To Rate or Not To Rate: Investigating Evaluation Methods for Generated Co-Speech Gestures

Pieter Wolfert*
IDLab, Ghent University - imec
Ghent, Belgium
pieter.wolfert@ugent.be

Taras Kucherenko
KTH Royal Institute of Technology
Stockholm, Sweden
tarask@kth.se

Jeffrey M. Girard
Department of Psychology, University of Kansas
Kansas, USA
jmgirard@ku.edu

Tony Belpaeme
IDLab, Ghent University - imec
Ghent, Belgium
tony.belpaeme@ugent.be

ABSTRACT

While automatic performance metrics are crucial for machine learning of artificial human-like behaviour, the gold standard for evaluation remains human judgement. The subjective evaluation of artificial human-like behaviour in embodied conversational agents is however expensive and little is known about the quality of the data it returns. Two approaches to subjective evaluation can be largely distinguished, one relying on ratings, the other on pairwise comparisons. In this study we use co-speech gestures to compare the two against each other and answer questions about their appropriateness for evaluation of artificial behaviour. We consider their ability to rate quality, but also aspects pertaining to the effort of use and the time required to collect subjective data. We use crowd sourcing to rate the quality of co-speech gestures in avatars, assessing which method picks up more detail in subjective assessments. We compared gestures generated by three different machine learning models with various level of behavioural quality. We found that both approaches were able to rank the videos according to quality and that the ranking significantly correlated, showing that in terms of quality there is no preference of one method over the other. We also found that pairwise comparisons were slightly faster and came with improved inter-rater reliability, suggesting that for small-scale studies pairwise comparisons are to be favoured over ratings.

CCS CONCEPTS

• **Human-centered computing** → **HCI design and evaluation methods**; **Human computer interaction (HCI)**.

KEYWORDS

virtual agents, user study, nonverbal behaviour, evaluation methodology

---

*Corresponding Author

1 INTRODUCTION

When we interact with embodied conversational agents, we expect a similar manner of nonverbal communication as when interacting with humans. One way to achieve more human-like nonverbal behaviour in conversational agents is through the use of data-driven methods, which learn model parameters from data and gained in popularity over the past few years [1, 17, 18, 40]. Data-driven methods have been used to generate lips synchronisation, eye gaze or facial expressions, however in this work we take co-speech gestures as a test bed for comparing evaluation methods. Data-driven methods are able to generate a wider range of gestures and behaviours, as behaviour is no longer restricted to pre-coded animation or procedurally generated behaviour, but instead are generated from models trained on large amounts of data of human movement. These behaviours are often used to drive conversational agents in both virtual and physical agents, as these improve interaction [3, 11, 21, 28, 30]. Co-speech gestures are traditionally divided into four dimensions: iconic gestures, beat gestures, deictic gestures, and metaphoric gestures [24]. The approach to produce each of these categories often differs but, using data-driven methods, it becomes possible to generate multiple categories of gestures with a single model.

The quality of generated human-like behaviour can be assessed using objective or subjective measures. *Objective measures* rely on an algorithmic approach to return a quantitative measure of the quality of the behaviour and are entirely automated, while *subjective measures* instead rely on ratings by human observers. Most recent papers on co-speech gesture generation report objective measures to assess the quality of the generated behaviour, with measures such as velocity diagrams or average jerk being popular [1, 16, 40]. These measures not only are easy to automate, but also allow comparisons across models. For example, Yoon et al. [40] trained models from other authors on the same dataset to compare objective metric results. For this reason, objective measures are

often preferred over subjective evaluations, as the latter are harder to compare due to their potentially high variability. Yet, subjective evaluations are crucial when evaluating the behaviour of agents interacting with humans. This is because social communication is much more complex than current objective measures are capable of capturing, and subjective evaluations are still considered to be the gold standard. There may also be a large subjective component to how observers interpret generated behaviour, which we would like to capture. Thus, the "final stretch" in quality evaluations still relies heavily on subjective evaluations [34].

While the value of subjective evaluations is widely accepted, there is little consensus on how to collect and analyse such evaluations in relation to the evaluation of data-driven generated stimuli. Recently, several authors working on data-driven methods for non-verbal behaviour generation moved from rating scales (e.g., having observers rate how human-like two generated stimuli are from 1-to-5) to the use of pairwise comparisons (e.g., having observers select which of two generated stimuli is more human-like) [17, 26, 36, 40]. For example, Yoon et al. [40] argued that *"co-speech gestures are so subtle, so participants would have struggled to rate them on a five- or seven-point scale"* and promoted the use of pairwise comparisons over rating scales. Relatively little empirical attention has been devoted to this methodological topic in regard to the evaluation of data-driven generated stimuli, however, and it is still unknown how much the methods actually differ in terms of usability and informativeness.

In the current study, we seek to explore the similarities and differences between the rating scale and pairwise comparison approaches. We take generated co-speech gestures as a test bed for our evaluations but note that these findings may also apply to stimulus evaluation in other areas. Our hypotheses, design, and methodology were pre-registered before the data was gathered[1]. We present short video clips to human participants, with each video clip showing an avatar displaying combined verbal and nonverbal behaviour. The movements are generated using three data-driven methods of varying quality and we expect the subjective evaluations to clearly reflect this difference. In order to gain more insight into the effectiveness of the two subjective evaluation methods, we formulated the following five hypotheses.

H1. The rank-order of stimuli implied by the pairwise comparisons and rating scales will be different.

H2. Pairwise comparisons will have higher inter-rater agreement than rating scales.

H3. Pairwise comparisons and rating scales will differ in terms of time-efficiency (e.g., the time it takes for a single participant to finish a single evaluation).

H4. Pairwise comparisons and rating scales will differ in terms of participant usage preference and usability (both qualitative and quantitative).

H5. Pairwise comparisons and rating scales will both find a difference between stimuli that have a pronounced quality difference, but will not have enough resolution to find a difference between stimuli that differ slightly in quality.

---

[1]https://osf.io/7d9fs

Our aim is to quantify the pros and cons of these subjective evaluation methods and to provide empirical recommendations for the community working on gesture generation on when to use each method. Although the concept of comparing these evaluation strategies is not novel on its own [8], it is novel in relation to the evaluation of gesture generation for ECAs. We hope that this work can both highlight similarities and differences between the evaluated methods, and function as a bridge between the different fields of psycho-metrics and gesture generation researchers.

2 RELATED WORK

In this section, we cover work that compared rating scale evaluations with pairwise comparisons, and look at their specific use in the field of gesture generation for embodied conversational agents. To our knowledge, there has not been a comparison between rating scales and pairwise comparisons for the evaluation of co-speech gestures in Embodied Conversational Agents (ECAs) in particular. However, subjective evaluation methods have been studied and compared in several other fields. We want to highlight that we believe that there is a difference in the type of stimuli that are evaluated: we consider data-driven behaviour in virtual agents. We are aware of the overlap there is with psychology and psychometrics, but want to zoom in specifically on the use of both rating and pairwise assessments of generated gesticular behaviour for ECAs.

There is a rich history of work in psychology on related topics. DeCoster et al. [5] compared analysing continuous variables directly with analysing them after dichotomisation (e.g., re-coding them as two-class variables such as high-or-low). Although there were a few edge cases where dichotomisation was similar to direct analysis, they demonstrated that dichotomisation throws away important information and concluded that the use of the original continuous variables is to be preferred in most circumstances. Simms et al. [32] randomly assigned participants to complete the same personality rating scales with different numbers of response options ranging from two to eleven. They found that including four or fewer response options often attenuates psychometric precision, and including more than six response options generally provides no improvements in precision. Finally, Rhemtulla et al. [29] demonstrated that treating rating scale data as continuous can be problematic (i.e., can result in biased estimates) for scales with fewer than five response options, which tend to be quite non-normally distributed. Such data thus requires specialised ordinal methods to analyse properly. Overall, the psychological literature thus suggests that rating scales with between five and seven response options would be preferable to rating scales with fewer response options. If we consider the pairwise comparison approach to be similar to a rating scale with two response options (e.g., better or worse), this would raise concerns about the approach's psychometric precision and normality.

Although not covered in this paper, another way of evaluating stimuli on a continuous scale is by using visually-aided rating (VAR) [12]. Visually, categories are still used as anchors in VAR, but specific scores are not visualised in comparison to Likert scales. This enables participants to quantify an ordering, from which it is still possible to derive a quantifiable rating. VAS-RRP is congruent

to VAR, except that in VAR the rating scale is placed vertically, and in VAS-RRP horizontally [33].

However, there have also been impassioned arguments in favour of ordinal and rank-based approaches (of which the pairwise comparison approach can be considered a simple variant) within the affective computing community in recent years [22, 38, 39]. The argument is that many subjective evaluations are inherently ordinal and cannot be adequately treated as continuous numbers or nominal categories and should instead be handled using rankings. If this argument is accurate, then the pairwise comparison approach would be preferable to the rating scale approach on theoretical grounds. There is also evidence that rank-based approaches might have some practical benefits over rating scale approaches, such as being faster to administer and more reliable over time. For example, Clark et al. [4] evaluated the perception of physical strength from images of male bodies using both pairwise comparisons and rating scales and found that the scores were closely correlated but that the pairwise comparisons were completed 67% faster. Other examples, like Elliot et al. [7] and Mueser et al.[25] found high correlations between rankings resulting from the evaluation of physical features in humans. Liang et al. [20] proposes a model to 'calibrate' self-reported user ratings for dialogue systems due to issues with validity and bias. In relation to biomedical image assessments, where evaluation considers the visual quality of the stimuli, Phelps et al. [27] found that pairwise comparisons and ranked Likert scores made for more accurate assessments in comparison to the use of non-ranked Likert scores. Burton et al. [2] compared rating scales with best-worst scaling, another variant of the rank-based approach. In this study, participants were asked to select the most attractive and least attractive faces in a series of images. The best-worst scaling approach showed better test-retest reliability than the rating scale approach.

One of the reasons the community would benefit from greater standardisation of subjective evaluations methods can be found in the recently organised GENEA gesture-generation challenge [18]. Invited researchers were asked to submit models trained on the same dataset containing human speech and co-speech gestures. All submissions were then compared using crowd-sourced subject evaluation, in which online participants were asked to rate each clip with a score between 1 and 100. The benefit of this method is that generation models from different authors can be tested at once, within the same framework and participant pool. Sticking to a single evaluation strategy makes it possible to compare work across models, such as in the GENEA challenge, and also across time.

Finally, a recently published preprint reviewed the literature on evaluation of gesture-generation systems [37]. The review found that stimuli were often evaluated using very different rating scales, such as likeability, naturalness, and gesture-timing. This variability makes comparisons difficult across papers and time.

3 METHODS

3.1 Experimental Design

In this study, we used 30 video stimuli[2] showing a gesticulating avatar provided by Kucherenko et al. [17], the stimuli are already available and have been used by other researchers [13]. The videos had a resolution of 640×480 pixels and a frame rate of 30 frames per second. Three types of videos were used: Full, NoSpeech and NoText. The *Full* videos were generated by a model trained on motion of a human actor with the model having access to both the audio speech and transcribed text; the *NoSpeech* videos were generated from a model only trained on motion and transcribed text; and the *NoText* videos were generated by a model trained on motion and speech audio only. Thirty videos were created per type and, in each triplet of videos (across type), the avatar spoke the same sentence to facilitate comparison. We have two study conditions: Full versus NoSpeech (which we denote Low Difference) and Full versus NoText (High Difference). We denote them this way because the former showed a small difference in the original study [17], while the latter showed a large difference. These conditions (Full. vs NoSpeech and Full. vs NoText) turned out to show significant differences in quality, and we assume that our subjective evaluations will reflect this.

Each participant in the current study was assigned to either the LowDiff or HighDiff condition. Following that, the participant was assigned to one of two ordering conditions:

PR: Pairwise Comparison approach for 10 videos drawn from a set of 30 videos, followed by the Rating Scale approach for the same 10 videos.

RP: Rating Scale approach for 10 videos drawn from a set of 30, and then Pairwise Comparison approach for the same 10 videos.

3.2 Participants

For this study, 130 participants were recruited on Prolific[3]. To ensure data quality, participants have to be a native speaker of English, have at least a 90% approval rating on the platform, and have participated in at least 100 other studies on the platform. Participants were assigned to conditions using block randomisation in order to maintain balanced conditions.

3.3 Technical Setup

From Prolific, participants were forwarded to a web application to evaluate the stimuli. This application was based on HEMVIP [14], which in turn was based on WebMushra [31] but adapted to work with video files. Since two evaluation strategies were evaluated, there were two interface versions.

The pairwise comparison interface (Figure 1) displays two videos side by side, with three options for evaluation displayed below the videos. For all conditions, the question was: 'In which video are the character's movements most human-like?' The three response options were: *left*, *right*, and *equal*. Participants were able to play both videos at the same time, but it is not explicitly mentioned in
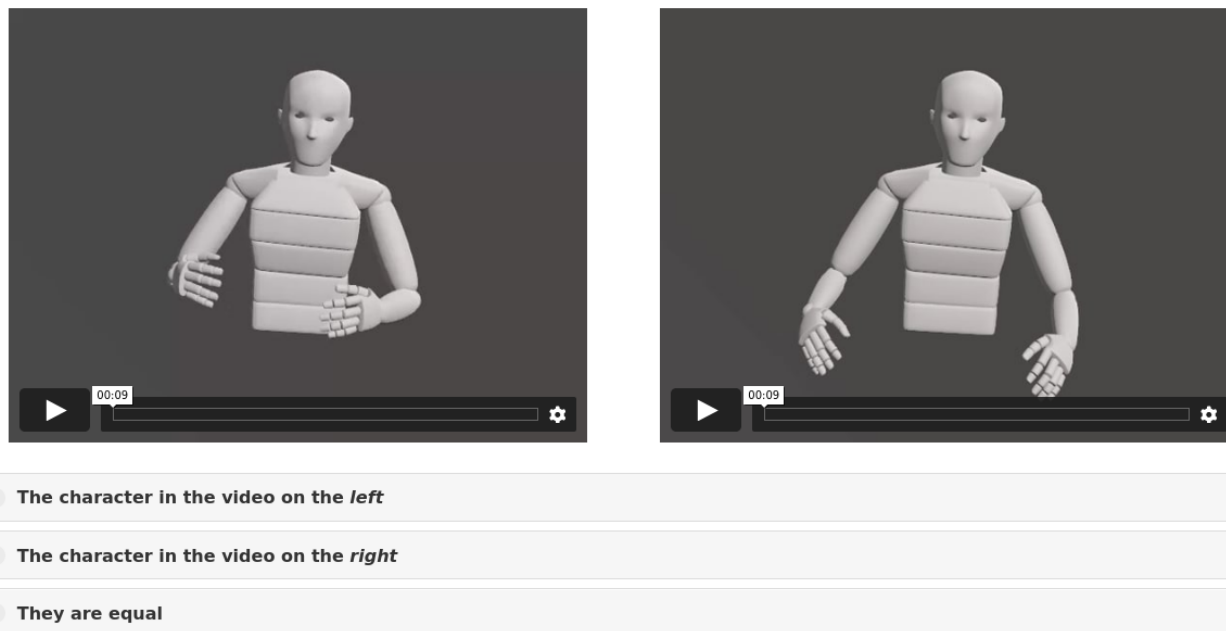
Figure 1: Interface for pairwise comparison evaluation

the instructions. After the participants watched both videos and selected a response option, they could continue to the next page.

The rating scale interface (Figure 2) displays a single video at a time, with a rating scale displayed below. For all conditions, the question was: 'How human-like was the agent in this video?' Response options ranged from 1 to 5 and were labelled *not at all*, *slightly*, *somewhat*, *moderately*, and *extremely*. Videos could only be watched one-at-a-time, and participants were only able to advance to the next page when both videos had been played and rated.

3.4 Experimental Procedure

After participants were assigned to the task on Prolific, they were forwarded to the online evaluation system. Here, they were assigned an internal participant ID that corresponds to a configuration file containing the stimuli and order of stimuli to show to the participant, and when to run attention checks. Each participant evaluated a total of 22 video pairs. These 22 video pairs correspond to 10 videos evaluated in a pairwise comparison approach, and 10 in a rating style approach. Two of the 22 video pairs contained an attention check. The order of evaluation (pairwise comparison vs. rating approach) was based on the assigned ordering condition. The position of the attention checks in the series of evaluation pairs was randomised, and there were two types of attention checks: one in which the response option to select was provided visually and one in which it was provided acoustically.

After evaluating the 22 video pairs, participants were presented with a questionnaire collecting their age, gender, nationality, level of education and experience with computers. This was followed with open questions related to the procedure they just completed, and whether they had a preference for pairwise comparison or rating scale evaluations. Once done with the study, successful participants were rewarded with 2.50 GBP (pay on average was 7.23 GBP per hour when taking into account the average duration of the task). The time each participant spent on each page of the experiment (and overall) was also recorded to allow us to evaluate efficiency.

4 ANALYSES

4.1 Hypothesis 1

To test the hypothesis that the two comparison methods would result in different rank-orderings of stimuli, we used a correlational approach. We first calculated each stimulus' average score across participants for each comparison method. Average scores using the rating scale method ranged from 1 to 5, and average scores for the pairwise approach ranged from −1 to 1 (on a scale where 1 = the stimulus was preferred over the alternative, 0 = the stimulus and alternative were equal, and −1 = the alternative was preferred over the stimulus). We then estimated the Kendall Rank-Order correlation [15] between these two series.

4.2 Hypothesis 2

To test the hypothesis that the pairwise comparison method would have higher inter-rater agreement than the rating scale method, we used two statistical approaches. First, we estimated intraclass correlation coefficients (ICCs) using Model 2A [23] and calculated

Rate the videos shown below

How human-like was the agent in this video?

	Video 1	Video 2
	Play	Play

5 Extremely
4 Moderately
3 Somewhat
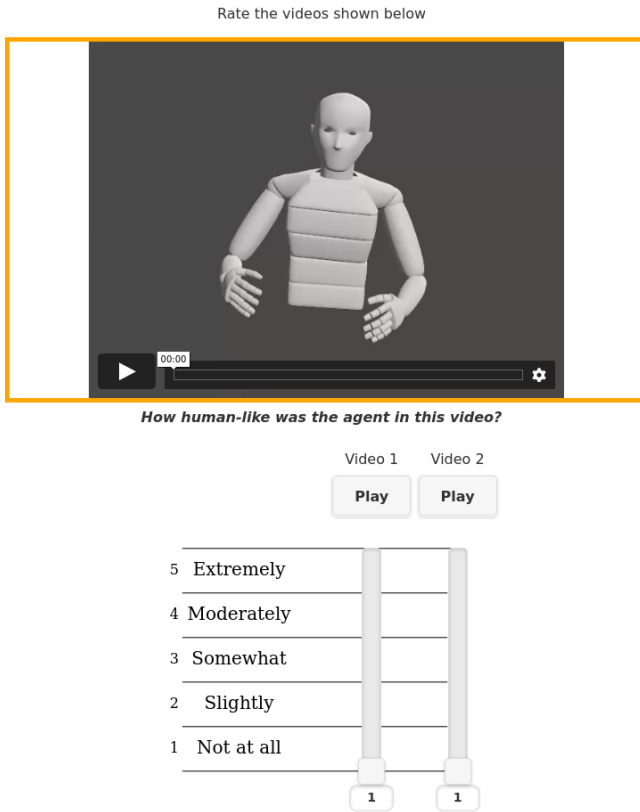2 Slightly
1 Not at all

1   1

Figure 2: Interface for rating scale evaluation

the absolute agreement of the average of 12 participants (i.e., the minimum number of participants assigned to any comparison). This approach estimates the reliability of the average of multiple participants' responses (which is what is used to compare video-generating methods), but assumes that the data approximates a continuous distribution (which is not the case for the pairwise method). As such, we also estimated chance-adjusted categorical agreement using quadratic-weighted kappa coefficients [10]. This approach is overly pessimistic in this case because it estimates the reliability of a single randomly selected participant's response, but it has the benefit of not assuming continuous data. In both cases, 2000 iterations of non-parametric bootstrapping [6] (with percentile-based confidence intervals and p-values) were used to compare the two approaches' inter-rater reliability.

4.3 Hypothesis 3

To test the hypothesis that the two comparison methods would differ in terms of time-efficiency (i.e., the time it takes a participant to complete a single comparison/page), we used a linear mixed effects modelling approach [9]. We estimated a model in which each page's completion time (in seconds) was regressed on a binary variable representing the comparison method. To control for practice and fatigue effects, we also regressed the completion time variable on a binary variable representing whether the comparison was during the first or second half of the experiment, and the method-by-half

interaction effect to allow the difference between comparison methods to differ between the first and second half of the experiment. Finally, to account for the clustering/nesting of comparisons within participants and videos, we included random intercepts for these variables and used Satterthwaite's approximation [19] to correct model degrees of freedom for small clusters.

4.4 Hypothesis 4

To test the hypothesis that participants would be more likely to prefer the pairwise comparison approach than the rating approach, we estimated an intercept-only logistic regression model to predict a binary variable representing whether each participant preferred the pairwise comparison approach over the rating comparison approach. We then back-transformed the intercept to probability units and tested whether it was significantly different from an equal preference of 50%.

4.5 Hypothesis 5

To test the hypothesis that the two comparison methods (i.e. rating scale and pairwise) would both find a difference in the case of a large difference in the quality of generated behaviour (i.e., Full vs. NoText stimuli) but not in the case of a small difference in the quality of generated behaviour (i.e. Full vs. NoSpeech stimuli), we used a linear mixed effects modelling approach [9]. We estimated a model in which the choice for the Full stimuli was regressed on other (NoText or NoSpeech) and order.

5 RESULTS

130 participants were recruited, of which 100 participants passed the attention checks. Of these, the mean age was 35.01 (SD=12.64), 55 identified as female, 45 as male. 68 of the participants were UK nationals, 22 were from the USA, 4 participants were Canadian, 2 Irish, 1 Australian, 1 Bulgarian, 1 Indian and 1 from New Zealand.

5.1 Hypotheses

5.1.1 Hypothesis 1. In Figure 3, we can see the relationship between the average pairwise scores and the average rating scores. We quantified the magnitude of this relationship using Kendall's Rank-Order Correlation. When we excluded trials where the two stimuli being compared were rated as equally human-like, we found a rank correlation of 0.44, 95% CI: [0.32, 0.55], $p < .001$. When we included trials where the two stimuli being compared were rated as equally human-like and assigned a pairwise score of 0, this correlation became 0.46, 95% CI: [0.35, 0.57], $p < .001$. Thus, although the two methods did not have exactly the same rank-ordering of stimuli, their rank-orderings were positively correlated (i.e., similar) to a high degree.

5.1.2 Hypothesis 2. Using the intraclass correlation approach, the inter-rater reliability coefficient was 0.62, 95% CI: [0.50, 0.69] for the rating scale method and 0.77, 95% CI: [0.71, 0.82] for the pairwise method; this difference was statistically significant ($\Delta = 0.15$, 95% CI: [0.06, 0.27], $p < .001$). Using the chance-adjusted categorical agreement approach, the quadratic-weighted kappa coefficient was 0.14, 95% CI: [0.09, 0.18] for the rating scale method and 0.23, 95% CI:
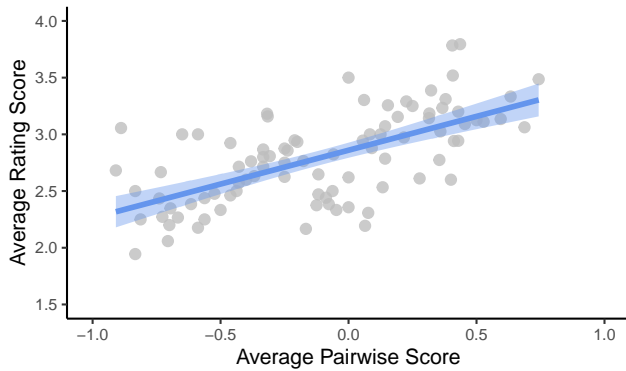
Figure 3: Relationship between average rating and pairwise scores. The two are positively correlated.

[0.18, 0.28] for the pairwise method; this difference was statistically significant ($\Delta = 0.09$, 95% CI: [0.03, 0.16], $p = .009$).

5.1.3 Hypothesis 3. The main effect of comparison method was significantly greater than zero, $B = 6.07$, 95% CI: [2.36, 9.77], $p = .002$ (see Figure 4). The unstandardised slope estimate of 6.07 means that pages were completed an average of around 6 seconds faster for the pairwise approach than for the rating approach. The main effect of ordering was not significantly different from zero ($p = .491$) and the type-by-ordering interaction effect was also not significantly different from zero ($p = .600$), which means that completion time did not significantly differ between the first and second half of the experiment and that the difference between comparison methods did not depend on which came first or second in the experiment.

If we want to know what the time difference would be for an entire experiment, we can multiply this page-level effect by the number of pages shown to participants. For 10 pages, as we did in this study, the experiment-level difference would be around 60 seconds.
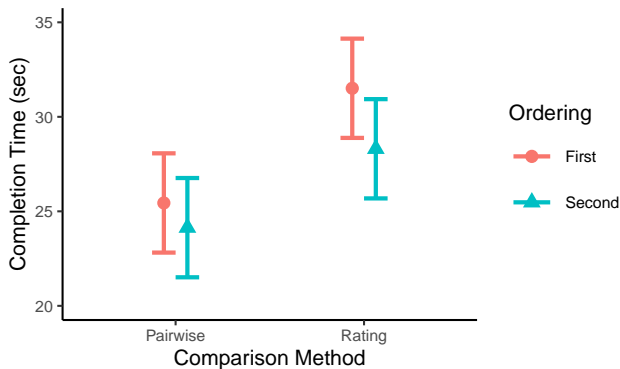

Figure 4: Completion time across conditions (error bars are 95% CIs), showing that the pairwise method is approximately 6 seconds faster per page than the rating method.

5.1.4 Hypothesis 4. The intercept for preference for the pairwise method was estimated at 56.0%, 95% CI: [46.2%, 65.5%] and was not significantly different from an equal preference of 50% ($p = .231$). Thus, we cannot conclude that participants reliably preferred one method over the other.

5.1.5 Hypothesis 5. For the rating scale method, the main effect of other was significantly greater than zero, $B = 0.66$, 95% CI: [0.40, 0.92], $p < .001$. This means that the extent to which the Full stimuli were rated higher than the other stimuli was greater for the HighDiff stimuli than for the LowDiff stimuli. In this model, neither the main effect of ordering ($p = .439$) nor the other-by-ordering interaction effect ($p = .860$) were significant. For the pairwise method, the main effect of other was significantly greater than zero, $B = 1.07$, 95% CI: [0.45, 1.69], $p < .001$. This means that the probability of preferring the full stimulus over the other stimulus was greater for the HighDiff stimuli than for the LowDiff stimuli. In this model, neither the main effect of ordering ($p = .750$) nor the other-by-ordering interaction effect ($p = .094$) were significant. Despite different scaling, the two methods had very similar results that matched our hypotheses and also matched the results from the original study we were reproducing [17] (see Figure 5).

6 DISCUSSION

In this study, we explored the differences in evaluating gesture motion stimuli with both pairwise comparisons and rating scales. Our aim was to gain a deeper understanding of when to use each approach. For this, we looked at the stimulus rankings both methods provided, their inter-rater reliability, the time it took participants to complete evaluations, participant preferences, and the conclusions both methods would yield regarding the comparison of gesture generation methods with high and low differences in quality.

The rank-ordering of stimuli between the pairwise comparisons and rating scales had a moderate positive correlation. We can conclude that in order to rank stimuli, in this instance co-speech gestures, there is not one approach that is preferred over the other; both are able to subjectively distinguish bad from good stimuli and this can be used to establish an order of quality.

When we take a look at the inter-rater reliability, we see a higher reliability for the pairwise method. This suggests that the pairwise method might be preferred over the rating scale method in terms of reliability.

When we look at which approach is faster, we can conclude that each comparison using the pairwise method was, on average, 6 seconds faster (25s instead of 31s) than each comparison using the rating scale method, which aligns with the findings of previous studies [4]. Although this difference was statistically significant (i.e., reliable), a difference of 6 seconds per comparison is likely too small to make much of a practical difference unless the number of comparisons being made by each participant was large (e.g., 100 or more).

Whether participants reliably preferred one comparison method over the other depended on which method they were assigned to use first. Those participants who used the rating scale method and then the pairwise method significantly preferred the pairwise method. However, those who used the pairwise method and then the rating scale method did not show a reliable preference for either
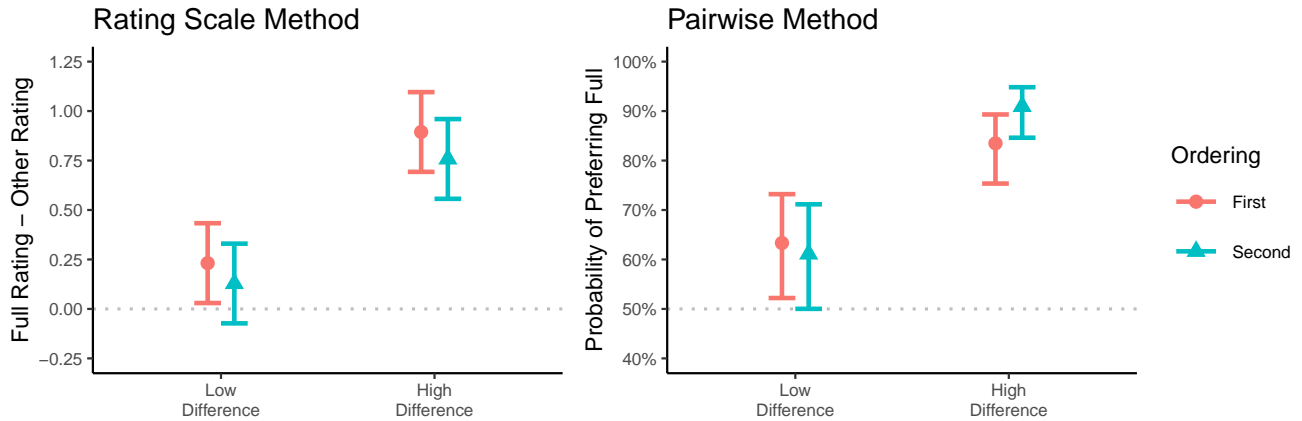
Figure 5: Comparison of generation methods by condition and evaluation method (error bars are 95% CIs)

method. This provides tentative evidence that the pairwise method may be more user-friendly.

In line with a previous study [17], we found that a high qualitative difference is indeed picked up by subjective evaluations. Not only does this hold for pairwise comparisons, but also for the rating scale approach. Both methods can provide similar results and are equivalent when comparing two or more conditions, for example two different models used to generate behaviour.

6.1 Limitations

For this study we gathered 2200 evaluations submitted by 100 participants. Due to the random drawing of stimuli, stimuli did not have the same number of responses. A pseudo-random spreading of stimuli over participants could have avoided this. The fact that participants could watch the videos simultaneously in the pairwise comparison interface but not in the rating scale interface may have contributed to the difference in average completion time between the two methods. We only considered 'human-likeness' in terms of assessing the quality of the generated gestures assessed by the participants, and are aware how limiting this question is in relation to the full spectrum of possible questions in relation to the evaluation of these stimuli. We opted for this strategy as the aim of this work was not to demonstrate which questions are most appropriate for the evaluation, but to compare the outcomes of two different evaluation strategies.

6.2 Recommendations

Based on our results, we have found no strong evidence to prefer one evaluation method over the other. The study does however allow us to make a number of recommendations for each method in relation to the domain of gesture generation, taking into account previous studies in other domains.

Pairwise comparisons may be better suited when a large number of stimuli are to be evaluated, as this not only results in a shorter study but is likely to avoid fatigue in participants. If only a small number of conditions are under consideration, then pairwise comparisons of conditions is practical, but as the number of

combinations grows with the faculty of the number of conditions ($\frac{n!}{2(n-2)!}$, with n the number of conditions) pairwise comparisons tend to become unwieldy for 4 or more conditions if we want to compare all versus all.

Rating scales may be more appropriate when fine-grained evaluations are needed, as ratings can not only be used to rate stimuli between conditions, but can also be used to rank stimuli within conditions. Ratings are also recommended when more than 3 conditions are under considerations, as the number of required ratings grows linearly with the number of conditions and stimuli. We would however like to emphasise the importance of providing anchors/labels for each response option in the rating scales [35]. When using rating scales, it is also recommended to calibrate participants' judgements by showing the participants poor and excellent stimuli during a brief training session. While the lack of calibration can somewhat be addressed by normalising participants' ratings, resolution and reliability are lost when participants are not properly trained before starting their rating task.

Finally, it is important to consider the type of information provided by each evaluation method. Rating scales provide information about the quality of each stimulus on an *absolute* scale, whereas pairwise comparisons provide information on a *relative* scale. Thus, you could use the pairwise comparison method to establish whether one method of generating human-like behaviour was reliably preferred over another. However, being 'better' is not always the same as being 'good'. For instance, one method could be considered 'poor' and the other 'very poor'; this would likely result in a big difference in pairwise comparisons, but it would be a mistake to conclude that the former was therefore high quality in absolute terms. This is where carefully crafted rating scales (and qualitative methods, such as interviews and free response boxes) can provide additional information about quality in general.

7 CONCLUSION

Objective evaluation measures of generated human-like behaviour often provide insufficient information to fully assess the quality of the behaviour. As such, these measures are often supplemented with

subjective evaluations. However, the field of gesture generation currently overlooks the amount of work that has been done in other fields that deal with subjective evaluations. This paper compared two popular methods, pairwise comparisons and rating scales, and found that both were equally effective to assess the quality of generated behaviour and provided surprisingly similar results in terms of rank-ordering of stimuli, inter-rater reliability, participant usability preferences, and the conclusions they yielded regarding the comparison of different stimuli-generation methods. We found that pairwise comparisons were slightly faster and showed somewhat higher inter-rater reliability, whereas the rating scale approach provided information on both absolute and relative quality and is capable of scaling better when comparing more than two stimuli at a time. These insights are increasingly relevant in a time where quantitative quality measures are used to drive research and development, especially when the use of data-driven methods tends to put draw attention away from subjective measures in favour of objective loss functions. Subjective measures are likely to remain the gold standard in evaluation studies and a better understanding of their capabilities benefits the study of multimodal behaviour generation.

8 ACKNOWLEDGMENT

REFERENCES

[1] Chaitanya Ahuja, Dong Won Lee, Ryo Ishii, and Louis-Philippe Morency. 2020. No Gestures Left Behind: Learning Relationships between Spoken Language and Freeform Gestures. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.findings-emnlp.170

[2] Nichola Burton, Michael Burton, Dan Rigby, Clare AM Sutherland, and Gillian Rhodes. 2019. Best-worst scaling improves measurement of first impressions. *Cognitive research: principles and implications* 4, 1 (2019), 1–10.

[3] Vijay Chidambaram, Yueh-Hsuan Chiang, and Bilge Mutlu. 2012. Designing persuasive robots: how robots might persuade people using vocal and nonverbal cues. In *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*. 293–300.

[4] Andrew P Clark, Kate L Howard, Andy T Woods, Ian S Penton-Voak, and Christof Neumann. 2018. Why rate when you could compare? Using the "EloChoice" package to assess pairwise comparisons of perceived physical strength. *PloS one* 13, 1 (2018), e0190393.

[5] Jamie DeCoster, Anne-Marie R. Iselin, and Marcello Gallucci. 2009. A Conceptual and Empirical Examination of Justifications for Dichotomization. *Psychological Methods* 14, 4 (2009), 349–366. https://doi.org/10/bh86w7

[6] Bradley Efron and Robert J. Tibshirani. 1993. *An Introduction to the Bootstrap*. Chapman and Hall, New York, NY.

[7] Lois Lawrence Elliott. 1958. Reliability of judgments of figural complexity. *Journal of experimental psychology* 56, 4 (1958), 335.

[8] Stephen T Fisher, David J Weiss, and René V Dawis. 1968. A comparison of likert and pair comparisons techniques in multivariate attitude scaling. *Educational and Psychological Measurement* 28, 1 (1968), 81–94.

[9] Andrzej Gałecki and Tomasz Burzykowski. 2013. Linear mixed-effects model. In *Linear Mixed-Effects Models Using R*. Springer, 245–273.

[10] Kilem Li Gwet. 2014. *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters* (fourth ed.). Advanced Analytics, Gaithersburg, MD.

[11] Jaap Ham, Raymond H Cuijpers, and John-John Cabibihan. 2015. Combining robotic persuasive strategies: the persuasive power of a storytelling robot that uses gazing and gestures. *International Journal of Social Robotics* 7, 4 (2015), 479–487.

[12] Kristiina Janhunen. 2012. A comparison of Likert-type rating and visually-aided rating in a simple moral judgment experiment. *Quality & Quantity* 46, 5 (2012), 1471–1477.

[13] Patrik Jonell, Taras Kucherenko, Ilaria Torre, and Jonas Beskow. 2020. Can we trust online crowdworkers? Comparing online and offline participants in a preference test of virtual agents. In *Proceedings of the 20th ACM International Conference on Intelligent Virtual Agents*. 1–8.

[14] Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, Taras Kucherenko, and Gustav Eje Henter. 2021. HEMVIP: Human evaluation of multiple videos in parallel. In *Proceedings of the International Conference on Multimodal Interaction*.

[15] Maurice G Kendall. 1938. A New Measure of Rank Correlation. *Biometrika* 30, 1 (1938), 81–93. https://doi.org/10/ch8zq6

[16] Taras Kucherenko, Dai Hasegawa, Naoshi Kaneko, Gustav Eje Henter, and Hedvig Kjellström. 2021. Moving fast and slow: Analysis of representations and post-processing in speech-driven automatic gesture generation. *Int. J. Hum. Comput. Interact.* (2021). https://doi.org/10.1080/10447318.2021.1883883

[17] Taras Kucherenko, Patrik Jonell, Sanne van Waveren, Gustav Eje Henter, Simon Alexandersson, Iolanda Leite, and Hedvig Kjellström. 2020. Gesticulator: A framework for semantically-aware speech-driven gesture generation. In *Proceedings of the International Conference on Multimodal Interaction*. 242–250.

[18] Taras Kucherenko, Patrik Jonell, Youngwoo Yoon, Pieter Wolfert, and Gustav Eje Henter. 2021. A large, crowdsourced evaluation of gesture generation systems on common data: The GENEA Challenge 2020. In *26th International Conference on Intelligent User Interfaces*. 11–21.

[19] Alexandra Kuznetsova, Per B Brockhoff, and Rune H B Christensen. 2017. lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software* 82, 13 (2017), 1–26. https://doi.org/10/dg3k

[20] Weixin Liang, James Zou, and Zhou Yu. 2020. Beyond user self-reported likert scale ratings: A comparison model for automatic dialog evaluation. *arXiv preprint arXiv:2005.10716* (2020).

[21] Kelsey Lucca and Makeba Parramore Wilbourn. 2018. Communicating to learn: Infants' pointing gestures result in optimal learning. *Child development* 89, 3 (2018), 941–960.

[22] Hector Martinez, Georgios Yannakakis, and John Hallam. 2014. Don't Classify Ratings of Affect; Rank Them! *IEEE Transactions on Affective Computing* 3045, c (2014), 1–1. https://doi.org/10/f6pnzt

[23] Kenneth O McGraw and S P Wong. 1996. Forming Inferences about Some Intraclass Correlation Coefficients. *Psychological Methods* 1, 1 (1996), 30–46. https://doi.org/10/br5ffs

[24] David McNeill. 1992. *Hand and mind: What gestures reveal about thought*. University of Chicago press.

[25] Kim T Mueser, Barry W Grau, Steve Sussman, and Alexander J Rosen. 1984. You're only as pretty as you feel: facial expression as a determinant of physical attractiveness. *Journal of Personality and Social Psychology* 46, 2 (1984), 469.

[26] Laura Pérez-Mayos, Mireia Farrús, and Jordi Adell. 2019. Part-of-speech and prosody-based approaches for robot speech and gesture synchronization. *Journal of Intelligent & Robotic Systems* (2019), 1–11.

[27] Andrew S Phelps, David M Naeger, Jesse L Courtier, Jack W Lambert, Peter A Marcovici, Javier E Villanueva-Meyer, and John D MacKenzie. 2015. Pairwise comparison versus Likert scale for biomedical image assessment. *American Journal of Roentgenology* 204, 1 (2015), 8–14.

[28] Pilar Prieto Vives, Alfonso Igualada Pérez, and Núria Esteve Gibert. 2017. Beat gestures improve word recall in 3-to 5-year-old children. *Journal of Experimental Child Psychology. 2017 Apr; 156: 99-112* (2017).

[29] Mijke Rhemtulla, this link will open in a new window Link to external site, Patricia É Brosseau-Liard, and Victoria Savalei. 2012. When Can Categorical Variables Be Treated as Continuous? A Comparison of Robust Continuous and Categorical SEM Estimation Methods under Suboptimal Conditions. *Psychological Methods* 17, 3 (2012), 354–373. https://doi.org/10/f395ws

[30] Maha Salem, Friederike Eyssel, Katharina Rohlfing, Stefan Kopp, and Frank Joublin. 2013. To err is human (-like): Effects of robot gesture on perceived anthropomorphism and likability. *International Journal of Social Robotics* 5, 3 (2013), 313–323.

[31] Michael Schoeffler, Sarah Bartoschek, Fabian-Robert Stöter, Marlene Roess, Susanne Westphal, Bernd Edler, and Jürgen Herre. 2018. webMUSHRA—A comprehensive framework for web-based listening tests. *Journal of Open Research Software* 6, 1 (2018).

[32] Leonard J Simms, Kerry Zelazny, Trevor F Williams, and Lee Bernstein. 2019. Does the Number of Response Options Matter? Psychometric Perspectives Using Personality Questionnaire Data. *Psychological Assessment* 31, 4 (2019), 557–566. https://doi.org/10/gfxv4h

[33] Yao-Ting Sung and Jeng-Shin Wu. 2018. The Visual Analogue Scale for Rating, Ranking and Paired-Comparison (VAS-RRP): a new technique for psychological measurement. *Behavior research methods* 50, 4 (2018), 1694–1715.

[34] Petra Wagner, Jonas Beskow, Simon Betz, Jens Edlund, Joakim Gustafson, Gustav Eje Henter, Sébastien Le Maguer, Zofia Malisz, Éva Székely, Christina Tånnander, and Jana Voße. 2019. Speech synthesis evaluation – State-of-the-art assessment and suggestion for a novel research program. In *Proc. SSW*, Vol. 10. ISCA, Vienna, Austria, 105–110. https://doi.org/10.21437/SSW.2019-19

[35] Bert Weijters, Elke Cabooter, and Niels Schillewaert. 2010. The effect of rating scale format on response styles: The number of response categories and response

category labels. *International Journal of Research in Marketing* 27, 3 (2010), 236–247.

[36] Pieter Wolfert, Taras Kucherenko, Hedvig Kjellström, and Tony Belpaeme. 2019. Should Beat Gestures Be Learned Or Designed?: A Benchmarking User Study. In *ICDL-EPIROB 2019 Workshop on Naturalistic Non-Verbal and Affective Human-Robot Interactions*.

[37] Pieter Wolfert, Nicole Robinson, and Tony Belpaeme. 2021. A Review of Evaluation Practices of Gesture Generation in Embodied Conversational Agents. *arXiv preprint arXiv:2101.03769* (2021).

[38] Georgios Yannakakis, Roddy Cowie, and Carlos Busso. 2021. The Ordinal Nature of Emotions: An Emerging Approach. *IEEE Transactions on Affective Computing* 12, 1 (2021), 16–35. https://doi.org/10.1109/TAFFC.2018.2879512

[39] Georgios Yannakakis and Héctor P. Martínez. 2015. Grounding Truth via Ordinal Annotation. In *2015 International Conference on Affective Computing and Intelligent Interaction (ACII)*. 574–580. https://doi.org/10/gjp74q

[40] Youngwoo Yoon, Bok Cha, Joo-Haeng Lee, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. 2020. Speech gesture generation from the trimodal context of text, audio, and speaker identity. *ACM Transactions on Graphics (TOG)* 39, 6 (2020), 1–16.