Rating vs. paired comparison for the judgment of dominance on first impressions

Fabrizio Nunnari, Alexis Heloir

Rating Vs. Paired Comparison for the Judgment of Dominance on First Impressions

Fabrizio Nunnari and Alexis Heloir

Abstract—This article presents a contest between the rating and the paired comparison voting in judging the perceived dominance of virtual characters, the aim being to select the voting mode that is the most convenient for voters while staying reliable. The comparison consists of an experiment where human subjects vote on a set of virtual characters generated by randomly altering a set of physical attributes. The minimum number of participants has been determined via numerical simulation. The outcome is a sequence of stereotypes ordered along their conveyed amount of submissiveness or dominance. Results show that the two voting modes result in equivalently expressive models of dominance. Further analysis of the voting procedure shows that, despite an initial slower learning phase, after about 30 votes the two modes exhibit the same judging speed. Finally, a subjective questionnaire reports a higher (63.8 percent) preference for the paired comparison mode.

Index Terms—Rating, paired comparison, virtual characters, dominance, submissiveness, personality trait

1 Introduction

Part of the research in Affective Computing investigates the attribution and simulation of emotions and personality in virtual characters. Since the perception of emotion and personality relies on mechanism which are hidden in some unexplained biological process, affective software systems must be trained and evaluated using subjective experiments.

The de-facto standard for subjective evaluation of affective stimuli is *rating* 5- to 7-point scales: a set of questions, with closed range answers, where subjects give an absolute judgment. A Likert-item, as originally defined by Rensis Likert [1], presents a question of agreement followed by a choice among 5 different levels marked "strongly disagree, disagree, neither agree nor disagree, agree, strongly agree". Variations (Likert-type items) provide a different granularity (7 or more choices) and might not even allow for a neutral answer (e.g., 10 levels). Alternatively, ratings can be provided on numerical scales presenting scores from 1 to 5, or any other integer maximum value. However, from recent investigations it appears that rating methods have several limitations [2] or even fundamentally wrong theoretical grounding [3]. In particular, rating methods have the tendency to collect votes at the center of the scale, present inconsistencies among sessions, and are prone to judgment drifts during a voting session.

Alternatively, *ranking* is a preference learning technique which solves the above-mentioned issues, and seems to better handle the high variance emerging from votes on affect-related topics, which is mainly due to the high subjectivity of the matter. In a ranking task, subjects are asked to sort a set of items by preference. When the number of items is too big, the sorting task might become too difficult or too long. *Paired comparison* (PC) is a special case of ranking involving only two items at a time. Given a set of items, a PC session consists of showing two items at a time to a panel of judges. Each judge must select which of the two is the "preferred" item, or state that he has no preference. The judges are not required to vote on all possible pair combinations. The outcome of a PC session associates a ranking value, or *estimate*, to each of the judged items, allowing for the choice of the most and the least preferred items, as well as an ordering among all the items.

The investigation presented in this paper builds on previous work of the authors (Nunnari and Heloir [4], [5]) who used pairwise comparison to collect data on the perception of dominance and trustworthiness from aesthetics of virtual characters. Despite of the positive results, the authors left unclear the advantages of ranking over a rating approaches. The goal of this paper is to investigate whether pairwise comparison can in general be considered as a valid alternative to rating methods. As described later in Section 2, a number of works have been published in the field of affective computing (and not only) directly comparing rating and ranking techniques, which led to the either inability to elect a winner or to contrasting results.

This paper contributes to such line of research by presenting a direct comparison between the *rating* against the *paired comparison* voting techniques in the specific context of measuring to what extent people perceive a virtual character as *dominant*. In particular, we present a user study where subjects, using the rating mode, look at the picture of a single virtual character and select on a 7-point scale how the character looks, from "very submissive" to "very dominant". In PC mode, subjects watch side-by-side the pictures of two different virtual characters and answer which of the two looks "more dominant".

- *Fabrizio Nunnari is with DFKI / MMCI, 66123 Saarbrücken, Germany. E-mail: fabrizio.nunnari@dfki.de.*
- *Alexis Heloir is with LAMIH UMR CNRS 8201/UPHF, 59300 Valenciennes, France. E-mail: alexis.heloir@uphf.fr.*

With respect to existing work, our experiment design controls simultaneously several aspects: i) votes are collected in a controlled environment through physical participation, thus limiting the biases normally introduced by crowd sourcing methods; ii) it sets a balance of the voting effort between the two modes; iii) it considers an a-priori estimation of the number of participants needed to minimize biases due to randomness in the study; iv) the duration of votes is accurately measured on a per-vote granularity. A comprehensive statistical analysis of the collected data measures the advantages and disadvantages of both approaches in terms of effectiveness, time usage, reliability, and user preference.

In the following, Section 2 gives references on the definition and perception of *dominance*, and on the history of *paired comparison* mode and its evaluation with respect to ratings. The experiment itself is presented in Section 3, while Section 4 presents some concluding remarks.

2 Related Work

The following sections present related work on the perception of dominance, which was selected as the pivotal personality trait due to the amount of work already conducted on it, followed by a review on the pairwise comparison mode and its relation with rating modes.

2.1 The Perception of Dominance

There has been recent work investigating the perception of personality from aesthetics, i.e., investigating if a human can correctly guess the personality of another subject from a judgement of the physical aspect [6]; results are modest but positive, Surprisingly, since the challenging part is that the study employed the widely used Five-Factor personality model (aka Big-5 or OCEAN model) [7].

In fact, the model of Openness to experience, Conscientiousness, Extraversion, Agreeableness, and Neuroticism was conceived to describe behavioral, cognitive, and emotional patterns, rather than visual aspects. Rather, Smith and Neff [8] have concluded that judgments on non interactive characters would better fit to the two-dimension model of *plasticity* and *stability*.

For judgements of personality based on aesthetics of static stimuli (i.e., static pictures in neutral poses, aka, zero-acquaintance first encounters), Oosterhof and Todorov found that the couple Dominance + Thustworthiness is a more appropriate model of *perceived* personality [9]. Dominance is since long time well recognized and was already present in the Cattell's 16 Personality Factors theory in the '70s [10], and more recent work extends the model of dominance and trustworthiness with an additional "youthful-attractiveness" dimension [11].

All this recent work was conducted using virtual characters, rather than pictures of real subjects, because generators of synthetic humans allow for a precise control of the numerical parameters associated to the visual output. Additionally, they all share the same approach: gather a set of pictures of real or virtual characters and ask to human subjects to vote for the perceived personality. The goal being to build a model able to predict the perceived personality of a picture.

Reversing the goal, there is work which instead consider a personality profile as input and produce a virtual character who would elicit the given profile when judged. Examples are in the field of behaviour control [8], [12], movement style [13], and generation of body shape [14]; all using the OCEAN model. Concerning models for zero-acquaitance, a generation of pictures from dominance and trustworthiness is proposed by Vernon *et al.* [15] and by Nunnari and Heloir [4], [5].

Concerning the correlation between perceived dominance and facial features, Toscano *et al.* [16] already reported that the more the eyebrows are inclined towards the center of the face (low inner brows), the more the subject is perceived as dominant. Other features positively influencing the perception of dominance are: short (vertically) eyes, chin length, less wide head, and the width of the nose and mouth. Windhager *et al.* [17] reported about the positive correlation between the perception of dominance and rounder facial shape, thicker eyebrows, smaller eyes, shorter nose, broadening of the lower face. Already back in 1981, Keating *et al.* [18] investigated on the facial features influencing the perception of dominance using a pairwise comparison. Positive correlations were found to receded hairlines, large jaws, and thin lips.

In our experiments, we included all of the above-mentioned elements, together with full body pictures of the characters, visibly modulating the body size and shape, in the attempt to disclose the correlation of more body-related features. We didn't find relevant past work on the perception on full-body features and the perception of dominance, with the exception of a study employing children as subjects [19], which found a positive correlation with height.

2.2 The Paired Comparison Voting Mode

The paired comparison technique was first considered back in the 19th century by Fechner [20], in the field of psychophysics, in order to deal with the subjectivity of humans exposed to experimental stimuli and to find a solution to the problem that subjects are not sensitive enough to small variations in stimuli. From the principles of Fencher, Thurstone [21] extracted the *law of comparative judgements* and promoted its application beyond psychophysics, such as in psychology and education.

In 1946, Guttman [22] investigated how to "determine a numerical value for each of a number of items which will best represent the comparisons in some sense." In other words, he investigated how to extract from a PC session numerical values to associate to each of the voted items, so that items can be sorted in a way that reflects user preferences. His solution is based on an iterative numerical analysis. In 1956, Bradley and Therry [23] proposed a solution to the problem addressed by Guttman using statistical techniques that associates each item of the compared set to an *estimate*.

Two indices measure the quality of the estimates: α and G. The separation reliability $\alpha \in [0, 1]$ (aka maximum-likelihood estimates reliability) measures the level of consistency of the votes between the subjects, and its interpretation is similar to a Cronbach's alpha. On the other hand, the separation index G quantifies the divergence between the most and the least preferred items [24, p. 268]. More detailed descriptions of the paired comparison method can be found in several works: [24], [25], [26], [27], [28], [29].

The method of paired comparison has been used in different fields, such as education [30], [31], [32] and forest science [33], [34].

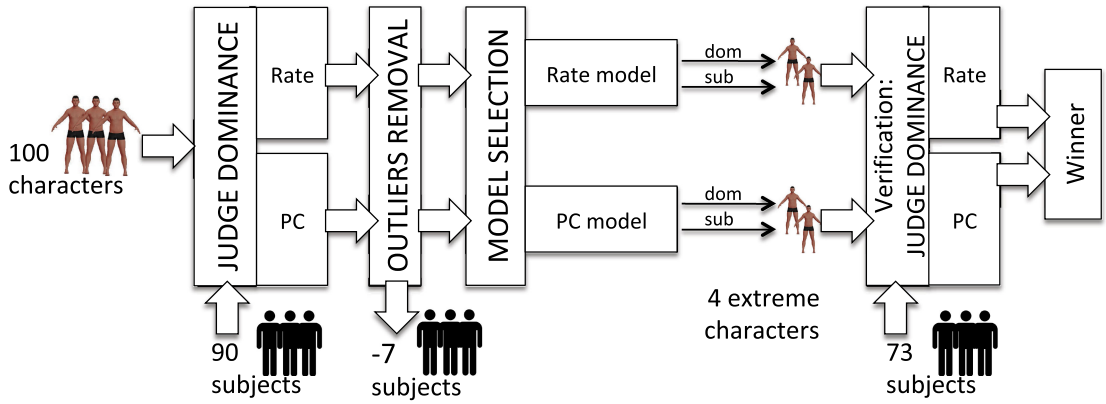

Fig. 1. The organization of the contest. Ninety human subjects judge the pictures of 100 virtual characters for the impression of dominance using the two voting modes: Rate and PC. After removing unreliable voters, a data analysis leads to the creation of two models, one for each voting mode. Each model is used to generate a fully dominant and a fully submissive character. Those 4 extreme stereotypes are again judged for the perception of dominance in order to elect the winner, i.e., the model(s) maximizing the perception of dominance or submissiveness.

In the realm of affective computing, Yang and Chen [35] used pairwise ranking to compare music excerpts (30 secs.) and rank them according to valence and arousal. They use a neural network to train a learn-to-rank model and organize the sound track in a bi-dimensional valence/arousal space.

This latter work inspired Baveye *et al.* [36] in the collection of ranking information (pairwise comparison) for the LIRIS-ACCEDE dataset: a collection of videos annotated for arousal and valence. Later, they converted the ranking scores into linear rating using Gaussian Regression [37]. To estimate the equivalence between ranking and rating methods, they involved 28 participants in the rating of 20 dataset samples on a 5-point scale. Results lead to a relatively high Spearman's rank correlation coefficient (0.751 for arousal and 0.795 for valence), the comparison did not linger on the equalization of the number of participants nor on the quantity of effort between the two voting conditions.

Lotfian and Busso [38] also investigated on different voting methods for the judgement of arousal and valence in audio clips. They compare the prediction performance of a model trained with pairwise data (analyzed through RankSVM) against a binary classification (using SVMs) and against a linear regression (using Support Vector Regressions, SVR). The results indicate that Ranking data lead to the most accurate model. However, the dataset for pairwise comparison was synthetically generated by procedurally comparing rated items. Still on the acoustic domain, Parthasarathy and Busso [39] collected sentence-level annotations of Arousal, Valence and Dominance from macro-acoustic features. When testing prediction models, they found that "preference-learning methods to rank-order emotional attributes trained with the proposed QAbased labels achieve significantly better performance than the same algorithms trained with relative scores obtained by averaging absolute scores across annotators." However, again, pairwise comparison data are not "natively" collected, but rather extracted from time-continuous emotional traces.

Karpińska-Krakowiak [40] performed a direct comparison between rating versus pairwise comparison techniques, and found that there is not significant differnce between the two methods. However, the experiments, were conducted asking users to vote on a very limited number of stimuli (five objects) for a very objective measurement criteria (height of trees and length of sticks).

Wood *et al.* [41] ran a set of experiments on the annotation of emotions in Tweeter text messages, including a comparison between ranking and pairwise comparison with more than 40K votes. They conclude that voting times are comparable and that rating modes lead to better annotators' agreement, but the votes were collected online and the voting time was just an estimate average among voting sessions.

3 Rating Versus Paired Comparison

The primary purpose of this experiment was to use both Rate and PC voting modes to extract, from pictures of virtual characters, the physical attributes (both facial and body features) that significantly affect the perception of dominance. The null hypothesis is that both voting modes will select the same list of physical attributes. While analysing the voting data, we compare the two approaches and report additional information on the consistency of the voters, an analysis on the duration, evidence of learning effects, and subjective preferences.

3.1 Experiment Overview

Fig. 1 depicts the organization of the experiment; it involved 90 subjects who gave a first impression judgement of dominance on a set of randomly generated virtual characters, the purpose being to build stereotype images of the most dominant and most submissive individuals. The judgements were collected in both Rate and Paired Comparison conditions for each participant. After an analysis of the voting time, seven of the participants were filtered out from the experiment. The remaining results, based on 83 participants, were used to compute two linear models (one per voting mode) predicting the perception of dominance from the physical aspect of the characters. The models were then simplified in order to select the phyisical attributes mostly influencing the perception of dominance. Finally, the simplified models were used to generate pictures of characters which would maximize or minimize the perception of dominance, yielding to four *extreme* characters. Those characters were used to run a smaller experiment aiming at verifying what of the two

voting modes led to characters expressing dominance (or submissiveness) the most.

3.2 Choosing the Number of Participants

The first problem is to find an appropriate minimum number N of pictures to vote on. As later described in Section 3.11, the generation of the characters is based on a linear regression where each of the voted-on pictures represents an "observation". We selected $N = 100$ as the number of different items to vote on. However, in order to keep the overall experiment (including both Rate and PC modes) below a 15-minute duration, each participant voted on a random selection of only 50 items.

As a design choice for the contest, participants were to provide the same amount of votes (50) in both Rate and PC modes. However, in the PC mode, $N = 100$ items lead to $N(N-1)/2 = 100 * 99/2 = 4950$ different pairs. Even though it is not necessary that each subject provides a vote for each of the possible pairs, the problem remains of determining how many subjects should be hired in order to provide reliable and statistically significant results.

We did not find in our literature review an established method to determine the minimum number of voters needed to reach a given confidence interval. Hence, we proceeded with a simulation approach, as described in the following. We simulated a Paired Comparison vote session on 100 items, 50 votes per subject, with an increasing number of subjects (from 10 to 100, with increments of 5), and different levels of *Agreement*. The items in the simulation are the natural numbers between 1 and 100. For every comparison, the "right choice" consists of giving preference to the greater number. When simulating the votes, the agreement level is used as the probability to select the greater number. The simulation ran with five levels of agreement: 0.5, 0.6, 0.7, 0.8, 0.9. Here, 0.5 denotes a completely random voting, while 0.9 denotes a very high consensus among the subjects who are voting. For each condition, the simulation ran on 10 different random voting sessions. The simulation does not consider the case of undecided preference.

Fig. 2 shows the result of the simulation. The top plot reports the variation of the separation index G, which as expected increases with the number of subjects involved in the experiment. Additionally, it is worth noting that the G stays well above 0 even in the case of completely random voting (Agreement = 0.5). Hence, it cannot help in determining a minimum number of voters. The reliability α (not plotted) reaches values above 0.9 after 30 subjects, suggesting this as a minimum number of voters.

For a deeper investigation, we also analyzed the progression of the minimum and maximum estimates among all 100 items. As expected (see Fig. 2, bottom), all the minimum estimates converge to 0.0 independently from the agreement level, while the maximum estimates increase together with the agreement level. The variation of the estimates converges up to 45 voters and then stabilizes, suggesting that ca. 50 voters are sufficient for reliable results.

Since our simulation doesn't consider *draws* as result of the comparisons, we aimed at hiring more subjects. For our study, we were able to recruit 114 subjects. Out of them, 90 were able to finish the study without technical problems.

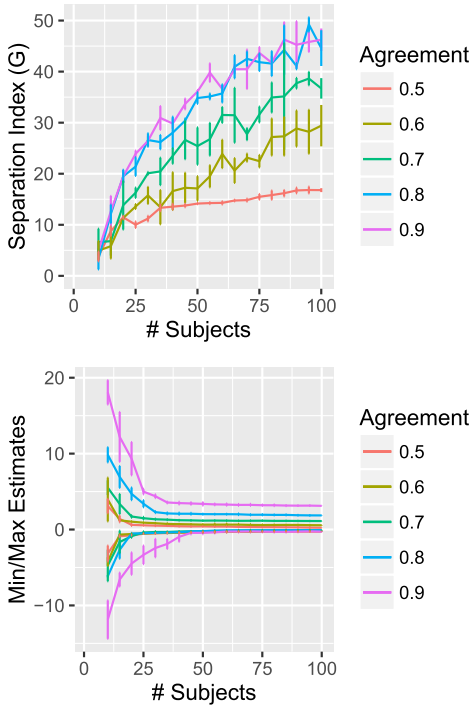

Fig. 2. The simulation of a paired comparison session among 100 items. Top: the resulting separation index G. Bottom: the minimum and maximum estimates among all items. Vertical bars indicate the variance across 10 simulations.

3.3 Experiment Design

The experiment consisted of asking human subjects to give a subjective judgement on the level of dominance of a virtual character. Each subject participated in two sessions, one for each voting *mode*: *Rate* (Fig. 3, left) and *PC* (Paired Comparison, Fig. 3, right). For each condition, there were 55 trials. Each trial consisted of viewing the face and the body of the virtual character(s) and giving a first-impression judgement of its Dominance. For the Rate mode, the judgement was given on a 7-level scale. For the PC mode, the judgment was given by comparing two characters side-by-side and expressing which of the two was more dominant. It was also possible to opt for a non-decision ("I don't know"). Of the 55 trials, 5 were repetitions. Every 10 trials, we presented a random repetition of a previously voted-on trial. Such repetitions were inserted in order to check the reliability of the voters. Half of the participants started voting in PC mode and the other half started in Rate mode (variable *mode-first*).

All the images (483x419 pixels) were displayed at a width of 10cm on monitors of comparable resolution (HD 1920x1080). During the experiment, participants could read on a printed paper the following definitions: Dominant "Has power and influence over others", and Submissive is "Ready to conform to the authority or will of others; meekly obedient or passive".

3.4 Vote Collection

The experiment was conducted in a controlled environment composed of a couple of large tables located along the corridor accessing the university dining hall of the Saarland University campus in Saarbrücken, Germany (see Fig. 4). On the tables, four workspaces with PC, mouse, and monitor were

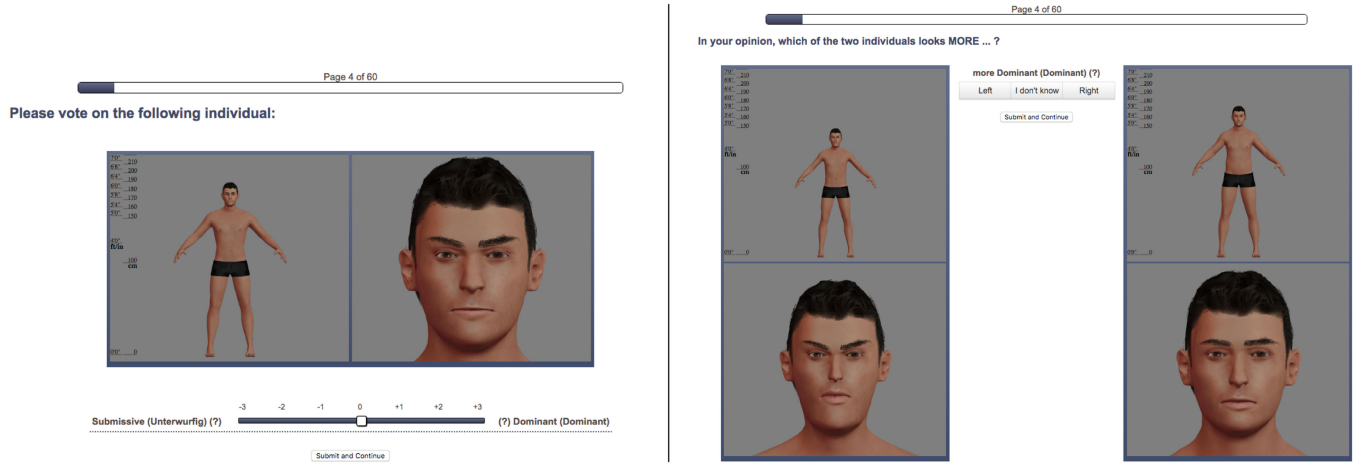

Fig. 3. Screenshots of the judgment trials in the two voting modes. Left: Rate; right: Paired Comparison.

facing the wall. We verified that the background noise of people chatting while walking through the corridor never represented a real source of distraction. Among the 114 invited people, 90 subjects finished the experiment without technical problems.

The subjects were recruited by randomly asking people walking through the corridor during the lunch break. Subjects were rewarded with a coupon of 2.85 Euro (full student meal). Each subject was introduced to the purpose of the experiment, invited to sit down, and told that the experimenters were going to stand near-by and would always be available in case of problems or questions.

The experiment was implemented as a sequence of 60 pages shown in a full-screen browser. The first three pages showed, in order, an introduction, instructions, and an example of how to vote. The initial instructions explicitly stated: "There is no correct answer. What matters is your opinion" and that the experiment was going to last between 10 and 15 minutes. Additionally, subjects were told that there is no real time limit to the duration of the experiment, but that each evaluation should last no more than 10 seconds. The following 55 screens represented the judgement trials. The last two screens asked for personal information (age, gender, nationality, education level) and then showed a goodbye message. All the text was shown in both English and German.

3.5 Material

The images of the virtual characters were generated with the open-source software MakeHuman (http://www.makehumancommunity.org/, v1.0.2), which was patched to allow for a batch generation of multiple images. MakeHuman follows a slider-based approach for the easy generation of virtual characters. A default androgynous character can be modified through the use of more than 200 sliders. Each slider modifies one macro characteristic (e.g., gender, height, weight, muscularity) or a detail of the body (e.g., shoulder width, nose length, eyebrows inclination, eye size).

Fig. 4. Experiment setup.

Similarly to our previous work [5], the characters for this experiment were generated by modulating 14 attributes, listed in Table 1 together with their minimum and maximum values according to the MakeHuman scale. We selected some of the attributes because of the literature recognizing them as triggering the perception of the dominance. The remaining attributes were selected because of their clear influence on the frontal aspect of the avatars. The min/max ranges were tuned by the authors to achieve visible changes in the characters while avoiding implausible or unnatural morphologies. As a reference, Fig. 5 shows the extreme characters that can be generated by respectively minimizing and maximizing all of the attributes at once.

One hundred characters were generated by uniformly randomizing their attribute values.

3.6 Outlier Filtering

According to our observations, subjects spending more than 40 seconds to accomplish a trial were either distracted by acquaintances passing through the corridor or by activities on their mobile phone. As a consequence, we filtered out 7 subjects from the data collection. The continuation of this analysis is based on the contributions of 83 subjects (50 male, 33 female). The remaining participants' average age was 26.8 (sd=6.4), mostly with German nationality (46). Other significant groups were Indian (5), Mexican (4), Colombian (3), Egyptian (3), and Pakistani (3). The remaining subjects were from 16 different nationalities.

3.7 Vote Consistency

As described earlier, every 10 judgements, each subject had to vote again on a page randomly selected from the previous 10. This kind of consistency check mechanism is generally used to verify the reliability of votes collected through crowd-sourcing methods, i.e., in unsupervised experiments.

TABLE 1
Experiment 1: The Attributes Modulating the Shape of the Virtual Characters

MakeHuman ID	Short Name	Description	min	max	default
chin/chin-bones	Chin bones	Chin lateral bones extension	0.5	1	0.5
chin/chin-height	Chin height	Distance between chin and lower lip	0.2	0.8	0.5
eyebrows/eyebrows-angle	Eyebrows angle	Eyebrows inclination	0.2	0.8	0.5
eyes/r-eye-size	Eye size	Size of both eyes (mirroring applied)	0.1	0.9	0.5
head/head-oval	Head ovality	Hard/soft forehead corners	0	0.8	0
macrodetails-height/Height	Height	From ca. 149cm to 201cm	0.25	0.75	0.5
macrodetails-universal/Muscle	Muscularity	Muscular tone of the body	0.2	0.8	0.5
macrodetails-universal/Weight	Weight	Overall mass of the body	0.2	0.8	0.5
mouth/mouth-scale-horiz	Mouth hscale	Mouth and lips width	0.1	0.9	0.5
mouth/mouth-scale-vert	Mouth vscale	Mouth and lips height	0.1	0.9	0.5
neck/neck-scale-horiz	Neck hscale	Neck width	0	1	0.5
nose/nose-scale-horiz	Nose hscale	Nose width	0.1	0.9	0.5
stomach/stomach-tone	Stomach tone	Belly in/out	0.2	1	0.5
torso/torso-vshape	Torso V-shape	Affects shoulder width	0	0.8	0.5

For each of the 5 repetitions, we measured how much a subject *drifted* away from the previous corresponding judgment. In Rate mode, the drift is measured between 0 (same judgement) and 6 (opposite side of the scale). In PC mode, the drift can be 0 (same judgment), 1 (from a preference to a non-preference, or vice versa), or 2 (opposite preference).

Fig. 6 shows the results for Rate (left) and PC (right) as bar plots of drifting counts (y axis) for each possible drift value (x axis). The sum of the counts is 415 (83 subjects by 5 repetitions).

Table 2 reports the drift count in Rate mode together with the mean and standard-deviation among all subjects. Subjects performed a consistent judgment (drift 0) in 152 cases (36.6 percent) with an average of 1.83 times per subject on 5 votes. Interestingly, they more often gave an adjacent score (drift 1) 185 times (44.6 percent) for an average of 2.23 times per subject. There were more inconsistent than consistent votes. The number of inconsistencies rapidly decreases for the successive categories, and none of the participants ever switched to the opposite side of the scale. A Fisher's exact test shows that this distribution is significantly different from the ideal distribution having a count of 415 for category drift_0 and 0 for all the others ($p < 0.001$).

Table 3 reports the drift count in PC mode together with the mean and standard deviation among all subjects. In PC mode, subjects were more consistent: there were 281 consistent votes (67.7 percent) with an average of 3.38 per individual out of 5 votes. However, they drifted to an adjacent and to an opposite decision roughly the same amount of times. A Fisher's exact test shows that this distribution is significantly different from the ideal distribution having a count of 415 for category drift_0 and 0 for the two others ($p < 0.001$).

The consistency check mechanism was designed with the purpose of spotting users who are possibly voting randomly. Given the controlled conditions of the experiment, the results were rather surprising. We were expecting a higher consistency, and it is clear that this mechanism can not be reliably used to exclude participants from the experiment.

We can only conclude that, at least in personality judgment, drifting votes are inevitable due to the subjective nature of the experiment. Future research can take the drift distribution profiles reported here as a reference, and use them to check for similarities in less controlled environments, like online crowd-sourcing platforms.

Fig. 5. The characters generated by simultaneously fully minimizing (left)

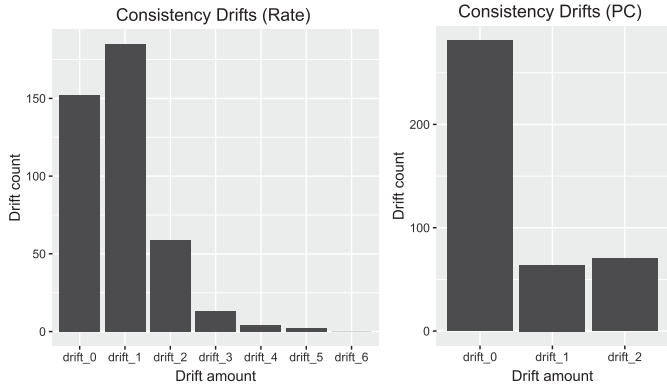

Fig. 6. The count of inconsistent votes for each possible drift value. For Rate (left) and PC (right) modes. Each column counts the drifts accumu-

TABLE 2
Experiment 1: Drifts for the Consistency Check in Rate Mode

	Drift Amount						
	0	1	2	3	4	5	6
Drift Count	152	185	59	13	4	2	0
Mean per subject	1.83	2.23	0.71	0.16	0.05	0.02	0
SD per subject	1.36	1.24	0.8	0.37	0.21	0.15	0

TABLE 3
Experiment 1: Drifts for the Consistency Check in PC Mode

	Drift Amount in PC		
	0	1	2
Drift Count	281	64	70
Mean per subject	3.38	0.77	0.84
SD per subject	1.21	0.97	1.02

3.8 Time Analysis

Fig. 7 shows the plots of the time needed for voting in both Rate and PC modes. The plots show all the trials by all users, except the last trial that could not be measured for technical reasons. Hence, the remainder of this section is based on the analysis of $83 * 54 = 4482$ trials.

In Rate mode, the average voting time was 5.2s (sd = 2.70), while in PC mode it was 5.77s (sd = 3.55). A Wilcoxon signed-rank test suggests that we can discard the null hypothesis that the two distributions have the same mean (W = 3008828, $p < 0.001$). We can hence state that the time to vote in PC mode is statistically significantly higher than the time needed for voting in Rate mode, by 13.2 percent.

We also analyzed the effect of counter-balancing. The categorical condition *mode-first* can be either *Rate-first* or *PC-first*. Out of our 83 subjects, 42 voted in Rate-first and 41 in PC-first. When considering the Rate voting mode, in the Rate-first condition the average voting time was 5.08s (sd = 2.75), while in PC-first the average time was 5.11s (sd = 2.64). A Mann-Whitney U test shows that the difference is not statistically significant (U = 2456899.5, p = 0.104). The voting speed for the Rate mode was independent from the first voting condition.

However, when considering the PC voting mode, in the Rate-first condition the average voting time was 4.95s (sd = 2.95), while in PC-first the average time was 6.61s (sd = 3.90), and a Mann-Whitney U test showed a statistically significant difference (U = 1746920, $p < 0.001$). Hence, it seems that when users start voting in PC as the first mode, they need more time to get used to the task of judging the pictures.

To better understand the behavior of voters in the combination mode = PC and mode-first = PC-first, we plotted the voting times of the 41 users involved, grouped by trial. As can be seen in Fig. 8, trial-by-trial the voting time decreases, suggesting a "learning effect".

Hence, we analyzed the behavior of subjects in the "long run", during the second voting session, by comparing two groups: a first group in mode = Rate, mode-first = PC-first and a second group with mode = PC, mode-first = Rate-first. In Rate mode the voting time was on average 5.11s (sd=2.64), while in PC mode the average voting time was 4.95s (sd=2.95). A Mann-Whitney U test shows a significant difference (U=2308182.5, $p < 0.001$) between the averages. Hence, when considering the second session only, the PC voting time is 3.2 percent faster than the Rate mode.

3.9 Learning Effect Duration

As pointed out in the previous section, and visible in Fig. 8, when subjects start voting in PC mode, their voting time progressively decreases. This is likely due to a *learning effect* on the judgment procedure. Since we did not find in the literature any well established method to measure the duration of the learning phase, we elaborated the strategy reported in Algorithm 1.

Algorithm 1. The Algorithm for Estimating the Trial Numbers Marking the End of the Learning Effect

Input: T, Number of trials.
Output: A list of integers, where each element is a candidate trial number indicating the end of the learning phase.
Data: trialTimes(), a function taking a list of trial numbers and returning their time measurements.
Data: MWUTest(), a function performing a Mann-Whitney U test and returning its resulting p-value.

```
1: learnEndTrials ← [] /* Output Accumulator */
2: for t ← 2 to T do
     /* Divide the trials into two groups: T1 and
        T2 */
3:   T1   [1, ..., t-1]
4:   T2   [t, ..., T]
     /* Check if the two groups have significantly
        different durations */
5:   p1 ← MWUTest(trialTimes(T1), trialTimes(T2))
6:   if p1 < 0.05 then
       /* Split T2 into T3 and T4 */
7:     tt ← ⌊(t + T)/2⌋
8:     T3   [t, ..., tt-1]
9:     T4   [tt, ..., T]
       /* If T3 and T4 have NOT a significant diff. */
10:    pp   MWUTest(trialTimes(T3), trialTimes(T4))
11:    if pp >= 0.05 then
         /* Mark t as the end of a learning
            phase */
12:      learnEndTrials   learnEndTrials + [t]
13:    end
14:  end
15: end
16: return learnEndTrials
```

The algorithm is based on the assumption that if a trial number t determines the end of a learning phase, there are: i) a significant difference in duration between the trials before t compared to the trials after t, and ii) the durations after t stabilize. The comparison between trial durations is done with a Mann-Whitney U test.

Applied on our set of 55 trials, the algorithm returned `learnEndTrials=[30,31]`, suggesting that the voting time decreased approximately until tasks 30 and 31 and then stabilized. Additionally, the voting times after the learning effect are comparable between Rate and PC modes.

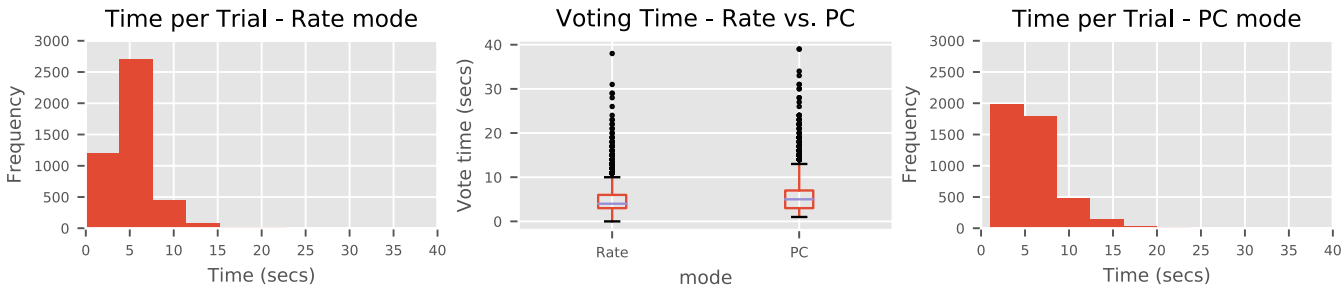

Fig. 7. The frequency distribution of the time needed to accomplish the voting trials in Rate mode (left) and PC mode (right). In the middle, a side-by-side comparison using boxplots.

Since the PC mode requires watching two pictures before expressing a judgement, we would have expected a higher speed for the Rate mode. However, it is possible that the time needed to judge two pictures is compensated for by the shorter time needed to choose a preference (left, none, right) rather than choosing an absolute value on the scale.

3.10 Subjective Preference

After performing the two judgement tasks (PC and Rate modes), subjects answered the following question: "In your opinion, in which of the two modalities was it easier to judge the characters?" Fifty-three out of 83 subjects (63.8 percent) expressed a preference for the Paired Comparison mode. A chi-squared test for goodness of fit against the uniform distribution (50-50 percent) shows a statistically significant difference in the preference tendency ($\chi = 6.37, p = 0.016$).

Although still far from a strong majority, this percentage might suggest to future researchers to adopt the PC method to make user studies more appealing for the subjects.

To analyze the effect of counter balancing, we performed a Pearson's chi-squared test on the contingency table shown in Table 4, which groups the subjects according to both their preference and their first mode. The result of the test ($\chi = 5.89e-01, p = 0.442$) suggests that the preference vote is not influenced by the voting order.

3.11 Model Selection

A model selection is the process of analyzing the votes collected from all 83 participants in order to extract the subset of attributes that contribute the most to the perception of dominance. The analysis started by performing a linear regression between the character's attributes (predictors) and the dominance value (measured variable).

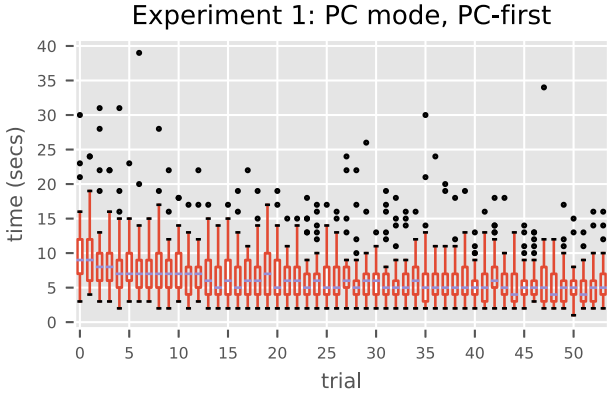

Fig. 8. The time needed to accomplish each trial in PC mode, as first experiment (mode-first=PC-first). The x axis reports the trial number, the y axis the completion time. Each boxplot summarizes the time measured from the 41 subjects.

For the Rate mode, the linear regression took into account 4,150 measurements (50 votes per 83 subjects), where each measurement mapped the values of the attributes of each virtual character with its voted dominance value, in the range $[-3, +3]$.

For the PC mode, the comparison data needed first to be converted into *estimates*, which is a value of perceived dominance on a scale relative to the other pictures. Then, the linear regression was conducted on 100 measurements, mapping physical attribute values of virtual characters into dominance estimates. We analyzed our PC data in R (https://www.r-project.org/), using the `prefmod` package (https://www.jstatsoft.org/article/view/v048i10[42], [43]) to compute the estimates and the `sirt` package (https://rdrr.io/cran/sirt/[44]) to computate the α and G indices. Both packages compute the estimates from PC data using the Bradley-Therry model [23]. The analysis resulted in: estimate $min/max = -1.0677/1.1330$, reliability index $\alpha = 0.9704$, and separation index $G = 5.8216$. Both α and G indicate a very high consistency of the data. The estimate value was fixed to 0 for the 100th picture, while all the other values spread around 0 of an amount that reflects the level of agreement between the judges.

The two linear models are then reduced via *backward elimination* via *p-minimization*: an iterative technique used to select only the most significant predictors. The model selection considers the p-value associated to each variable (as a result of the linear regression) and discards the variable with the highest p-value above a threshold α. The algorithm iterates until there are no variables with p-value $\geq \alpha$. In this work, we used $\alpha = 0.05$.

Table 5 shows the models selected for both Rate and PC voting modes. The bottom line of the table reports the final

TABLE 4
Contingency Table for Preference

		first-mode		
		PC-first	Rate-first	
Preferred	PC	24	29	53
	Rate	17	13	30
		41	42	83

TABLE 5
The Two Models Selected From the Votes Collected in Both Rate and PC Mode

Attribute	Voting mode	
	Rate	PC
chin/chin-bones-in | out	*	-
chin/chin-height-min | max	***	*
eyebrows/eyebrows-angle-up | down	***	***
eyes/r-eye-size-small | big	-	-
head/head-oval	-	***
macrodetails-height/Height	***	***
macrodetails-universal/Muscle	**	***
macrodetails-universal/Weight	-	-
mouth/mouth-scale-horiz-incr | decr	-	-
mouth/mouth-scale-vert-incr | decr	-	-
neck/neck-scale-horiz-less | more	***	***
nose/nose-scale-horiz-incr | decr	***	*
stomach/stomach-tone-decr | incr	-	-
torso/torso-vshape-less | more	***	***
Selected	8	8
Adjusted R^2	0.2427	0.8435

*The inner cells report the significance resulting from the final linear fitting (***= $p < 0.001$, **= $p < 0.01$, *= $p < 0.05$). Where a dash (−) appears, the attribute was excluded from the model.*

adjusted squared correlation factor. The linear models were computed with the `lm` function of the `R` language (v3.1).

It is worth noticing that the much higher correlation factor of the PC mode does not indicate a better performance. In fact, in the Rating mode, the linear regression is run against the raw voting data, with all its variance. For the PC mode, however, the linear regression is run on the estimated dominance values, whose variance has already been absorbed by the computation of the estimates and expressed in terms of reliability.

3.12 Discussion

For boths modes, eight attribute were retained. For the Rate voting mode, they are: chin bones protrusion, chin height, eyebrows inclination, height, muscularity, neck width, nose width, and torso V-shape. In PC voting mode, the selected model does not include the chin bones' protrusion but adds the head ovality. For both modes, four attributes relate to the full bodies (neck width, heights, muscularity, and torso V-shape) and the remaining four to the face.

Concerning the body, height is not surprisingly correlated to dominance. However, we were expecting a stronger correlation for height in the PC mode with respect to the Rate mode, mainly because in PC mode the characters are shown side-to-side. Possibly, even in rate mode subjects are able to perceive different heights thanks to the fixed camera position among images. We found no previous work directly measuring the correlation between full body features and dominance, but both muscularity, neck size, and torso V-shape (i.e., the width of the shoulder) are intuitively correlated to a measure of masculinity, which is in fact perceived through the same attributes of dominance [17].

For the face, the contribution of eyebrows angle, head ovality, and chin bones confirm previous finding (see Section 2.1). For the nose width, we couldn't find any previous work directly associating it to dominance perception. Finally, it is not straightforward to explain why the modulation of head ovality (in practice, widening the full face) was noticed only in the PC mode, while the modulation of chin bones (in practice raise the height of the jaw) are more influencing in the Rate mode.

Fig. 9. Example of two inconsistently rated characters.

Since the two models are not perfectly identical, we compared them with a further verification test, presented in the next section.

To get a glimpse on the connection between the physical attributes influencing the perception of dominance and the discrepancies in consistency checks (Section 3.7), we visually analyzed the (pair of) individuals with highest discrepancies and report here some examples. Fig. 9 shows two characters that have been rated with a drift of 5. Not surprisingly, it looks like both characters present at the same time features characterizing both high and low dominance. The character on the left presents very inclined eyebrows, but is also very short. For the the character on the right, the face presents less elements of dominance, but the body is tall, muscular, and V-shaped.

Fig. 10 shows a pair that has been voted at opposite preference (drift 2) during the check. The character on the left

Fig. 10. Example of one inconsistently ranked pair.

Fig. 11. The four characters used for the verification experiment.

has inclined eyebrows and wide nose. The character on the right is slightly shorter, but eyebrows are slightly more inclined and the jaw more prominent. In this case, we could observe that the two characters present a disjoint set of features influencing dominance. This suggests that, between the two voting trials, the same subject was influenced by a different sub-set of physical attribute.

A careful numerical analysis on the (pair of) characters which led to inconsistent decision could help understanding in which conditions the combination of physical attributes lead to ambiguous judgements.

3.13 Verification Experiment

In order to verify whether the two models derived by the previous experiment lead to different judgements of dominance, we ran a verification experiment using only 4 virtual characters, illustrated in Fig. 11, which were created by maximizing and minimizing all of the attributes of the two selected models. The hypotheses are that: i) the two dominant stereotypes (`Rate-dom` and `PC-dom`) are rated higher than the two submissive ones (`Rate-sub` and `PC-sub`), and ii) one of the two voting system leads to more expressive characters (the meaning of "expressive" being the ability to generate characters perceived as more dominant or more submissive).

One month after the end of the previous experiment, 73 new subjects (52 male, 21 female), on average 25.7 years old (sd = 5.0), voted for the 4 virtual characters in Rate mode and for 4 random pairs in PC mode, in counterbalanced order. Thirty-three of them were German, eight Indian, six Greek, five Chinese, four Mexican, three Russian, and the others from a mix of eight different nationalities. The experiment took place just after another unrelated, longer study that asked each participant for judgments of 50 virtual characters. Hence, we assume that the subjects were already well acquainted with the voting task and any learning effect was exhausted.

TABLE 6
Results of the Verification Experiment: p-values of the Tests Comparing Couples of Stereotype Generated Characters

Pair	Rate comparison	PC comparison
*-dom vs. *-sub	$< 2e-16$ (***)	$< 2e-16$ (***)
Rate-dom vs. PC-dom	0.0211 (*)	0.620
Rate-sub vs. PC-sub	0.0629	0.2607

For the rating session, we compared the scores among all possible pairs via a Mann-Whitney U test and recorded the p-valuess. For the PC session, we extracted the p-values directly from the `prefmod` output on the regression statistics. Table 6 summarizes the results. Our analyses confirmed the expectation that the two dominant characters were significantly perceived as more dominant than the two submissive characters. However, we verified that, when rated, among the two most dominant characters, the one generated from the PC model appears more submissive than the other (Rate-dom mean rate = 2.0, sd = 1.13; PC-dom mean rate = 1.74, sd = 1.08), but with a mild significance of 0.0211. This statistical difference doesn't emerge in the PC voting mode nor for the submissive stereotypes.

Overall, this verification experiment suggest that the PC voting mode is as good as the well-established Rate voting mode in identifying the facial and body cues contributing in the perception of dominance.

4 Conclusion and Possible Extensions

This paper presented a direct comparison between the Rating and the Paired Comparison voting modes in judging the perception of dominance in virtual characters. The study consisted of two experiments. The first used the two voting modes to gather human judgements and to build virtual stereotypes of submissive and dominant characters. The second, a verification experiment, assessed that the two voting modes led to two equivalently expressive models of dominance.

Even though they should be experimentally verified, the method and the results of this work are likely applicable to any personality trait and to different subjective preference

criteria that can be judged from first impressions, such as beauty, attractiveness, scariness, and the like.

Further analysis of the voting data shows a learning effect in the PC mode that is exhausted after about 30 votes. After that, the judging speed is similar between the two modes, possibly because the longer time needed to view and judge two pictures is compensated for by a quicker three-way decision (left, don't know, right) versus choosing a score between 1 and 7. This last hypothesis can not be verified with the data collected in this experiment, but could be confirmed by monitoring user activity with eye-tracking devices.

The voting sessions included stimuli repetitions, which was intended to check for the reliability of voters. The data shows a significant and systematic inconsistency, making it impossible to arbitrarily establish thresholds for exclusions of voters. Given the relatively controlled nature of the voting sessions, we can only claim that the inconsistency profiles reported in this work are natural expressions of human variability in judgements. Future research might consider these profiles as bottom lines for a consistency check when running experiments in non-controlled environments, such as online crowd-sourcing platforms.

Finally, subjective questionnaires reported a higher (63.8 percent) preference for the paired comparison mode, suggesting consideration of PC as a valid alternative for the sake of subjects' satisfaction. Still, it might be preferred to use the ranking mode when time is a strong constraint and the number of votes lies below 30 trials.

Overall, these results suggest that pairwise comparison can be considered as a valid alternative in future research involving human judgements of virtual characters. Further research is needed to establish if the results presented in this paper can be extended to animated (interactive) characters and to other media and stimuli, such as the judgement of videos, music, and text.

A limitation of this study can be seen in the fact that, in contrast to many recent criticisms [2], [3], in the models selection (Section 3.11) the rating scores (1 to 7) are used to fit a linear model, thus assuming an equal distancing among the ordered rating labels. To deal with this issue, the analysis could be conducted by transforming the ratings into ranked representations, and then produce the classification models. Symmetrical, a second model can be constructed on dominance estimates, by using ratings as they are and PCs converted to estimates. These two models could then be compared.

Acknowledgments

The authors would like to thank our assistant Nicolas Erbach for his precious assistance both with the software implementation and during the user studies, and the reviewers for their useful and constructive comment. The research reported in this paper was partially supported by the "Multimodal Computing and Interaction" Cluster of Excellence at Saarland University.

References

[1] R. Likert, "A technique for the measurement of attitudes," *Archives Psychol.*, vol. 22, no. 140, 1932, Art. no. 55.

[2] H. Martinez, G. Yannakakis, and J. Hallam, "Don't classify ratings of affect; rank them!," *IEEE Trans. Affective Comput.*, vol. 5, no. 3, pp. 314–326, Jul.–Sep. 2014.

[3] G. N. Yannakakis and H. P. Martínez, "Ratings are overrated!" *Front. ICT*, vol. 2, Jul. 2015. [Online]. Available: http://journal.frontiersin.org/Article/10.3389/fict.2015.00013/abstract

[4] F. Nunnari and A. Heloir, "Generating virtual characters from personality traits via reverse correlation and linear programming," in *Proc. 16th Int. Conf. Auton. Agents Multiagent Syst.*, 2017, pp. 1661–1663.

[5] F. Nunnari and A. Heloir, "Generation of virtual characters from personality traits," in *Proc. 17th Int. Conf. Intell. Virtual Agents*, 2017, pp. 301–314.

[6] L. P. Naumann, S. Vazire, P. J. Rentfrow, and S. D. Gosling, "Personality judgments based on physical appearance," *Pers. Soc. Psychol. Bull.*, vol. 35, no. 12, pp. 1661–1671, Sep. 2009. [Online]. Available: http://psp.sagepub.com/cgi/doi/10.1177/0146167209346309

[7] R. R. McCrae and O. P. John, "An introduction to the five-factor model and its applications," *J. Pers.*, vol. 60, no. 2, pp. 175–215, Jun. 1992. [Online]. Available: http://doi.wiley.com/10.1111/j.1467-6494.1992.tb00970.x

[8] H. J. Smith and M. Neff, "Understanding the impact of animated gesture performance on personality perceptions," *ACM Trans. Graph.*, vol. 36, no. 4, pp. 1–12, Jul. 2017. [Online]. Available: https://dl.acm.org/doi/10.1145/3072959.3073697

[9] N. N. Oosterhof and A. Todorov, "The functional basis of face evaluation," *Proc. Nat. Acad. Sci. United States America*, vol. 105, no. 32, pp. 11 087–11 092, 2008. [Online]. Available: http://www.pnas.org/content/105/32/11087.abstract

[10] R. B. Cattell, H. W. Eber, and M. Tatsuoka, *Handbook for the 16 Personality Factor Questionnaire*. Champaign, IL, USA: Institute for Personality and Ability Testing, 1970.

[11] C. A. M. Sutherland, J. A. Oldmeadow, I. M. Santos, J. Towler, D. M. Burt, and A. W. Young, "Social inferences from faces: Ambient images generate a three-dimensional model," *Cognition*, vol. 127, no. 1, pp. 105–118, 2013. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0010027712002739

[12] T. Doce, J. Dias, R. Prada, and A. Paiva, "Creating individual agents through personality traits," in *Proc. Int. Conf. Intell. Virtual Agents*, 2010, vol. 6356, pp. 257–264. [Online]. Available: http://www.springerlink.com/index/10.1007/978-3-642-15892-6_27

[13] F. Durupinar, M. Kapadia, S. Deutsch, M. Neff, and N. I. Badler, "PERFORM: Perceptual approach for adding OCEAN personality to human motion using laban movement analysis," *ACM Trans. Graph.*, vol. 36, no. 1, pp. 1–16, Oct. 2016. [Online]. Available: http://dl.acm.org/citation.cfm?doid=2996392.2983620

[14] S. Streuber *et al.*, "Body talk: Crowdshaping realistic 3D avatars with words," *ACM Trans. Graph.*, vol. 35, no. 4, pp. 54:1–54:14, Jul. 2016. [Online]. Available: http://doi.acm.org/10.1145/2897824.2925981

[15] R. J. W. Vernon, C. A. M. Sutherland, A. W. Young, and T. Hartley, "Modeling first impressions from highly variable facial images," *Proc. Nat. Acad. Sci. United States America*, vol. 111, no. 32, pp. E3353–E3361, 2014. [Online]. Available: http://www.pnas.org/content/111/32/E3353.abstract

[16] H. Toscano, T. Schubert, and A. N. Sell, "Judgments of dominance from the face track physical strength," *Evol. Psychol.*, vol. 12, no. 1, pp. 1–18, 2014. [Online]. Available: http://urn.nb.no/URN:NBN:no-44049

[17] S. Windhager, K. Schaefer, and B. Fink, "Geometric morphometrics of male facial shape in relation to physical strength and perceived attractiveness, dominance, and masculinity," *Amer. J. Hum. Biol.*, vol. 23, no. 6, pp. 805–814, 2011. [Online]. Available: http://dx.doi.org/10.1002/ajhb.21219

[18] C. F. Keating, A. Mazur, and M. H. Segall, "A cross-cultural exploration of physiognomic traits of dominance and happiness," *Ethol. Sociobiol.*, vol. 2, no. 1, pp. 41–48, Jan. 1981. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/0162309581900212

[19] C. Batres, D. E. Re, and D. I. Perrett, "Influence of perceived height, masculinity, and age on each other and on perceptions of dominance in male faces," *Perception*, vol. 44, no. 11, pp. 1293–1309, Nov. 2015. [Online]. Available: http://journals.sagepub.com/doi/10.1177/0301006615596898

[20] G. T. Fechner, *Elemente der Psychophysik*. Leipzig, Germany: Breitkopf und Härtel, 1860.

[21] L. L. Thurstone, "A law of comparative judgment," *Psychol. Rev.*, vol. 34, no. 4, pp. 273–286, 1927. [Online]. Available: http://content.apa.org/journals/rev/34/4/273

[22] L. Guttman, "An approach for quantifying paired comparisons and rank order," *The Ann. Math. Statist.*, vol. 17, no. 2, pp. 144–163, 1946. [Online]. Available: http://www.jstor.org/stable/2236035

[23] R. A. Bradley and M. E. Terry, "Rank analysis of incomplete block designs," *Biometrika*, vol. 39, pp. 324–335, 1952.
[24] T. Bramley, "Paired comparisons methods," in *Techniques for Monitoring the Comparability of Examination Standards*. London, UK: Qualification and Authority, 2007, pp. 246–294. [Online]. Available: https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/487059/2007-comparability-exam-standards-i-chapter7.pdf
[25] M. G. Kendall and B. B. Smith, "On the method of paired comparisons," *Biometrika*, vol. 31, no. 3/4, pp. 324–345, 1940. [Online]. Available: http://www.jstor.org/stable/2332613
[26] H. A. David, *The Method of Paired Comparisons*, vol. 12. London, United Kingdom: Hodder Arnold, 1988.
[27] T. C. Brown and G. L. Peterson, "An enquiry into the method of paired comparison: Reliability, scaling, and thurstone's law of comparative judgment," U.S. Department of Agriculture, Forest Service, Rocky Mountain Research Station, Fort Collins, CO, Tech. Rep. RMRS-GTR-216 WWW, 2009.
[28] J. V. Verkuilen, *Regression Models for Paired Comparisons*. Ann Arbor, MI, USA: ProQuest, 2007.
[29] M. Garner and G. Engelhard, "Rasch measurement theory, the method of paired comparisons, and graph theory," *Objective Meas., Theory Pract.*, vol. 5, pp. 259–286, 2000.
[30] S. Heldsinger and S. Humphry, "Using the method of pairwise comparison to obtain reliable teacher assessments," *The Australian Educ. Researcher*, vol. 37, no. 2, pp. 1–19, 2010. [Online]. Available: http://dx.doi.org/10.1007/BF03216919
[31] M. Maida, K. Maier, and N. Obwegeser, "Pairwise comparison techniques for preference elicitation: Using testretest reliability as a quality indicator," in *Proc. Int. Conf. Inf. Resour. Manage.*, 2012. [Online]. Available: http://aisel.aisnet.org/confirm2012/65/
[32] M. Garner and G. Engelhard, "Using paired comparison matrices to estimate parameters of the partial credit Rasch measurement model for rater-mediated assessments," *J. Appl. Meas.*, vol. 10, no. 1, pp. 30–41, 2009.
[33] G. J. Buhyoff and W. A. Leuschner, "Estimating psychological disutility from damaged forest stands," *Forest Sci.*, vol. 24, no. 3, pp. 424–432, 1978. [Online]. Available: http://www.ingentaconnect.com/content/saf/fs/1978/00000024/00000003/art 00016
[34] T. C. Brown, D. Nannini, R. B. Gorter, P. A. Bell, and G. L. Peterson, "Judged seriousness of environmental losses: Reliability and cause of loss," *Ecol. Econ.*, vol. 42, no. 3, pp. 479–491, Sep. 2002. [Online]. Available: http://linkinghub.elsevier.com/retrieve/pii/S0921800902001714
[35] Y. Yang and H. H. Chen, "Ranking-based emotion recognition for music organization and retrieval," *IEEE Trans. Audio, Speech, Lang. Process.*, vol. 19, no. 4, pp. 762–774, May 2011.
[36] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "LIRIS-ACCEDE: A video database for affective content analysis," *IEEE Trans. Affective Comput.*, vol. 6, no. 1, pp. 43–55, Jan.–Mar. 2015.
[37] Y. Baveye, E. Dellandréa, C. Chamaret, and L. Chen, "From crowdsourced rankings to affective ratings," in *Proc. IEEE Int. Conf. Multimedia Expo Workshops*, 2014, pp. 1–6.
[38] R. Lotfian and C. Busso, "Practical considerations on the use of preference learning for ranking emotional speech," in *Proc. IEEE Int. Conf. Acoust. Speech Signal Process.*, 2016, pp. 5205–5209. [Online]. Available: http://ieeexplore.ieee.org/document/7472670/
[39] S. Parthasarathy and C. Busso, "Preference-learning with qualitative agreement for sentence level emotional annotations," in *Proc. Annu. Conf. Int. Speech Commun. Assoc.*, 2018, pp. 252–256. [Online]. Available: http://www.isca-speech.org/archive/Interspeech_2018/abstracts/2478.html
[40] M. Karpinska-Krakowiak, "Ratings or pairwise comparisons? An experimental study on scale usability," *Econ. Environ. Stud.*, vol. 18, no. 46, pp. 653–664, Jun. 2018. [Online]. Available: http://www.ees.uni.opole.pl/content/02_18/ees_18_2_fulltext_11.pdf
[41] I. Wood, J. McCrae, V. Andryushechkin, and P. Buitelaar, "A comparison of emotion annotation approaches for text," *Information*, vol. 9, no. 5, May 2018, Art. no. 117. [Online]. Available: http://www.mdpi.com/2078–2489/9/5/117
[42] R. Dittrich and R. Hatzinger, "Fitting loglinear Bradley-Terry models (LLBT) for paired comparisons using the R package prefmod," *Psychol. Test Assessment Model.*, vol. 51, no. 2, 2009, Art. no. 216.
[43] R. Hatzinger and R. Dittrich, "Prefmod: An R package for modeling preferences based on paired comparisons, rankings, or ratings," *J. Statist. Softw.*, vol. 48, no. 10, pp. 1–31, 2012. [Online]. Available: http://www.jstatsoft.org/v48/i10
[44] D. R. Hunter, "MM algorithms for generalized Bradley-Terry models," *The Ann. Statist.*, vol. 32, no. 1, pp. 384–406, 2004. [Online]. Available: http://www.jstor.org/stable/3448514

Fabrizio Nunnari received the master's degree in computer science from the University of Torino, Italy, in 2001, and the PhD degree on the use of 3D data visualization for collaborative work from the University of Torino, Italy, in 2005. He is currently a senior researcher with the German Research Center for Artificial Intelligence (DFKI) in Saabrücken, Germany. Between 2006 and 2012, he worked as researcher and lead developer at the Virtual Reality and Multimedia Park in Torino, Italy. Since 2020, he is part of the Affective Computing group and works on the generation of interactive virtual humans and on the synthetic animation of virtual interpreters for sign language.

Alexis Heloir received the master's degree in computer science from the University of Lille, France, in 2004, and the PhD degree from the University of Southern Brittany, France, in 2008 on the specification and design of intelligible signing avatars. He is currently an assistant professor with the Université Polytechnique Hauts de France – LAMIH Laboratory (UMR CNRS 8201). He collaborated with the German Research Center for the Artificial Intelligence (DFKI) since 2008. Between 2012 and 2017, he lead the SLSI independant research group at Multimodal Computing Excellence Cluster in Saarbrücken. His current research interests include 3D and tangible user interfaces, the control and the animation of embodied conversational agents as well as the generation of intelligible sign language utterances using avatars.