]. Image-based avatars use warping and blending under the guidance of coarse geometry to generate real-time facial animations with fine-scale details [10]. End-to-end approaches for facial speech animation are becoming more common as supervised learning can simulate complex, non-linear relationships [53].

Prior research on capture and reconstruction has worked on eye tracking [1, 7, 32, 59], face tracking [31, 47, 58], and body tracking [2, 3, 20, 51] of movements while a person is wearing a head-mounted display (HMD). Some systems are good at recreating full facial animation from audio alone [16, 47]. Advanced systems, like Holoportation [40] or Codec Avatars [12] demonstrate the possibility of photorealistic avatars. However, they require complex apparatus and do not work off-the self when wearing a HMD, which may be a reason why photorealistic avatars have not been adopted more widely in commercial systems.

Despite the general intuition that stylized avatars might not suit certain situations, they still rank higher on aspects such as friendliness, trustworthiness, and appeal compared to more realistic avatars when uncanny valley effects are not properly addressed [15, 34, 42]. Stylized avatars can produce a strong enfacement illusion by presenting realistic facial animations, in particular lip sync [23], even when they do not look like the real person.

Regardless of the avatar used, deciding whether an avatar's animation is of sufficient quality for its intended use-case remains an issue. Avatars' facial animations can also be affected by noise in the driving signals and network problems. Animations can be be "driven" using video capture, dialogue, audio, or a combination of the three [5]. Motions driven by tracking individual video frames can cause discontinuities in animations and lead to jitter [61]. Such animations can be constrained using physics-based priors [46] or smoothed. However, these risk removing nuances in behavior [30]. Disruptions of network signals (e.g. latency) have long been shown to disrupt turn-taking in video meetings [6], and similar problems are apparent in immersive tech [22].

2.2 Perception of avatars

Research has looked at "budgeting" quality of eye gaze, blinking, mouth animation, and microexpressions [36], and also at desyncronization of lip and arm motions [13]. Avatars can transmit personality traits [48], and also have different cultural attachment [14]. The realism of animations will affect the appeal of a virtual face [29]. Therefore, facial parameters should be heavily considered in the design of avatars [39].

Eyes are a key aspect of facial emotions. Spatial and temporal aspects of naturally occurring blinks are driven by an asymmetry in the eye's closing and opening, affecting perceived naturalness [55]. This finding might also change the way blinks are compressed and executed in systems that use avatars.

The mouth also plays crucial role in expressing and perceiving emotional cues. Incorrect mouth movement can result in the incorrect interpretation of affective states and impact the sense of presence in virtual interactions [38]. Erroneous mouth movement during speech can also reduce the impact of positive effects, like phonemic restoration [22]. Morphological information cannot be removed without impacting perceptions, e.g. both posed and spontaneous smiles are perceived as less genuine when compressed via a linear model applied to the blendshapes [56].

Mouth movements can also be used to control emotional transitions, and they interact with other facial movements [57]. Different parts of the face can be temporally aligned to elicit emotions, e.g. in work on hiding true emotions, mouth movements have been found to retrospectively conceal micro-expressions in the eyes [26]. These effects are also culturally dependent [56], e.g. some Asian cultures focus emotional attention on eyes while some Western cultures focus on the mouth [62, 64].

2.3 Avatars, animation noise, and ecological validity

Compared to evaluations of avatars in terms of scales of realism, research on the impact of animation noise for avatars is surprisingly sparse. The focus of research is often on noise-resilience of models [17, 18, 44]. The impact of noise on the communicative experience of avatars is less well understood. When it is investigated, the focus tends to be on one's own avatar. Evaluations of the impact of noise on the animation of interlocutors' avatars or the avatars of others in communicative situations are underexplored [54]. Alexanderson [4] investigated how noise in the signal could affect how clearly communicative expressions could be interpreted by an observer. However, these studies do not provide direct evidence for thresholds of acceptability for avatar facial animation noise.

Much of the research above also tends to be low in ecological validity. The most common work in which avatars are used in scenarios requiring something akin to ecological validity is in the use of VR for simulation of problematic social situations, such as sexual harassment training [45, 49]. In that research, agent-avatars play roles in a larger immersive scenarios. While clearly laudable, such research does not explore the effects of noise on agent or human avatars, as that would defeat the purpose of the simulation effect.

In sum, there is a clear gap in the fields' understanding of threshold of acceptability for facial animation noise of avatars as representations of people, and, further, a gap in the fields' understanding of how findings will be impacted by low versus high ecologically validity in studies. We address these gaps in this paper by reporting on two studies investigating participants' responses to different types and levels of facial animation noise in low and high ecological validity conditions.

3 Materials and Methods

3.1 Recording

To create a relatable situation, we used motion capture to animate two avatars to be viewed in a scenario to be viewed in AR on a Hololens2. In our plot we set a scene in which an architect and a client are involved in a house remodelling conversation (Figure 1). This scenario allowed us to introduce conflict as part of a professional conversation that was also easy to understand. Our script included multiple emotions and provided a balanced talk time for both characters. Capturing professional actors helped achieve varied and natural facial expressions.

To counterbalance gender we recorded two pairs of videos with the male and female actors switching the client and architect roles. Recording was completed inside a professional capture volume (sized 33'7" x 35'9") with 60 Prime-41 Cameras from OptiTrack. We also used 2 Head Mounted Cameras (HMC) Mark IV's from Faceware. Audio was captured with DPA 4080 mics. This was compiled by Telestream Lightspeed, and fine-tuned as our stylized avatars contained 91 blendshapes.

3.2 Types of Noise

We varied facial animation frame rate (fps), and the following noise types: Asymmetry of the Face, Brow Jitter, Eye Closure, Mouth Closure, and Mouth Jitter (Table 1). For sample-to-sample jitter, we used the root mean square (RMS) of the change in the position and orientation values reported by the points represented in Figure 1D. RMS indicates the size of jumps that are made from frame to frame in the output in 1-30 Hz, and thus indicates the velocity of the noise in the position and orientation measures [37]. The larger the RMS, the more visible the jitter artifacts become. RMS for a data segment of n samples is given by:

$$RMS_m = \sqrt{\frac{1}{n}\sum_{i=1}^{n-1}\Delta m_1^2}$$

where Δm_i is the difference between samples i and i+1 of the measure m under consideration.

3.3 Experimental Procedure

Participants (N=56) were recruited in Washington, USA by a research recruitment firm. All participants provided informed con-

Table 1: Types of Noises and Ranges

Name, range [units] (2sf)	How noise is applied	Metric calculation
Asymmetry, 1.6-4.2 [mm]	Opens and rolls the mouth on one side, representative of real HMD driven FT artefact.	Measures average asymmetric offset of vertices left and right of the lower lip center in meters.
Brow Jitter, 0.042-0.63 [mm]	Pre-recorded brow jitter representative to the device is added to the sequence.	Measures RMS distance travelled by a vertex on the inner edge of left eyebrow in 1-30Hz in meters.
Closing Eye, 11-12 gap [mm]	Stops eye closing at a level (e.g., gap between eyelids is always at least 1mm).	Measures 5% lower percentile of the absolute distance between vertices in the middle of the left upper and lower eyelids in meters.
Closing Mouth, 2.2-10 gap [mm]	Stops mouth closing at certain level (e.g., gap between lips is always at least 1mm).	Measures 5% lower percentile of the absolute distance between middle vertices on the lips in meters.
Mouth Jitter, 0.23-0.84 [mm]	Simulated lip jitter representative to audioface is added to the system.	Measures RMS distance travelled by a vertex in the middle of the lower lip in 1-30Hz in meters.
Frame Rate, 5, 10, 15, 30 fps [frames/sec]	Key frames set at given FPS and interpolated betwee.n	Frames rendered per second (discrete values).

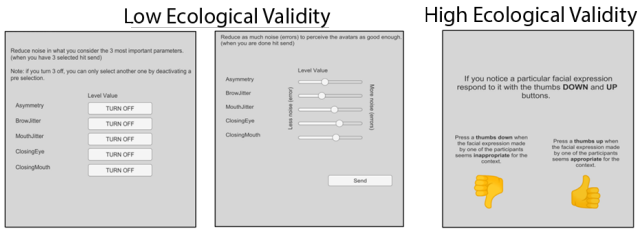

Figure 2: Companion app for the 2 experiments, measuring thresholds of perception, acceptability, and sensitivity to the noises in low and high ecological validity conditions.

sent and the experimental procedure was approved by an Institutional Review Board (anonymized for review). Participants completed a short demographic questionnaire and the Toronto Empathy questionnaire [52]. An experimenter explained how to use a HoloLens2 and participants then observed different versions of the pre-recorded avatar interactions. A tablet computer was connected to the HoloLens2 and used by participants to control the experimental protocols. An experimenter was always available, visible through a glass door, but not in the room (due to COVID-19 restrictions).

The HoloLens2 rendering app and companion experimental control app were created with Unity and connected through ZeroMQ. At each step, the companion app sent the ID of the video to be played, the type of noise introduced, and the video time. The HoloLens2 app was therefore controlled through the tablet companion app at all times. The tablet also recorded all experimental data and responses.

3.4 Experiments

Two experiments were conducted. Experiment 1 used low ecological validity conditions, i.e., where types of noise were detached from their conversational setting, and even shown in isolation from others. Experiment 2 used a high ecological validity condition, i.e., where the facial distortions were seen in a social setting (Figure 2). Studies with high ecological validity require vastly different design approaches than studies with low ecological validity. We highlight the key considerations in Table 2 and then describe each experiment in more depth. All conditions were counterbalanced in our experiments.

3.4.1 Experiment 1 - Low Ecological Validity Conditions

In Experiment 1 we played only separate and pre-selected 20-second clips of the script during which both characters had the opportunity to talk. These clips showed a large range of facial expressions, but had low ecological validity because they were taken from the conversation without regard to context. Participants had to increase the level of noise in each of the categories until the face looked acceptable for them. Thresholds for an acceptable level of noise for each type of noise when presented individually.

In this experiment, we tested the types of noise in isolation and in combination. Therefore participants undertook the following tasks:

- Isolated. Using a slider on the companion app, participants were asked to increase the magnitude of noise in the avatar until it felt like too much for the participant. They did this for only one parameter at a time: mouth jitter, brow jitter, mouth closure, eye closure, facial asymmetry, and fps.
- Combined. Participants were then allowed to reduce the pre-selected thresholds using sliders again, looking for an acceptable level for each type of noise when presented simultaneously.

Additionally, all types of noise were presented simultaneously at the levels participants had pre-selected in isolation and were asked to select the three noises they considered a priority to be reduced, whatever were the most annoying to them.

3.4.2 Experiment 2 - High Ecological Validity Condition

In Experiment 2, we replayed the full conversations. While watching the conversations, participants were asked to imagine themselves as a second client. And were asked to respond by pressing a 'thumbs-up' button when they saw any avatar make a facial expression they deemed 'appropriate' for the context, and a 'thumbs-down' button when they saw an 'inappropriate' facial expression. They could press the buttons as many times they wanted.

Every 20 seconds the conversation paused and the participant had to press a 'Continue' button to resume the conversation. We used that moment to introduce a randomly-assigned new noise parameter and level. This break also helped maintain participants' attention levels and ensure breaks were taken if necessary.

In each break, we added particular facial animation degradation one at time, at different strengths. Both avatars always had the same

Table 2: Key differences in design choices when designing Low and High Ecological Validity experiments.

Aspect	Experiment 1 - Low Ecological Validity	Experiment 2 - High Ecological Validity
Stimuli	20-second conversation clips without context.	Full conversations with context.
Task	Adjust the level of acceptable noise for different facial animation parameters. This enables quick evaluation of many levels and is an effective way to tune the user's acceptability threshold, but lacks context.	To employ context in the evaluations, participants were asked to focus on the conversation and provide feedback by pressing "thumbs-up" or "thumbs-down" for appropriate or inappropriate facial expressions respectively.
Noise	Active participant-driven threshold selection Applied in isolation and in combination.	Passive, introduced one at a time at different strengths without participant's knowledge.
Baseline	No noise applied. Allowed participants to quickly evaluate a large number of noise values, making it practical to enable them to tune the noise from zero to high values in the low ecologically valid condition.	Direct comparisons of realisable noise levels, ranging from just under to just above the thresholds estimated from small number of initial testers. The high ecologically valid condition limited the number of noise levels that could be tested to only four values per noise type. Since zero noise level is impossible in practice, the decision was made to allocate the noise level budget to the evaluation of values that are achievable in practice.

type and level of degradation. For example, if we were exploring facial asymmetry through that 20-second block, participants only saw different levels of asymmetries. Then they would see the same with mouth jitter and so on. Each type of degradation was presented at 4 different levels. Each script played for 20 seconds before pausing. After all 4 levels were shown, a different type of degradation was presented. The order of presentation of degradation types and levels was randomized for each block.

The presentation order of the pre-recorded avatar conversation videos (total of 4 videos) in which the architect could be a woman or a man, and the outcome could be agreement or disagreement, was randomized to reduce ordering effects. All participants saw the full four versions of the scripts in action.

4 Results

4.1 Experiment 1 - Low Ecological Validity

In our first experiment, when comparing the numerical results on the noise acceptance values, we found that when all types of noise were shown simultaneously, participants significantly lowered their acceptability thresholds. This reduction in thresholds could have been caused by the combined effects of facial animation noise having a bigger effect on the experience. However, we cannot rule out that the ordering effects could have predisposed study participants to lower their thresholds.

4.1.1 Priorities of Noises

When all types of noise were shown in combination, and before asking participants to adjust the sliders, participants were asked to remove 3 out of 5 types of noise, prioritizing those which were the most annoying. Almost all participants removed mouth jitter, with the mouth-closing parameter gathering the second-highest number of responses. Eye-closing noise was turned off by the fewest participants. Asymmetry of the face and brow jitter were similarly rated in the priority list, ranking third (Figure 5).

4.2 Experiment 2 - High Ecological Validity

In our second experiment, we presented the same types of facial animation noises in the context of the full social situation from which the short clips were drawn. Each full script lasted 1:35 seconds. Each segment of conversation lasted 20 seconds. When looking at the engagement levels of the participants with the companion app (Figure 6), response rates for 'appropriate' facial expressions were significantly higher than for 'inappropriate' facial expressions ($F(1) = 86.7$, $p < 0.0001$). This validates the recording and blendshapes of the facial expressions, because they were considered mostly correct and appropriate.

Table 3: List of thresholds for acceptable noise levels where 90% of the participants deemed the experience acceptable.

Parameter	Isolated	Combined	Social Cont.
Asymmetry	0.8mm	0.1mm	3.2mm
Brow Jitter	0.1mm	0.03mm	0.6mm
Closing Eye	2.7mm	0.01mm	12.5mm
Closing Mouth	1.7mm	0.01mm	3.4mm
Mouth Jitter	0.05mm	0.01mm	0.25mm
FPS	15 FPS (87.5%)	N/A	15 fps (90%)

The points at which thresholds of acceptable noise when presented in a social setting were slightly different from the points used on the sliders of Exp. 1. The points when users started reporting more 'inappropriate' facial expressions than 'appropriate' ones can be seen as the interaction effect, aka, the crossing point between the graphs of Figure 4.

Numerical values of the results can be observed in Table 3, and videos of these threshold points are available in the supplementary materials for illustration. The tolerance for noise dropped significantly when in the low ecological validity task in combination. However, tolerance in the high ecological validity condition was much higher across all parameters. For FPS, in which the metric was discrete (not continuous), we find that 15 fps was generally treated as 'good enough' inside the AR experience, representing 87.5% of participants' responses in the isolation mode, and 90% in the social context experiment.

4.2.1 Gender Effects

The script manipulations in the 4 different conditions alternated the architect and client roles between the man and the woman, as well as the outcome of the discussion, which could end in agreement or disagreement. The outcome of the discussion thus changed the affective valence of the whole conversation, making it more positive or negative.

Participant response rates varied by condition of the discussion. A repeated-measures ANOVA on user engagement with the reporting

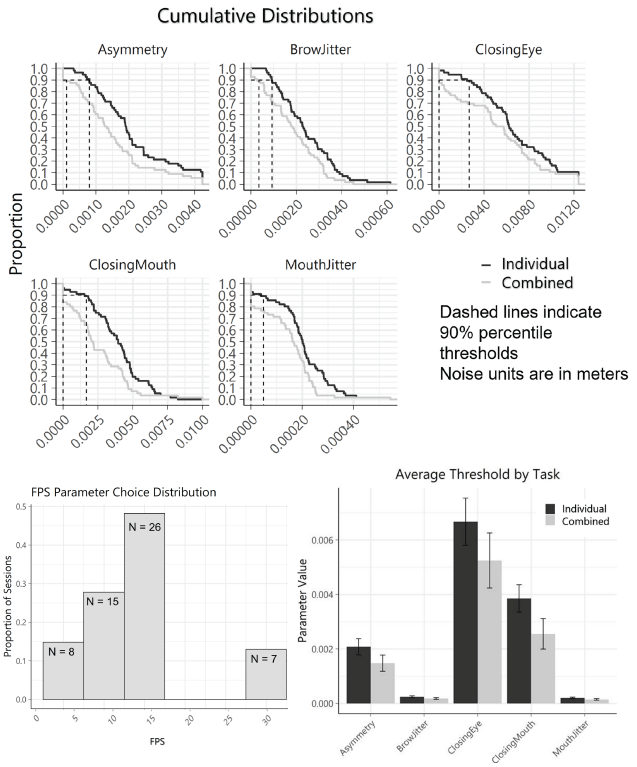

Figure 3: Noise acceptability based on users' slider selections when shown in isolation or in combination. The cumulative graph shows thresholds by noise type. When the facial animation noise was higher fewer participants found the experience within their threshold of acceptability. Dashed lines indicate 90% percentile thresholds. Noise units are in meters. For the particular case of FPS where the data was not continued, the data is presented separated.

tool showed differences in the response rate that depended on the condition of the video, i.e. the gender of the characters and the outcome of the video (F(3)=3, p=0.03). "Female architect agreeing" received the highest number of 'appropriate' responses and "Female architect disagreeing" received the fewest 'appropriate' responses. These differences illuminate the possibility that user biases affected how they perceived facial expressions (Figure 6).

4.3 Empathy Effects

When taking the empathy questionnaire responses of the participants into account, we found that participants with lower empathy scores reported higher rates of 'inappropriate' responses when mouth jitter was present ($p < 0.05$) in the high ecological validity experiment. Overall, participants with higher empathy scores had higher rates of 'appropriate' responses to facial expressions.

Empathy did not affect chosen noise thresholds in the low ecological validity condition. The thresholds for different types of facial animation noise did not differ significantly for the participants with the top 25% empathy scores and participants with the bottom 25%.

5 Discussion

Running high and low ecologically valid conditions on the same cohort, on the same device, with the same assets, brings to light how experimental design impacts the results of a parameter evaluation. This is due to the fact that the necessary differences in the design approaches may introduce confounding variables and complicate direct comparisons. Our primary goal is to bring awareness to the community regarding the range of results that can be obtained

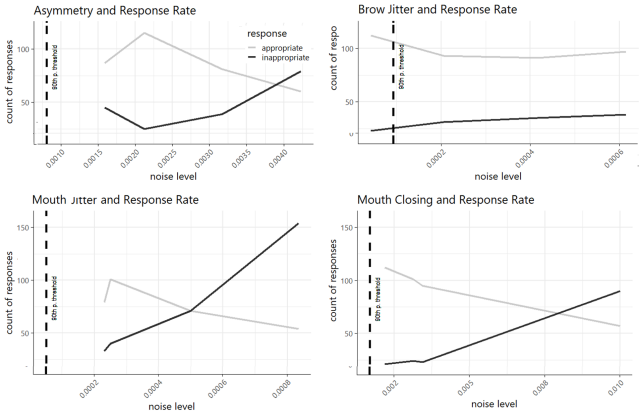

Figure 4: Responses to noise during social setting.

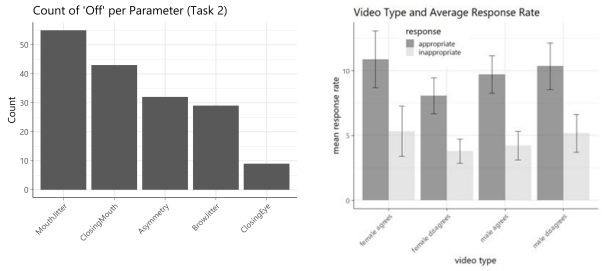

Figure 5: (Top) Participants selected the following noises as priority for them to remove when they were all in combination. (Bottom) Engagement, measured as responses count depending on whether participants found the facial expressions 'appropriate' or 'inappropriate' for the different recorded conditions.

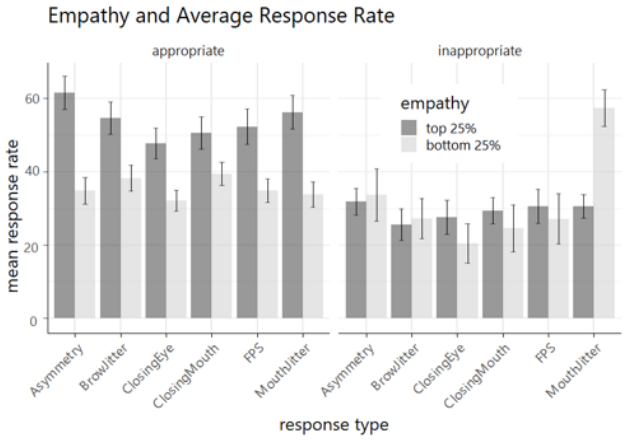

Figure 6: Mean Response Rate by Response Type and Empathy Level; grey indicates highest empathy quantile, faded indicates lowest empathy quantile.

depending on the experimental design, and to establish a common understanding of the levels of uncertainty that should be taken into account when analysing the results. It is our hope that this will lead to more reliable and accurate findings in future research.

The high ecological validity experiment embedded participants in the social setting of a client-expert discussion and asked participants to point out when facial expressions seemed inappropriate, without mentioning noise. Empathy turned out to be a strong factor in the results in such a setting. On the other hand, empathy was much less of a factor in the classic psycho-physics setting of participants specifically tuning noise acceptance thresholds, which was our low ecological validity experiment. In the low ecological validity experiment we found an effect that if noise was presented in isolation, it was tolerated more than when different types of noise were presented simultaneously. Those values seemed to be persistent across Character (empathy) variability.

In sum, as noise levels increase, the quality of participants' experiences decrease. Our results show that mouth jitter had the greatest impact on avatar acceptability, as it was prioritized by participants in both low and high ecological validity conditions. Participants showed higher tolerance for other types of facial animation noise. Crucially, though, the level of ecological validity changes the usefulness of these results for researchers and developers of avatars. When focused on a social task, participants seemed to be less sensitive to the increase of noise levels than if focused on adjusting the acceptable noise in isolation. Beyond task, gender and empathy also appeared to impact results in complex ways.

In the high ecological validity experiment we found that participant engagement was impacted by the gender assigned to the client or the expert, as well as the outcome and tone of the script being positive or negative. When the architect presented as female, her agreement received the highest number of 'appropriate' responses; but when she disagreed, it received the lowest number of 'appropriate' responses. These differences suggest that socio-cultural norms were affecting how participants perceived the experience – in this case possible biases against women in professional contexts [8].

Participant's empathy did not influence the chosen noise thresholds in the low ecological validity experiment. By contrast, in the high ecological validity experiment, participants with lower empathy scores had higher rates of 'inappropriate' responses when mouth jitter was present. This is consistent with other findings that point to the role of mouth movements in emotional interpretation. For example, when perceiving social situations, individuals with lower empathy show greater focus on mouths, bodies, and objects, and that fixation (dwell) times on mouths have been shown to be powerful predictors of degree of social competence [28]. Conversely, participants with higher empathy score had overall higher rates of 'appropriate' responses to facial expressions, showing a clear impact of empathy on evaluation of facial animation noise.

Given these results, we believe that there is clear need for balancing low and high ecological validity studies, especially for evaluating issues such as avatar acceptance when end results is intended to be using in complex human situations such as communication. We suggest the following general recommendations:

- Use cheaper, less ecologically valid experiments to gain an understanding of key trends and interactions between components. However, avoid using these experiments to establish technical requirements, set standards, or draw general conclusions on user sensibility.

- Use more expensive, more ecologically valid experiments to validate srong hypotheses, gain a deeper understanding of experience factors that have not yet been taken into account, and if necessary, set technical requirements.

- Communication being a gestalt experience, many factors will interplay. Neglecting understating of their interactions will lead to unreliable results

On this last point of the gestalt experience of communication, we further suggest that there are three broad factors that need to be taken into account, which we call the 'Triple-Cs': Context, Culture, and Character.

Context refers to the circumstances and setting in which people engage with one another [11]. Presenting an avatar in context ultimately means presenting it in a social setting in which it will be fully understood and assessed as representing social meaning. When presented out of social settings, avatars, pipelines or their animations can be presented as study parameters, that may or not be isolated, or presented in combination. The findings on parameter values for an evaluation will therefore depend on the circumstances of the experiment. The closer to a real setting, the more ecologically valid the results will be. However, a social setting will also add factors which interact with the other aspects of the experiments and characteristics of the users: Culture and Character.

Culture, in the anthropological sense [60], refers to a learned system of shared beliefs. Members of different cultural groups will share distinct ways of perceiving the world and possess certain assumptions that will be shared with others. These and other cultural constructs, such as gender roles or biases, will elicit different outcomes from an evaluation. The Proteus effect [63], also referred to as mimicry [21] has shown that people in VR will tend to behave in stereotypical ways, for example participants trying to play the drums while wearing a cocktail suit will result in less rhythmic music than embodying a more casual avatar [27]. In our experiment we did also find strong impact of the gender of the avatar in the high ecologically valid experiment.

Character can be understood as the personality of an individual, including intra-cultural variability. Although two individuals might or might not come from the same culture, they might respond very differently to avatars when presented in context [19]. The most widely used system of personality traits is the OCEAN Five-Factor Model [33]. Most recently, empathy, as trans-cultural trait, has been considered more reliable and less culturally correlated than the Big-Five [24].

The Triple-C factors do make research more complex. High ecological validity studies that use them will need to be able to differentiate their effects from one another and on dependent variables. However, given our nascent findings on gender and empathy, we also argue that high validity experimentation requires considering how the Triple-C factors matter to achieving balanced, generalizable, and replicable results that evaluate and reduce the effects of bias.

6 LIMITATIONS

Our experimental design choices were influenced by external constraints that are common to this type of experiment. These included the need to recruit participants from diverse backgrounds but a fixed budget per participant for recruitment, limiting the number of participants. Participants also have a limited attention to be able to meaningfully find and rate differences, and as such our versions of low and high ecological validity were themselves constrained. Additionally our setup was a social setting but participants only observed the conversation, while taking part on it could have a different layer to the results and affect also the ecological validity of the experiment.

We acknowledge that since we did not include a gender measure in the low validity experiment, it is unclear how gender bias impacted specific noise types. A limitation particular to the noise study was that we did not collect 0% animation noise baseline results. We made the decision not collect these these results because 0% animation noise is not achievable in practice and that taking the time to collect it would reduce our ability to collect the more meaningful data of the noise values that are actually achievable. Instead, the lowest

perceptible version of each form of animation noise type was judged by animation experts and experiments run using that as a baseline. Neverhteless, the results are robust within participants and as relative noise values.

7 Conclusion

A way to compare and contrast avatars is to explore whether outcomes are different in a high ecological validity or low ecological validity settings. There will also be different responses to whether participants are asked if something is appropriate for a social context than if participants are asked to find a threshold for a parameter (e.g. level of mouth jitter) that they find acceptable. Different results might occur depending of the style of avatar, and also be sub-product of the demographics of participants (as purported users) and how cultural and personality traits impact particularly high ecological validity settings. While low ecological validity tests may be less affected by such factors, ignoring this difference of context evaluation will have significant impact not only on the results and in the experience, but also on research, undermining future replicability or extrapolation of results.

References

[1] K. Ahuja, R. Islam, V. Parashar, K. Dey, C. Harrison, and M. Goel. Eyespyvr: Interactive eye sensing using off-the-shelf, smartphone-based vr headsets. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 2(2):1–10, 2018.

[2] K. Ahuja, E. Ofek, M. Gonzalez-Franco, C. Holz, and A. D. Wilson. Coolmoves: User motion accentuation in virtual reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, 5(2):1–23, 2021.

[3] K. Ahuja, V. Shen, C. M. Fang, N. Riopelle, A. Kong, and C. Harrison. Controllerpose: Inside-out body capture with vr controller cameras. In *CHI Conference on Human Factors in Computing Systems*, pp. 1–13, 2022.

[4] S. Alexanderson. *Performance, Processing and Perception of Communicative Motion for Avatars and Agents*. PhD dissertation, KTH Royal Institute of Technology, Stockholm, 2017. Publisher: KTH Royal Institute of Technology.

[5] D. Aneja, R. Hoegen, D. McDuff, and M. Czerwinski. Understanding conversational and expressive style in a multimodal embodied conversational agent. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–10, 2021.

[6] G. Berndtsson, M. Schmitt, P. Hughes, J. Skowronek, K. Schoenenberg, and A. Raake. *Methods for Human-Centered Evaluation of MediaSync in Real-Time Communication*, pp. 229–270. Springer International Publishing, Cham, 2018. doi: 10.1007/978-3-319-65840-7_9

[7] D. Borland, T. Peck, and M. Slater. An evaluation of self-avatar eye movement for virtual embodiment. *IEEE transactions on visualization and computer graphics*, 19(4):591–596, 2013.

[8] K. Broadbent, G. Strachan, and G. Healy, eds. *Gender and the Professions: International and Contemporary Perspectives*. Routledge, New York, 2017.

[9] J. K. Burgoon, V. Manusov, and L. K. Guerrero. *Nonverbal communication*. Routledge, 2021.

[10] C. Cao, H. Wu, Y. Weng, T. Shao, and K. Zhou. Real-time facial animation with image-based dynamic avatars. *ACM Transactions on Graphics*, 35(4), 2016.

[11] H. Cappelen and J. Dever. *Context and Communication*. Oxford University Press, Oxford, 2016.

[12] H. Chu, S. Ma, F. D. la Torre, S. Fidler, and Y. Sheikh. Expressive telepresence via modular codec avatars. In *European Conference on Computer Vision*, pp. 330–345. Springer, 2020.

[13] T. Collingwoode-Williams, M. Gillies, C. McCall, and X. Pan. The effect of lip and arm synchronization on embodiment: A pilot study. In *2017 IEEE Virtual Reality (VR)*, pp. 253–254, March 2017. doi: 10.1109/VR.2017.7892272

[14] T. D. Do, S. Zelenty, M. Gonzalez-Franco, and R. P. McMahan. Valid: A perceptually validated virtual avatar library for inclusion and diversity. *arXiv preprint arXiv:2309.10902*, 2023.

[15] G. C. Dobre, M. Wilczkowiak, M. Gillies, X. Pan, and S. Rintel. Nice is different than good: Longitudinal communicative effects of realistic and cartoon avatars in real mixed reality work meetings. In *CHI Conference on Human Factors in Computing Systems Extended Abstracts*, pp. 1–7, 2022.

[16] P. Edwards, C. Landreth, M. Popławski, R. Malinowski, S. Watling, E. Fiume, and K. Singh. Jali-driven expressive facial animation and multilingual speech in cyberpunk 2077. In *Special Interest Group on Computer Graphics and Interactive Techniques Conference Talks*, pp. 1–2, 2020.

[17] S. E. Eskimez, R. K. Maddox, C. Xu, and Z. Duan. Noise-resilient training method for face landmark generation from speech. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 28:27–38, 2020. doi: 10.1109/TASLP.2019.2947741

[18] A. Genay, A. Lécuyer, and M. Hachet. Being an avatar "for real": A survey on virtual embodiment in augmented reality. *IEEE Transactions on Visualization and Computer Graphics*, 28(12):5071–5090, 2022. doi: 10.1109/TVCG.2021.3099290

[19] M. Gonzalez-Franco, P. Abtahi, and A. Steed. Individual differences in embodied distance estimation in virtual reality. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 941–943. IEEE, 2019.

[20] M. Gonzalez-Franco, Z. Egan, M. Peachey, A. Antley, T. Randhavane, P. Panda, Y. Zhang, C. Y. Wang, D. F. Reilly, T. C. Peck, et al. Movebox: Democratizing mocap for the microsoft rocketbox avatar library. In *2020 IEEE International Conference on Artificial Intelligence and Virtual Reality (AIVR)*, pp. 91–98. IEEE, 2020.

[21] M. Gonzalez-Franco and J. Lanier. Model of Illusions and Virtual Reality. *Frontiers in Psychology*, 8, 2017.

[22] M. Gonzalez-Franco, A. Maselli, D. Florencio, N. Smolyanskiy, and Z. Zhang. Concurrent talking in immersive virtual reality: on the dominance of visual speech cues. *Scientific reports*, 7(1):1–11, 2017.

[23] M. Gonzalez-Franco, A. Steed, S. Hoogendyk, and E. Ofek. Using facial animation to increase the enfacement illusion and avatar self-identification. *IEEE transactions on visualization and computer graphics*, 26(5):2023–2029, 2020.

[24] T. Guilera, I. Batalla, C. Forné, and J. Soler-González. Empathy and big five personality model in medical students and its relationship to gender and specialty preference: a cross-sectional study. *BMC medical education*, 19(1):1–8, 2019.

[25] D. Higgins and R. McDonnell. A preliminary investigation of avatar use in video-conferencing. In *2021 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 540–541. IEEE, 2021.

[26] M. Iwasaki and Y. Noguchi. Hiding true emotions: micro-expressions in eyes retrospectively concealed by mouth movements. *Scientific reports*, 6(1):1–10, 2016.

[27] K. Kilteni, I. Bergstrom, and M. Slater. Drumming in Immersive Virtual Reality: The Body Shapes the Way We Play. *IEEE Transactions on Visualization and Computer Graphics*, 19(4):597–605, Apr. 2013. Conference Name: IEEE Transactions on Visualization and Computer Graphics. doi: 10.1109/TVCG.2013.29

[28] A. Klin, W. Jones, R. Schultz, F. Volkmar, and D. Cohen. Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Archives of general psychiatry*, 59(9):809–816, 2002.

[29] E. Kokkinara and R. McDonnell. Animation realism affects perceived character appeal of a self-virtual face. In *Proceedings of the 8th ACM SIGGRAPH Conference on Motion in Games*, pp. 221–226, 2015.

[30] J. Lee, J. Chai, P. S. Reitsma, J. K. Hodgins, and N. S. Pollard. Interactive control of avatars animated with human motion data. In *Proceedings of the 29th annual conference on Computer graphics and interactive techniques*, pp. 491–500, 2002.

[31] H. Li, L. Trutoiu, K. Olszewski, L. Wei, T. Trutna, P.-L. Hsieh, A. Nicholls, and C. Ma. Facial performance sensing head-mounted display. *ACM Transactions on Graphics (ToG)*, 34(4):1–9, 2015.

[32] S. Marwecki, A. D. Wilson, E. Ofek, M. Gonzalez Franco, and C. Holz. Mise-unseen: Using eye tracking to hide virtual reality scene changes in plain sight. In *Proceedings of the 32nd Annual ACM Symposium on User Interface Software and Technology*, pp. 777–789, 2019.

[33] R. R. McCrae and P. T. Costa. Validation of the five-factor model of personality across instruments and observers. *Journal of personality and social psychology*, 52(1):81, 1987.

[34] R. McDonnell, M. Breidt, and H. H. Bülthoff. Render me real? investigating the effect of render style on the perception of animated virtual humans. *ACM Transactions on Graphics (TOG)*, 31(4):1–11, 2012.

[35] Microsoft. Join a meeting as an avatar in Microsoft Teams - Microsoft Support, 2023.

[36] M. Murcia-López, T. Collingwoode-Williams, W. Steptoe, R. Schwartz, T. J. Loving, and M. Slater. Evaluating virtual reality experiences through participant choices. In *2020 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 747–755, 2020. doi: 10.1109/VR46266.2020.00098

[37] D. C. Niehorster, L. Li, and M. Lappe. The accuracy and precision of position and orientation tracking in the htc vive virtual reality system for scientific research. *i-Perception*, 8(3):2041669517708205, 2017. PMID: 28567271. doi: 10.1177/2041669517708205

[38] S. Y. Oh, J. Bailenson, N. Krämer, and B. Li. Let the avatar brighten your smile: Effects of enhancing facial expressions in virtual environments. *PloS one*, 11(9), 2016.

[39] C. Oh Kruzic, D. Kruzic, F. Herrera, and J. Bailenson. Facial expressions contribute more than body movements to conversational outcomes in avatar-mediated virtual environments. *Scientific reports*, 10(1):1–23, 2020.

[40] S. Orts-Escolano, C. Rhemann, S. Fanello, W. Chang, A. Kowdle, Y. Degtyarev, D. Kim, P. L. Davidson, S. Khamis, M. Dou, et al. Holoportation: Virtual 3d teleportation in real-time. In *Proceedings of the 29th annual symposium on user interface software and technology*, pp. 741–754, 2016.

[41] P. Panda, M. J. Nicholas, M. Gonzalez-Franco, K. Inkpen, E. Ofek, R. Cutler, K. Hinckley, and J. Lanier. Alltogether: Effect of avatars in mixed-modality conferencing environments. In *2022 Symposium on Human-Computer Interaction for Work*, pp. 1–10, 2022.

[42] V. Phadnis, K. Moore, and M. G. Franco. Avatars in work meetings: Correlation between photorealism and appeal, 2023.

[43] V. Phadnis, K. Moore, and M. Gonzalez-Franco. The work avatar face-off: Knowledge worker preferences for realism in meetings. In *ISMAR*. IEEE, 2023.

[44] C. F. Purps, S. Janzer, and M. Wölfel. Reconstructing Facial Expressions of HMD Users for Avatars in VR. In M. Wölfel, J. Bernhardt, and S. Thiel, eds., *ArtsIT, Interactivity and Game Creation*, Lecture Notes of the Institute for Computer Sciences, Social Informatics and Telecommunications Engineering, pp. 61–76. Springer International Publishing, Cham, 2022. doi: 10.1007/978-3-030-95531-1_5

[45] S. Rawski, J. Foster, and J. Bailenson. Sexual Harassment Bystander Training Effectiveness: Experimentally Comparing 2D Video to VR Practice. *Academy of Management Proceedings*, 2022(1):11526, Aug. 2022. Publisher: Academy of Management. doi: 10.5465/AMBPP.2022.139

[46] D. Rempe, T. Birdal, A. Hertzmann, J. Yang, S. Sridhar, and L. J. Guibas. Humor: 3d human motion model for robust pose estimation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11488–11499, 2021.

[47] A. Richard, C. Lea, S. Ma, J. Gall, F. De la Torre, and Y. Sheikh. Audio- and gaze-driven facial animation of codec avatars. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pp. 41–50, 2021.

[48] K. Ruhland, K. Zibrek, and R. McDonnell. Perception of personality through eye gaze of realistic and cartoon models. In *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, pp. 19–23, 2015.

[49] S. Sadeh-Sharvit, J. Giron, S. Fridman, M. Hanrieder, S. Goldstein, D. Friedman, and S. Brokman. Virtual Reality in Sexual Harassment Prevention: Proof-of-Concept Study. In *Proceedings of the 21st ACM International Conference on Intelligent Virtual Agents*, IVA '21, pp. 87–89. Association for Computing Machinery, New York, NY, USA, Sept. 2021. doi: 10.1145/3472306.3478356

[50] E. Sifakis, A. Selle, A. Robinson-Mosher, and R. Fedkiw. Simulating speech with a physics-based facial muscle model. In *Proceedings of the 2006 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pp. 261–270, 2006.

[51] B. Spanlang, J.-M. Normand, D. Borland, K. Kilteni, E. Giannopoulos, A. Pomés, M. González-Franco, D. Perez-Marcos, J. Arroyo-Palacios, X. N. Muncunill, et al. How to build an embodiment lab: achieving body representation illusions in virtual reality. *Frontiers in Robotics and AI*, 1:9, 2014.

[52] R. N. Spreng, M. C. McKinnon*, R. A. Mar, and B. Levine. The toronto empathy questionnaire: Scale development and initial validation of a factor-analytic solution to multiple empathy measures. *Journal of personality assessment*, 91(1):62–71, 2009.

[53] S. Taylor, T. Kim, Y. Yue, M. Mahler, J. Krahe, A. G. Rodriguez, J. Hodgins, and I. Matthews. A deep learning approach for generalized speech animation. *ACM Transactions on Graphics (TOG)*, 36(4):1–11, 2017.

[54] N. Toothman and M. Neff. The impact of avatar tracking errors on user experience in vr. In *2019 IEEE Conference on Virtual Reality and 3D User Interfaces (VR)*, pp. 756–766, 2019. doi: 10.1109/VR.2019.8798108

[55] L. C. Trutoiu, E. J. Carter, I. Matthews, and J. K. Hodgins. Modeling and animating eye blinks. *ACM Transactions on Applied Perception (TAP)*, 8(3):1–17, 2011.

[56] L. C. Trutoiu, E. J. Carter, N. Pollard, J. F. Cohn, and J. K. Hodgins. Spatial and temporal linearities in posed and spontaneous smiles. *ACM Transactions on Applied Perception (TAP)*, 11(3):1–15, 2014.

[57] L. C. Trutoiu, J. K. Hodgins, and J. F. Cohn. The temporal connection between smiles and blinks. In *2013 10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, pp. 1–6. IEEE, 2013.

[58] M. Volonte, E. Ofek, K. Jakubzak, S. Bruner, and M. Gonzalez-Franco. Headbox: A facial blendshape animation toolkit for the microsoft rocketbox library. In *2022 IEEE Conference on Virtual Reality and 3D User Interfaces Abstracts and Workshops (VRW)*, pp. 39–42. IEEE, 2022.

[59] E. Whitmire, L. Trutoiu, R. Cavin, D. Perek, B. Scally, J. Phillips, and S. Patel. Eyecontact: scleral coil eye tracking for virtual reality. In *Proceedings of the 2016 ACM International Symposium on Wearable Computers*, pp. 184–191, 2016.

[60] R. Williams. *Culture and society, 1780-1950*. Columbia University Press, 1983.

[61] A. Winkler, J. Won, and Y. Ye. Questsim: Human motion tracking from sparse sensors with simulated avatars. In *SIGGRAPH Asia 2022 Conference Papers*, pp. 1–8, 2022.

[62] H. Yamamoto, M. Kawahara, M. Kret, and A. Tanaka. Cultural differences in emoticon perception: Japanese see the eyes and dutch the mouth of emoticons. *Letters on Evolutionary Behavioral Science*, 11(2):40–45, 2020.

[63] N. Yee and J. Bailenson. The Proteus Effect: The Effect of Transformed Self-Representation on Behavior. *Human Communication Research*, 33(3):271–290, July 2007. doi: 10.1111/j.1468-2958.2007.00299.x

[64] M. Yuki, W. W. Maddux, and T. Masuda. Are the windows to the soul the same in the east and west? cultural differences in using the eyes and mouth as cues to recognize emotions in japan and the united states. *Journal of Experimental Social Psychology*, 43(2):303–311, 2007.

[65] Zoom. Using Avatars in meetings and webinars, 2023.