Can we trust online crowdworkers?
Comparing online and offline participants in a preference test of virtual agents

Patrik Jonell*
KTH Royal Institute of Technology
pjjonell@kth.se

Taras Kucherenko*
KTH Royal Institute of Technology
tarask@kth.se

Ilaria Torre
KTH Royal Institute of Technology
ilariat@kth.se

Jonas Beskow
KTH Royal Institute of Technology
beskow@kth.se

ABSTRACT

Conducting user studies is a crucial component in many scientific fields. While some studies require participants to be physically present, other studies can be conducted both physically (e.g. in-lab) and online (e.g. via crowdsourcing). Inviting participants to the lab can be a time-consuming and logistically difficult endeavor, not to mention that sometimes research groups might not be able to run in-lab experiments, because of, for example, a pandemic. Crowdsourcing platforms such as Amazon Mechanical Turk (AMT) or Prolific can therefore be a suitable alternative to run certain experiments, such as evaluating virtual agents. Although previous studies investigated the use of crowdsourcing platforms for running experiments, there is still uncertainty as to whether the results are reliable for perceptual studies. Here we replicate a previous experiment where participants evaluated a gesture generation model for virtual agents. The experiment is conducted across three participant pools – in-lab, Prolific, and AMT – having similar demographics across the in-lab participants and the Prolific platform. Our results show no difference between the three participant pools in regards to their evaluations of the gesture generation models and their reliability scores. The results indicate that online platforms can successfully be used for perceptual evaluations of this kind.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in HCI**; *User studies*.

ACM Reference Format:

---

*Both authors contributed equally to this research.

---

1 INTRODUCTION

More and more perceptual studies in the Human-Computer Interaction field are done using online crowdsourcing platforms, such as Amazon Mechanical Turk[1] (AMT) and Prolific[2] [30, 33]. As there is no way to control the environment and experimental setting of the online workers, they can do other activities simultaneously or ignore instructions (such as wearing headphones). This can in turn lead to poor quality of study results. Checking attentiveness of online workers is common practice for perceptual studies [4, 10] and often leads to discarding a large number of participants [16, 20]. Passing attention checks does not always imply good concentration however, since online workers can simply learn to just pass the attention checks [25].

One way to investigate reliability of online participants is to compare them with in-lab (offline) participants. Offline participants are believed to be more attentive because they are in a controlled environment with fewer distractions and often with an experimenter present in the same room [28].

Prior work has been investigating differences between online and offline study participants [3, 7, 15, 18, 27], but most of them compared text-based survey research (questionnaire studies). In this work we focus on perceptual studies which differ from questionnaire studies in that they involve stimuli of other modalities than text. While it might be straightforward to prove attentiveness and reliability for text-based studies, the same is not always the case for perceptual studies, where there might not be a right or wrong answer to a question.

In this paper we replicate the study conducted by Kucherenko et al. [21] to investigate the differences in performing subjective perceptual studies between an in-lab setting and two crowdsourcing platforms. Specifically, we consider a preference test between two gesture generating models for a virtual agent where video artifacts have been produced for both models. The study was repeated three times: in-lab, using AMT, and on Prolific. In order to compare the participants in the three different pools, we evaluate the difference in the preference score given to the two gesture generation models,

---

[1]www.mturk.com
[2]www.prolific.co

the difference in inter- and intra-rater agreement, and the number of attention checks passed.

Our main research question is: *Do in-lab participants perform differently from participants on crowdsourcing platforms in a subjective audio-visual preference experiment?*

2 BACKGROUND AND RELATED WORK

In this section, we review previous work on analysing and comparing the quality of the data obtained from online workers with that from in-lab participants.

2.1 Improving quality of online studies

Improving the quality of the data obtained from online workers is an active field of research [1, 9, 26]. One way of detecting participants who might not be paying attention is to use instructional manipulation checks (IMCs). These were first introduced by Oppenheimer et al. [29], and are one of the most common ways to detect "cheaters" or inattentive participants. As the name suggests, IMCs are manipulations of the instructions which are used to detect if participants read the instructions carefully. An IMC could, for example, be an instruction which tells the participant to ignore a specific question, click "other" or write "I read the instruction" as an answer. Since IMCs were first introduced there have been other methods developed to detect inattentive participants or "dirty data" (reviewed in [8]), but using IMCs is still the standard technique.

Berinsky et al. [1] examined several strategies to enforce workers to be more attentive on different tasks. They found that some of them (especially training workers) produced a strong effect on the IMC passage rate, but that did not translate into higher-quality data. Apart from that, it has been shown that online workers are on the lookout for attention checks [25].

There are ongoing debates on whether attention checks should be used. Hauser et al. [14] argued that attention checks might distort the results, especially if they are very different from the original task. On the other hand, Kung et al. [22] experimentally showed that common attention checks do not affect scale validity in several classical experiments. In other words, previous research suggests that attention checks could be used, but with care [4].

Another way to ensure that participants provide high quality data is by screening them. This can be done by removing participants who are deemed to be unfit according to some criteria, such as by providing several wrong answers to questions with known answers [26] or by giving many identical answers in a row [6]. Several experiments indicated that screening can have an impact on statistical results [9, 26]. Screening therefore is commonly used for online studies, and can be done during or after the study.

2.2 Comparing online and offline participants

The type of data a study seeks to collect could also influence whether screening methodologies are more or less successful. For example, a qualitative study such as a market research, where participants have to create elaborate, free-text answers, might need different attention checks than an online questionnaire, where participants have to respond using multiple-choice answers.

Several researchers have been investigating differences between online and offline participants for questionnaire studies [7, 15, 18].

Hauser and Schwarz [15] used IMCs to test the attentiveness of participants when they read instructions before filling out questionnaires. For this study they used in-lab participants which were using their own computers and were not supervised. In three studies, AMT workers were consistently more likely to pass IMCs than the in-lab participants. Kees et al [18] did a similar study but found no differences between AMT and in-lab participants in terms of their performance in the tests.
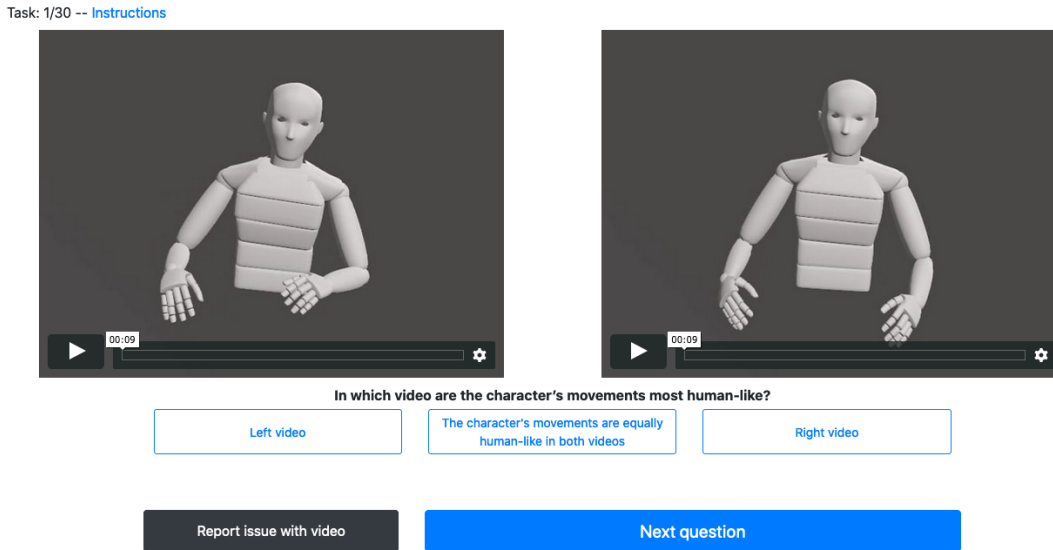
Several studies have indicated that online participants can reproduce in-lab results for different perceptual studies [3, 11, 19, 24, 27]. Lansford et al. [24] found similar results for the online and offline participants in terms of perceptual-training benefits while having different demographics. Germine et al. [11] showed that for challenging cognitive and perceptual experiments, online participants perform similarly to in-lab participants in different cognitive ability tests, even when those self-selected online participants are anonymous, uncompensated, and unsupervised. For a detailed review of perceptual studies, we refer the reader to Woods et al. [32].

One particularly interesting study is that of Burmenia et al. [2]. They conducted a perceptual study on emotion annotation in videos using AMT. They proposed and evaluated a novel filtering method which uses online quality assessment, stopping the evaluation when the quality of the worker drops below a threshold. They did, however, not compare in-lab with online participants.

The most similar study to ours is that of Byun et al. [3]. They compared in-lab experts with crowdworkers on AMT in a speech perception task. They had certain stimuli which were expected to yield a certain result and filtered out workers that did not score above chance at those tasks. The main difference to this work is the fact that we are using not only audio, but audio-visual stimuli with longer duration (10s). The current study also differs from Byun et al.'s in that we do not compare expert judgments, but rather layman judgments, and control for some of the demographic attributes when possible. Additionally we have a subjective task with no pre-defined "correct" answers.

In the field of non-verbal behavior generation for virtual agents, subjective evaluation is required to assess the quality of the models. Most of the modern methods in this field conduct subjective evaluations using online crowdsourcing platforms, such as AMT [16, 21, 33]. Many of them do screening based on attention checks. Yoon et al. [33] excluded participants who could not pass attention check questions or gave too vague answers to questions about subjective impressions, resulting in the exclusion of 28% of participants. Jonell et at. [16] discarded 43% of participants who did not pass the attention checks. Kucherenko et al. [21] used attention checks which were realized by distorting either the audio or the video. Participants who failed to report the majority of those samples as having an issue were discarded. Most of the participants (79%) did not finish the experiments as they either dropped out or failed a majority of the attention checks. Those results put in question the reliability of online workers for audio-visual perception studies.

The present study investigates if online participants are as reliable as in-lab participants. To the best of our knowledge, this is the first study which makes this investigation using virtual agents.

Can we trust online crowdworkers?
Comparing online and offline participants in a preference test of virtual agents

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

Figure 1: The user interface for evaluating the videos. Up in the left the participant is able to see how many trials are left. The two videos are played independently from each other. The participant has to chose one of the three alternatives as an answer in order to click "Next question" or click on the "report video as broken" button.

3 METHOD

The study used a mixed design with two independent variables: a) participant pool (between-subjects: in-lab, Prolific, AMT) and b) gesture generation model (within-subject: 'No PCA' and 'No text').

The main independent variable was the participant pool. The participant pools vary in how much control they provide, with the in-lab study granting a higher degree of control, but over a limited set of participants with limited demographics. On the other side of the spectrum we find AMT, which affords little control but a large amount of participants with a wide spread in demographics. There are also other services, such as Prolific, having fewer workers than AMT but providing more fine-grained control; for example, contrary to AMT, Prolific allows screening participants based on their language fluency.

The second independent variable is the gesture generation model used, which is described further in Subsection 3.2.

The two main dependent variables are:

(1) **Preference score**
whether participants indicate that one video is more human-like than the other, or that they are both equally human-like. Thus, this is a variable with 3 levels. From the preference score we also derive inter-rater intraclass correlation coefficients (ICC) and intra-rater ICC scores.

(2) **Number of attention checks passed**
There are two types of attention checks: audio-video based attention check (AV attention check), where the participant was instructed to mark a video as broken when they hear or see an instruction to do so; and same video attention check (SV attention check), where the exact same two videos were played, and the participant was expected to say that there was no difference between the two stimuli.

For exploratory analysis we also consider the following:

(1) **Time spent on each rating**
The time elapsed from the moment the two videos are shown and the moment the preference is indicated, in seconds.

(2) **Comment field length**
At the end of the experiment participants could leave a free-text comment about the experiment. The length of these comments (in characters) was a measure of their engagement with the experiment.

We hypothesize that:

- **H1**) Preference for the two models will be significantly different between the results obtained in-lab and the results obtained from Prolific and AMT.
- **H2**) In-lab participants will pass more attention checks during the experiment than online participants.
- **H3**) The inter-rater agreement, estimated using ICC, will be significantly higher in the results obtained in-lab than the results obtained from online workers.

This work was pre-registered using the OSF platform: osf.io/dxwak. There are however a few changes made with respect to the pre-registration. The main difference is that the measure for inter-rater agreement was changed to use ICC instead of Cohen's kappa. We also removed a fourth hypothesis (H4), since it became irrelevant.

3.1 Procedure

A web-based evaluation platform was implemented which was used across all three participant pools. There were some minor differences across each pool to accommodate some differences in how the recruiting was performed. The participants went through the following steps:

(1) In-lab: short description and field to input name,
    Prolific: Demographic input for age, gender, employment,
    and education
(2) Instructions
(3) Five training trials
(4) 21 trials as in Figure 1
(5) Demographic questions

Every participant first completed a training phase to familiarize themselves with the task and interface. This training consisted of five items with video segments not present in the study, showing the participants what kind of videos they may encounter during the study. Then each participant was asked to evaluate 21 same-speech video pairs: 15 pairs randomly sampled from a pool of 28 segments and 6 pairs that were intended as attention checks, described in Section 3.3. The videos were presented side by side and could be replayed separately as many times as desired. For each pair, participants indicated which video they thought best corresponded to the given question – *In which video are the character's movements most human-like?* – there was also an option to state that they perceived both videos to be equally human-like. An example of how a trial looked like is shown in Figure 1.

The video pairs for each participant were randomly sampled from a pool of videos, while the placement for the AV attention checks was counter-balanced. Also, the relative position of the videos within each pair (left or right) for each trial was randomized. Each participant was randomly assigned to a specific order of videos in the experiment. The same set of 24 fixed orders for trials were used among all the three experimental conditions. For all three conditions we recruited 24 participants to allow for counterbalancing of the order of placement for the AV attention checks.

3.2 Stimuli

The stimuli were generated by gesture generation models from Kucherenko et al. [21], which are neural networks trained to generate gestures based on speech using a dataset of human gesticulating (see the paper [21] for more details). For our experiments we have used the following two variations of the model: "No PCA" (which uses both acoustic and semantic features as an input and could hence generate complex gestures) and "No text" (which uses acoustic features only as an input and hence were mainly generating beat gestures). Please see samples at: vimeo.com/showcase/7571619.
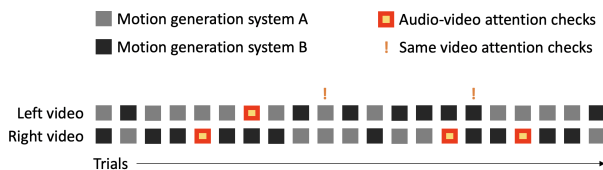
Figure 2: Illustration of attention check placement showing an experimental session for a participant. An experiment consisted of 21 trials, each showing a video on the left and on the right. The videos were showing an avatar generated using gesture generation system A or B or an AV attention check.

3.3 Attention checks

The attention checks were developed so that they would be similar to the actual task in order to prevent affecting the results [14]. For the four AV attention checks we picked four separate video pairs used only for attention check and added either a text or a synthesized voice telling the participant to report the video as broken (two of them had a text, and two had an auditory instruction). These were positioned in one out of four non-overlapping segments (spanning all of the 21 trials) by randomly choosing a place within that segment. The order of attention checks (such as "audio1", "video2", "video1", "audio2") was counterbalanced in a Latin Square fashion.

The two SV attention checks, which presented the exact same two videos (which were not used for the rest of the study) were placed at the 10th and the 16th trial-position for all experimental sessions. Here, an attentive rater should answer "no difference". Figure 2 illustrates an example of how attention checks were placed within an experiment.

3.4 Participants

Participants were recruited from three participant pools: in-lab, Prolific and AMT. In total 72 participants were recruited with 24 participants in each participant pool. Since Prolific allows for controlling for a wide range of participant characteristics, it was to a large extent possible to replicate certain demographics of the in-lab study participants (age, gender, education level, student status), see Table 1. Unfortunately, the same was not possible for AMT.

3.4.1 In-lab. In-lab participants were recruited through Facebook posts and the study was performed in person, always using the same laptop in the same room where one of the researchers was present during the whole experiment. Participants could adjust the volume during the training stage. They received minimal verbal instructions (asked to sit down, put on headphones and follow the on-screen instructions). All the participants were residing in Stockholm, Sweden. No exclusion criteria was used, but the participants were told that they would not get the reward if they failed too many attention checks. This was just used to motivate them to pay attention in the study, in reality everyone would have received their reward; in any case, none of the participants failed too many attention checks. The reward for participation was a movie ticket voucher (average price of a movie ticket in Stockholm is 9.4 USD).

3.4.2 Prolific. As the second participant pool, we recruited participants using a crowd-sourcing platform called Prolific based on the same basic demographics as the in-lab study. We requested participants with the same four demographic characteristics as in-lab participants: age, gender, education level and employment status. Payment was 3.90 + 5.5 USD (5.5 USD was given as a bonus upon completion if the participant passed the majority of the audio-video attention checks). When the Prolific participants started our experiment they were asked to fill out a form with the four demographic questions mentioned above. The participants did not know which criteria had given them access to the experiment; and in case they did not match with the criteria provided in their profile, they were not allowed to partake in the experiment.

Can we trust online crowdworkers?
Comparing online and offline participants in a preference test of virtual agents

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

	AMT	In-lab	Prolific
Number of participants	24	24	24
Age (avg±std)	41 ± 13	28 ± 6	28 ± 6
Gender (f/m/o)	12/11/1	10/13/1	10/13/1
Exp. with technology (1-5) (avg±std)	3.8 ± 1.0	4.3 ± 0.9	4.0 ± 0.8
Number of past similar studies (avg±std)	0.6 ± 2.0	0.8 ± 1.4	0.3 ± 0.6
Number of different nationalities	3	16	12
Having a higher education[1]	58%	96%	96%
Currently students	0%	66%	46%

Table 1: Demographics of the participants broken down over the three participant pools.
[1]Higher education is defined as having completed a bachelor's degree or higher

3.4.3 Amazon Mechanical Turk. As the third pool, participants were recruited through the crowd-sourcing platform AMT. Payment was 3.90 + 5.5 USD (5.5 USD was given as a bonus upon completion if the participant passed the majority of the audio-video attention checks). Requirements to partake in our experiment was to have finished at least 10,000 previous HITs, have an approval rate of at least 98% and to be located in the United States.

AMT provides a limited way of controlling demographics, thus it was not possible to adequately replicate the demographics of the participants from the other pools.

4 RESULTS

The results from the experiment are presented below. The data was analysed using R version 3.6.3. The analysis was performed in a double-blind fashion, such that the authors had obfuscated the participant pool and preference score variables, and revealed them after all analyses were done. Before the comparison of preference score and comparison of attention checks were done, a pre-analysis was conducted in order to determine whether these measures were correlated or not. The outcome of the pre-analysis would determine whether they would be analyzed together (in case they were correlated) or separately (in case they were not correlated).

Participants could mark any video as having issues during the experiment. All trials with videos that were marked as having issues were excluded from the analyses. In case of missing data, we removed the trial (row) from the participant, but still used the rest of the data from the participant. Therefore, the final number of trials analysed was 359 for Prolific, 360 for AMT, and 358 for in-lab.

The preference score can have one of 3 possible values: -1 (preference for the 'No text' Model), 0 (equal preference) and 1 (preference for the 'No PCA' Model). Any references to "preference score" refers to this coding of the data.

4.1 Pre-analysis

In order to determine whether to analyze the two dependent variables (DVs) together or separately, we calculated the correlations between the two DVs (average number of AV attention checks passed and preference strength) using the Pearson correlation coefficient. The preference strength was defined as: for each participant, rows with ties (value 0) were removed, so that only values of -1 and 1 remained. Afterwards we calculated the absolute value of the average preference score: $abs(average(preference\_score))$, which we call *preference strength*, which is a continuous variable between 0 and 1.
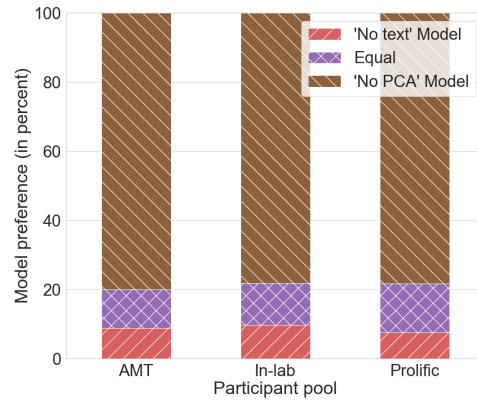

Figure 3: Participants' preferences toward the two different gesture generation models per participant pool.

This score indicates how strong the opinion of a user is: the closer the preference is to one, the stronger the preference is to any of the two gesture generation conditions. The cut-off for the correlation coefficient $r = 0.3$ was determined following standards in Psychology [5]. It was found that for all the conditions the correlation coefficient r was lower than the cut-off value. Hence we regarded the two variables as not correlated and analyzed them separately. The preference strength was only used for the pre-analysis, for all the comparisons described below the preference score was used.

4.2 Comparison of preference score

We fitted a cumulative link mixed model via likelihood ratio test with preference score as dependent variable, participant pool as predictor, and rater and sample (refers to a unique pair of videos) as random intercepts; however, we found no significant effect of participant pool ($\chi^2(2) = 0.54, p = 0.77$). A post-hoc g*power analysis was made, resulting in a power of 0.95.

As a separate test, we compared inter-rater and intra-rater reliability. An analysis using bootstrapped ICC was used for the inter-rater and intra-rater agreement using the "agreement" R-package [12] and the dimensional ICC method using Model 2A [13]. Analysis showed that the confidence intervals are overlapping (as seen in Figure 4) and that there was no statistically significant difference between the three participant pools in either of the cases [31].
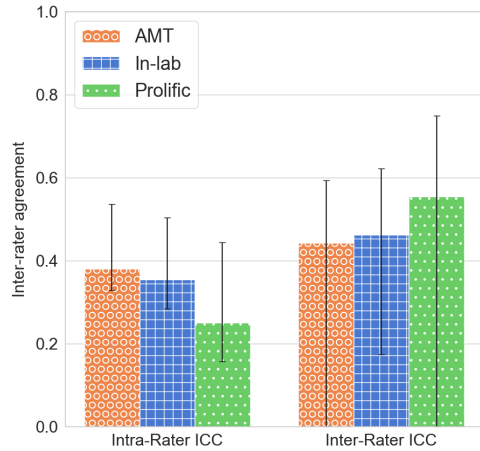
Figure 4: Inter-rater and Intra-rater ICC for each model separately broken down per participant pool. Higher values corresponds with higher agreement between the raters. Error bars show 95% confidence intervals.

4.3 Comparison of passed attention checks

4.3.1 Audio-video attention checks. The AV attention checks were coded as 0 (failed) or 1 (passed). Then for each participant, we calculated the average attention score, by summing all the 0 and 1 scores and dividing them by the number of attention checks. This was a value between 0 (high failure rate) and 1 (low failure rate). Most of the participants (69 out of 72) passed all audio-video attention checks during the experiment and thus there was no difference between the participant pools in terms of passing AV attention checks. The three participants that did not pass all of the AV attention checks failed only on the video-based attention checks and belonged to the in-lab participant pool. Therefore we concentrated our analyses on the SV attention checks.

4.3.2 Same video attention checks. The SV attention checks were coded as 0 (failed) or 1 (passed). Then for each participant, we calculated the average of passed attention check as in Section 4.3. Compared to the AV attention checks, a higher degree of participants (32 out of 72) failed either one or both of the SV attention checks. The results contained a large number of zero-values, we therefore fit a zero-inflated regression model on the same SV attention checks via maximum likelihood estimation ($\chi^2(2) = 0.17, p = 0.91$) and observe that there is no statistically significant difference between the participant pools. These results are visualized in Figure 5.

4.4 Trial duration analysis

As an exploratory analysis we considered the difference between the participant pools in terms of the duration to complete each trial, since it can be used to measure attentiveness [8]. We performed the analysis by fitting a linear mixed-effects model using the lmerTest package [23], with task duration as dependent variable, participant pool as predictor, and participant and sample as random intercepts.
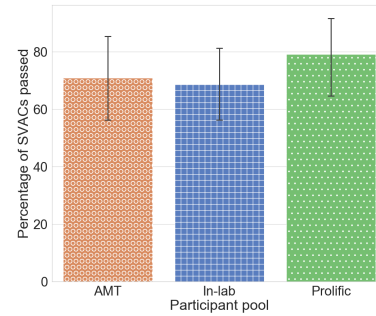
Figure 5: Percentage of passed SV attention check broken down per participant pool. A higher number indicates higher attention. Error bars show 95% confidence intervals.
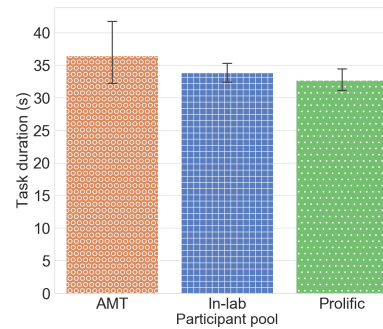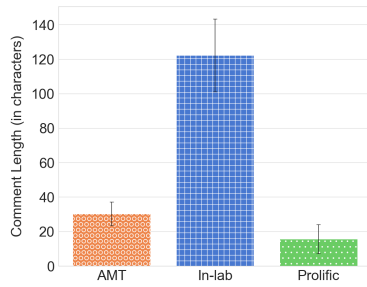

Figure 6: Average trial duration in seconds broken down per participant pool. Error bars show 95% confidence intervals.

We found no statistical significance between the participant pools ($\chi^2(2) = 1.23, p = 0.54$), meaning that participants spent approximately the same amount of time on the trials in each participant pool. The AMT participant pool showed a higher variance than the two other participant pools as can be seen by the larger confidence interval in Figure 6.

4.5 Comment field analysis

As an additional exploratory analysis, we also looked at the length (in characters) of the optional comment at the end of the experiments as it might reveal information on how engaged participants in the different pools were.

Since several participants left the field empty, thus resulting in a value of 0, we fit a zero-inflated regression model on the comment length via maximum likelihood estimation, with participant pool as predictor. This model was significant, suggesting that participants' comments differed in the three participant pools ($\chi^2(2) = 800.99, p < .001$). As a post-hoc analysis we calculated the estimated marginal means (EMMs) using Tukey correction for multiple comparisons. The results are illustrated in Figure 7. We found that in-lab participants wrote significantly longer comments than both AMT participants ($z = 4.120, p < .001$) and Prolific

Can we trust online crowdworkers?
Comparing online and offline participants in a preference test of virtual agents

IVA '20, October 19–23, 2020, Virtual Event, Scotland Uk

Figure 7: Average comment length broken down per participant pool. Error bars show 95% confidence intervals.

($z = 4.671, p < .001$), while there was no statistically significant difference between AMT and Prolific ($z = 1.364, p = 0.36$).

5 DISCUSSION

We compared three participant pools; in-lab, Prolific, and AMT in terms of their preference scores, inter/intra-rater agreement, and attentiveness when comparing two gesture generation models.

From the results we see that there is no difference in neither of the three measures across the three participant pools, thus we reject the three hypothesis we set out from the beginning; H1, H2, and H3. Our result is consistent with previous findings on other perceptual tasks [11, 24, 32]. The preference scores obtained across the three participant pools were also consistent with the experiment performed in the work of Kucherenko et al. [21]. We therefore conclude that reliable results can be obtained from both in-lab participants and online workers.

Comparing the preference score alone might not give a complete picture of the differences between the two groups and thus we also investigated the inter- and intra-rater agreement. The Inter-rater agreement gives a measure of how consistent an individual participant is with the other participants, while the intra-rater agreement gives a measure of how consistent each worker are with themselves. We compared these measures as well, and found that participants are equally consistent in both of these regards over all of the three participant pools. This supports our conclusion that the participants in all conditions were equally reliable.

In terms of the number of passed attention checks there was however a large difference between our experiment and that of Kucherenko et al [21]. In the previous study a big portion of the AMT workers (over 75%) did not finish the study, either due to timing out or failing a majority (more or equal to two) of the AV attention checks, while in the present study participants from the AMT pool (as well as those from the other pools) never failed more than one AV attention check, and no one timed out. There were two major differences between Kucherenko et al.'s study and the present one; how the AV attention checks were designed and the size of the reimbursement. The AV attention checks in the current work explicitly either displayed a text message or spoke using a synthetic voice instructing the user to mark the trial as broken. In Kucherenko et al.'s work the AV attention checks were not explicit, instead the quality of the audio or video was degraded to such a level

where they were unusable, and participants were asked to report any broken videos that prevented them from making a judgement. These implicit AV attention checks probably led to more participants failing them. They might even have been a cause of frustration for participants, thus decreasing their intrinsic motivation to take part in the task. The second main difference between this experiment and that of Kucherenko et al. lies in the monetary reward, which was considerably higher in the present study (28.2 USD/h vs 9 USD/h). This seems to suggest that the reimbursement level can have a strong effect on how attentive the participants were and that when having high reward AV attention checks might not be necessary, as participants might be more motivated to perform well. Giving a more appealing monetary reward might have increased participants' extrinsic motivation to complete the task. Both intrinsic and extrinsic motivation are important for collecting high quality data [17].

Another point of interest is that the in-lab participants provided longer comments in the optional comment field of the experiment, which might suggest that in-lab participants put more effort and commitment into the task. Participants did not differ significantly in terms of time spent on each trial, however the AMT did show a considerably higher variance. This is interesting and warrants further investigation.

5.1 Limitations

The results from our study seem promising for researchers who want to use crowd-sourcing as a method for evaluating stimuli and performing perceptual experiments. However, there are two main points which the authors would like to highlight as limitations to the current experiment.

In this study we aimed at reimbursing the participants in an as comparable way as possible. Due to restrictions by the University it was not possible to reimburse in-lab participants monetarily. Therefore each participant was given a cinema ticket voucher upon successful completion of the study. The participants on Prolific and AMT were given the monetary value of the cinema ticket (approx. 9 USD). The reimbursement of 9 USD for 20min task is relatively high for Prolific and AMT and, as previously discussed, could be an important factor when comparing with previous work [17].

The reimbursement on both AMT and Prolific was divided in two steps, one for finishing the study, and a second part for finishing the study "successfully". The wording was intentionally ambiguous to not give away that there were attention checks, and all participants who failed several AV attention checks would not have been reimbursed. The two levels were set up as we wanted to reimburse even "cheating" participants since we intended to use their data, but we also did not want to pay "cheaters" the full amount. The in-lab participants were informed that they would not receive the reward if they did not finish the experiment successfully (however, they would still receive it even if they failed). These differences could potentially be confounds.

5.2 Future Work

Many previous experiments have used lower reimbursement levels compared to the present study, and has reported high failure rates for attention checks [16, 21, 33]. In the present study none of the

participants failed more than one AV attention check, which seems to indicate that reimbursement might be an important factor. As a next step we plan to analyze the effect on the results in a similar audio-visual perception test by varying the reimbursement levels.

Furthermore it would be interesting to investigate the influence of different types of attention checks that could be used in this type of perceptual experiments and how they would affect the results.

Another interesting direction for future work would be to conduct the same experiment with *experts*, i.e. people who work professionally in the field of gesture analysis, in order to validate the finding based on "correct" answers from experts.

6 CONCLUSION

This paper presented an experiment where a comparison was made between three participant pools; in-lab, Prolific, and Amazon Mechanical Turk. The experiment was a subjective preference test of videos generated by two gesture generation models. The results showed that there was no significant difference in several measures (preference score, attention checks passed and inter/intra-rater agreement) between the three participant pools. These results indicate that online workers can successfully be used instead of in-lab participants for audio-visual perception experiments similar to the one outlined in this paper, significantly easing the task of recruiting participants. The results have to be interpreted with some caution however, as the effect of reimbursement level is not fully understood, and would need further investigation.

ACKNOWLEDGEMENT

REFERENCES

[1] Adam J Berinsky, Michele F Margolis, and Michael W Sances. 2016. Can we turn shirkers into workers? *Journal of Experimental Social Psychology* 66 (2016).
[2] Alec Burmania, Srinivas Parthasarathy, and Carlos Busso. 2015. Increasing the reliability of crowdsourcing evaluations using online quality assessment. *IEEE Transactions on Affective Computing* 7, 4 (2015), 374–388.
[3] Tara McAllister Byun, Peter F Halpin, and Daniel Szeredi. 2015. Online crowdsourcing for efficient rating of speech: A validation study. *Journal of communication disorders* 53 (2015), 70–83.
[4] Janelle H Cheung, Deanna K Burns, Robert R Sinclair, and Michael Sliter. 2017. Amazon Mechanical Turk in organizational psychology: An evaluation and practical recommendations. *Journal of Business and Psychology* 32, 4 (2017), 347–361.
[5] Jacob Cohen. 2013. *Statistical power analysis for the behavioral sciences.* Academic press.
[6] Paul T Costa Jr and Robert R McCrae. 2008. *The Revised NEO Personality Inventory (NEO-PI-R).* Sage Publications, Inc.
[7] Matthew JC Crump, John V McDonnell, and Todd M Gureckis. 2013. Evaluating Amazon's Mechanical Turk as a tool for experimental behavioral research. *PloS one* 8, 3 (2013), e57410.
[8] Paul G Curran. 2016. Methods for the detection of carelessly invalid responses in survey data. *Journal of Experimental Social Psychology* 66 (2016), 4–19.
[9] Justin A DeSimone and PD Harms. 2018. Dirty data: The effects of screening respondents who provide low-quality data in survey research. *Journal of Business and Psychology* 33, 5 (2018), 559–577.
[10] Avi Fleischer, Alan D Mead, and Jialin Huang. 2015. Inattentive responding in MTurk and other online samples. *Industrial and Organizational Psychology* 8, 2 (2015), 196–202.
[11] Laura Germine, Ken Nakayama, Bradley C Duchaine, Christopher F Chabris, Garga Chatterjee, and Jeremy B Wilmer. 2012. Is the Web as good as the lab? Comparable performance from Web and lab in cognitive/perceptual experiments. *Psychonomic bulletin & review* 19, 5 (2012), 847–857.
[12] Jeffrey Girard. 2020. *agreement.* R package version 0.0.0.9002.
[13] Kilem L Gwet. 2014. *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters.* Advanced Analytics, LLC.
[14] David J Hauser, Phoebe C Ellsworth, and Richard Gonzalez. 2018. Are manipulation checks necessary? *Frontiers in psychology* 9 (2018).
[15] David J Hauser and Norbert Schwarz. 2016. Attentive Turkers: MTurk participants perform better on online attention checks than do subject pool participants. *Behavior research methods* 48, 1 (2016), 400–407.
[16] Patrik Jonell, Taras Kucherenko, Erik Ekstedt, and Jonas Beskow. 2019. Learning Non-verbal Behavior for a Social Robot from YouTube Videos. In *ICDL-EpiRob Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions, Oslo, Norway, August 19, 2019.*
[17] Nicolas Kaufmann, Thimo Schulze, and Daniel Veit. 2011. More than fun and money. Worker Motivation in Crowdsourcing-A Study on Mechanical Turk.. In *AMCIS*, Vol. 11. Detroit, Michigan, USA, 1–11.
[18] Jeremy Kees, Christopher Berry, Scot Burton, and Kim Sheehan. 2017. An analysis of data quality: Professional panels, student subject pools, and Amazon's Mechanical Turk. *Journal of Advertising* 46, 1 (2017), 141–155.
[19] Steven Komarov, Katharina Reinecke, and Krzysztof Z. Gajos. 2013. Crowdsourcing Performance Evaluations of User Interfaces. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Paris, France) *(CHI '13).* ACM, New York, NY, USA, 207–216. https://doi.org/10.1145/2470654.2470684
[20] Taras Kucherenko, Dai Hasegawa, Gustav Eje Henter, Naoshi Kaneko, and Hedvig Kjellström. 2019. Analyzing Input and Output Representations for Speech-Driven Gesture Generation. In *International Conference on Intelligent Virtual Agents (IVA '19),* Vol. 19. ACM, Paris, France, 97–104. https://doi.org/10.1145/3308532.3329472
[21] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexanderson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the ACM International Conference on Multimodal Interaction.*
[22] Franki YH Kung, Navio Kwok, and Douglas J Brown. 2018. Are attention check questions a threat to scale validity? *Applied Psychology* 67, 2 (2018), 264–283.
[23] Alexandra Kuznetsova, Per B Brockhoff, and Rune Haubo Bojesen Christensen. 2017. lmerTest package: tests in linear mixed effects models. *Journal of statistical software* 82, 13 (2017).
[24] Kaitlin L Lansford, Stephanie A Borrie, and Lukas Bystricky. 2016. Use of crowdsourcing to assess the ecological validity of perceptual-training paradigms in dysarthria. *American Journal of Speech-Language Pathology* 25, 2 (2016), 233–239.
[25] Matt Lovett, Saleh Bajaba, Myra Lovett, and Marcia J Simmering. 2018. Data quality from crowdsourced surveys: A mixed method inquiry into perceptions of amazon's mechanical turk masters. *Applied Psychology* 67, 2 (2018), 339–366.
[26] Michael R Maniaci and Ronald D Rogge. 2014. Caring about carelessness: Participant inattention and its effects on research. *Journal of Research in Personality* (2014).
[27] Róisín McNaney, Mohammad Othman, Dan Richardson, Paul Dunphy, Telmo Amaral, Nick Miller, Helen Stringer, Patrick Olivier, and John Vines. 2016. Speeching: mobile crowdsourced speech assessment to support self-monitoring and management for people with Parkinson's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM, 4464–4476.
[28] Babak Naderi, Ina Wechsung, and Sebastian Möller. 2015. Effect of being observed on the reliability of responses in crowdsourcing micro-task platforms. In *2015 Seventh International Workshop on Quality of Multimedia Experience (QoMEX).*
[29] Daniel M Oppenheimer, Tom Meyvis, and Nicolas Davidenko. 2009. Instructional manipulation checks: Detecting satisficing to increase statistical power. *Journal of experimental social psychology* 45, 4 (2009), 867–872.
[30] Najmeh Sadoughi and Carlos Busso. 2019. Speech-driven animation with meaningful behaviors. *Speech Communication* 110 (2019), 90–100.
[31] Margarita Stolarova, Corinna Wolf, Tanja Rinker, and Aenne Brielmann. 2014. How to assess and compare inter-rater reliability, agreement and correlation of ratings: an exemplary analysis of mother-father and parent-teacher expressive vocabulary rating pairs. *Frontiers in psychology* 5 (2014), 509.
[32] Andy T Woods, Carlos Velasco, Carmel A Levitan, Xiaoang Wan, and Charles Spence. 2015. Conducting perception research over the internet: a tutorial review. *PeerJ* 3 (2015), e1058.
[33] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2019. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In *International Conference on Robotics and Automation (ICRA '19).* IEEE.