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

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 Holoporation [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 “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 desynchronization 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 and executed in 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:

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

where $\Delta m_i$ 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 search recruitment firm. All participants provided informed cons-
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 a time, at different strengths. Both avatars always had the same...

---

**Table 1: Types of Noises and Ranges**

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

---

**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.**
When all types of noise were shown in combination, and before were shown simultaneously, participants significantly lowered their thresholds. Randomization for each block. 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.

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.

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

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Isolated</th>
<th>Combined</th>
<th>Social Cont.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Asymmetry</td>
<td>0.8mm</td>
<td>0.1mm</td>
<td>3.2mm</td>
</tr>
<tr>
<td>Brow Jitter</td>
<td>0.1mm</td>
<td>0.03mm</td>
<td>0.6mm</td>
</tr>
<tr>
<td>Closing Eye</td>
<td>2.7mm</td>
<td>0.01mm</td>
<td>12.5mm</td>
</tr>
<tr>
<td>Closing Mouth</td>
<td>1.7mm</td>
<td>0.01mm</td>
<td>3.4mm</td>
</tr>
<tr>
<td>Mouth Jitter</td>
<td>0.05mm</td>
<td>0.01mm</td>
<td>0.25mm</td>
</tr>
<tr>
<td>FPS</td>
<td>15</td>
<td>N/A</td>
<td>15 fps (90%)</td>
</tr>
</tbody>
</table>

(87.5%)

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
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.
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 acceptability 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 strong 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 understanding 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. Nevertheless, the results are robust within participants and as relative noise values.

7 Conclusion
A way to compare and contrast avatars is to explore whether out-validity settings. While low ecological validity tests may be less 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