




How does familiarity with a voice affect trait judgements?

Nadine Lavan^{1*} , Mila Mileva^{2,3} and Carolyn McGettigan^{1*}

¹Department of Speech, Hearing and Phonetic Sciences, University College London, UK

²Department of Psychology, University of York, UK

³School of Psychology, University of Plymouth, UK

From only a single spoken word, listeners can form a wealth of first impressions of a person's character traits and personality based on their voice. However, due to the substantial within-person variability in voices, these trait judgements are likely to be highly stimulus-dependent for unfamiliar voices: The same person may sound very trustworthy in one recording but less trustworthy in another. How trait judgements differ when listeners are familiar with a voice is unclear: Are listeners who are familiar with the voices as susceptible to the effects of within-person variability? Does the semantic knowledge listeners have about a familiar person influence their judgements? In the current study, we tested the effect of familiarity on listeners' trait judgements from variable voices across 3 experiments. Using a between-subjects design, we contrasted trait judgements by listeners who were familiar with a set of voices – either through laboratory-based training or through watching a TV show – with listeners who were unfamiliar with the voices. We predicted that familiarity with the voices would reduce variability in trait judgements for variable voice recordings from the same identity (cf. Mileva, Kramer & Burton, *Perception*, 48, 471 and 2019, for faces). However, across the 3 studies and two types of measures to assess variability, we found no compelling evidence to suggest that trait impressions were systematically affected by familiarity.

We can rapidly form first impressions about the personality of unfamiliar others just by hearing their voice or seeing their face. These first impressions have been shown to follow two fundamental dimensions – trustworthiness (valence) and dominance (McAleer, Todorov & Belin, 2014; Oosterhof & Todorov, 2008). While the accuracy of trait judgements in relation to an individual's true character, ability, or personality is low at best (Klofstad & Anderson, 2018; Olivola & Todorov, 2010; Todorov et al., 2015), trait judgements have been shown to be consistent across different raters. This suggests that they reflect general, stereotyped aspects of person perception (McAleer, *et al.*, 2014; Todorov, Said, Engell, Oosterhof, 2008). Even though first impressions tend to not be accurate, they are important because they have been shown to predict behaviour and influence decision-making (see Olivola, Funk, & Todorov, 2014, for a review) in a number

This is an open access article under the terms of the Creative Commons Attribution License, which permits use, distribution and reproduction in any medium, provided the original work is properly cited.

*Correspondence should be addressed to Nadine Lavan or Carolyn McGettigan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, UK.
(emails: n.lavan@ucl.ac.uk; c.mcgettigan@ucl.ac.uk).

of different contexts, such as election outcomes (Ballew & Todorov, 2007; Klofstad, 2016; Klofstad, Anderson, & Peters, 2012; Mileva, Tompkinson, Watt, & Burton, 2020; Sussman, Petkova, & Todorov, 2013; Todorov, Mandisodza, Goren, & Hall, 2005), applicant success at job interviews (see Harris & Garris, 2008, for a review), and court sentencing (Wilson & Rule, 2015; Zebrowitz & McDonald, 1991).

Recent work in the face perception literature has shown that explicit trait judgements are largely stimulus-dependent when using variable images (i.e., images including different emotional expressions, hairstyles, lighting, and viewpoints), further underlining the limited accuracy of these judgements for unfamiliar identities. Here, trait judgements for different images of the same unfamiliar face vary substantially: The same person may look very trustworthy in one image but rather untrustworthy in another. Indeed, the degree of within-person variability in judgements often is on par with or exceeds the degree of between-person variability (e.g., Mileva, Young, Kramer, & Burton, 2019; Sutherland, Young & Rhodes, 2017; Todorov & Porter, 2014; but Mahrholz, Belin & McAleer, 2018, for evidence of stable trait judgements across different stimuli [read words; read sentences]).

This within-person variability is not restricted to the visual domain but is also a prominent feature in human voices, such that the same person's voice can sound dramatically different from situation to situation (e.g., shouting over background noise, singing, or laughing; Lavan, Burton, Scott & McGettigan, 2019). Within-person variability has been shown to dramatically affect perceptual judgements of voice identity when listeners were not familiar with a voice: In a series of voice sorting studies, unfamiliar listeners were unable to accurately perceive speaker identity from naturally varying voice recordings (i.e., excerpts taken from across a television series; Lavan, Burston & Garrido, 2019; Lavan, Burston, Ladwa, Merriman, Knight & McGettigan, 2019; Lavan, Merriman, Ladwa, Burston, Knight & McGettigan, 2019). Specifically, they perceived variable recordings of the same voice identity as a number of different people, thus misinterpreting within-person variability as between-person variability. Strikingly, the effects of within-person variability on voice identity perception were much reduced when listeners were familiar with the voices: Familiar listeners were able to accurately perceive recordings of the same person as a single identity, despite the substantial within-person variability. These differences in behaviour between the two groups have been ascribed to listeners having access to stable and robust representations of familiar voices, which enables them to link variable stimuli back to a single identity (e.g., Burton, Kramer, Ritchie & Jenkins, 2016; Lavan, Burton *et al.*, 2019; Lavan, Knight & McGettigan, 2019, for faces).

Based on the evidence above, it could also be predicted that trait judgements for familiar and unfamiliar voices should differ from each other, with listeners being less susceptible to the effects of within-person variability when dealing with familiar voices. Evidence to suggest that this is the case for faces has recently been reported by Mileva, Kramer and Burton (2019). In this study, the authors presented participants with 4 images of each of 40 A-list and 40 foreign celebrities (representing familiar and unfamiliar identities, respectively) and collected judgements for 5 social traits (trustworthiness, dominance, attractiveness, distinctiveness, and extraversion). Using Procrustes analyses, the authors showed that the variability in judgements in their 5-dimensional social trait space for familiar faces is indeed smaller compared to the variability in judgements for images of unfamiliar faces. Trait perception from familiar identities thus indeed appears to be less vulnerable to the effects of within-person variability.

In the current study, we asked if and how familiarity with a person would affect trait judgements from voices. Specifically, we asked whether trait judgements for familiar

voices would be less variable than trait judgements for unfamiliar voices, since familiar listeners should be able to have a more stable percept of a talker's identity and their associated trait attributes. We furthermore asked which aspects of familiarity could affect trait perception: Can familiarity alone, in the absence of any semantic (valenced) knowledge, affect judgements or is semantic knowledge about a person essential? Are laboratory-based training paradigms sufficient or are more naturalistic learning environments required? To answer these questions, we conducted a series of three experiments in which we manipulated the types of familiarity listeners had with the voices (familiarized through laboratory-based training [Experiments 1 and 2] vs previously familiar through watching a TV show [Experiment 3]) as well as the type of knowledge familiar listeners had access to (no semantic knowledge [Experiments 1 and 2] vs semantic knowledge [Experiments 1 and 3]).

Experiment 1

In Experiment 1, we sought to contrast trait judgements of listeners who had been trained to recognize a set of voices with trait judgements of listeners who were unfamiliar with the voices. We additionally aimed to assess the influence of semantic knowledge about a person on trait judgements made by familiar(ized) listeners. For this purpose, we collected trait judgements from a group of listeners that received no training, thus rating the voices they were completely unfamiliar with ('No Training'). We furthermore trained 3 groups of listeners to recognize a set of voices in a between-subjects design where each group completed different versions of a voice learning task. One group of listeners learned to recognize the different voices only by associating them with a name – thus modelling a case of 'familiarity only', gained in the absence of semantic knowledge about the person beyond their name ('Neutral Training'). The other two groups completed 'valenced' training paradigms, where listeners learned to recognize the voices by name, with each training stimulus being accompanied with either a positively or negatively valenced vignette ('Positive Training' and 'Negative Training', respectively). These vignettes were used as a model for the semantic knowledge that is usually acquired when becoming familiar with a person in naturalistic settings (e.g., through social interactions). All listeners then completed a trait judgement task, providing trustworthiness ratings for novel voice recordings of the 4 identities that were part of the training paradigms for the three familiar listener groups. Here, we focused on trustworthiness as it is considered the primary dimension of social evaluation (over, e.g., dominance; Cuddy, Fiske & Glick, 2008; Sutherland, Rhodes, Burton & Young, 2019).

Following our prediction that familiarity (and semantic knowledge) should affect the perception of traits from voices, we expected that 1) valenced learning should lead to an overall shift in trustworthiness perceptions relative to neutral or no training (manipulation check) and 2) variability should be reduced for familiar listeners in the valenced training groups compared to the unfamiliar listeners in the No Training group. We had no specific directional hypothesis for how variability in judgements would be affected in the Neutral Training group: If familiarity alone – that is, the formation of an identity-specific representation – is sufficient to reduce variability in the absence of semantic knowledge, the Neutral Training group should behave in similar ways to the valenced training paradigms. If familiarity alone is not sufficient, the variability of judgements in this group should be more similar to the unfamiliar listeners in the No Training group.

Method

Participants

A total of 124 participants (*mean age* = 27.1 years, *SD* = 6.5 years, 74 female) were included in the final sample for this study (31 participants \times 4 training groups). This sample size was deemed appropriate based on the sample sizes usually used for studies of trait perception in the face and voice perception literature. Before arriving at this final sample, 12 participants were excluded based on preregistered exclusion criteria: 6 failed the vigilance trials (see *Materials and Procedure*), five did not learn to recognize the voices with sufficient accuracy for our cut-off of 50% correct at the end of training (chance = 25%; see *Materials and Procedure*), and one participant's judgements on the main task were more than 3 SDs above the group mean. Participants were recruited via the online recruitment service Prolific (www.prolific.co) and tested online using the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine et al., 2019). All participants were native speakers of English, aged between 18 and 40 years, had no reported hearing difficulties, a high acceptance rate (>90%) on Prolific, and had not taken part in any studies using similar stimulus materials. Ethical approval for this study was obtained from the departmental ethics committee.

Materials and procedure

After providing informed consent, all listeners completed a screening task to ensure they were wearing headphones and could hear the sounds played to them (Woods et al., 2017). Following this screening, listeners assigned to the different training groups completed two training tasks, a brief recognition task and finally a task in which they were asked to rate the perceived trustworthiness of the voices they had just learned. Listeners in the No Training group completed the trait rating task only.

Auditory stimuli were extracted from the LUCID corpus (Baker & Hazan, 2011). This corpus includes voice recordings of 40 young adult speakers (20 male, 20 female) of Standard Southern British English. We selected 4 female voices from the corpus as the set of identities used in this experiment.

For the training tasks, each of the 4 voice identities was represented by 25 stimuli (100 stimuli in total). To include substantial within-person variability in our stimulus sets for each person, stimuli were sampled from a range of different speaking styles and speaking situations, across a number of recording sessions. Specifically, 10 stimuli were extracted from unscripted, conversational speech (5 stimuli produced in adverse speaking conditions, leading to 'clear' speech to enhance the intelligibility of the speech, 5 stimuli in conversational speech without any manipulations). The linguistic content varied across stimuli and was considered to be of neutral valence in content (e.g., 'Do you have two seagulls in the air?'; 'Yes, which has two bees on it'; 'One's recycled, one is waste'). A further 10 stimuli were read sentences (5 stimuli 'clear' speech, 5 stimuli 'normal' read speech; e.g., 'The woman stopped to pay a bill', 'Wasps and bees are part of the summer'). Finally, 5 stimuli were recordings of semi-spontaneous speech elicited via a picture naming task (e.g., 'I can see a [ITEM]'). All stimuli were normed for intensity using PRAAT. Training stimuli were on average 2.2 seconds (*SD* = 0.5 seconds) in duration.

In the first training task, listeners were passively exposed to randomly ordered blocks of stimuli from each of the four voice identities (all 100 stimuli presented once, via 2 blocks of 12 or 13 stimuli per speaker) while a name was displayed on the screen (e.g., 'This is Anna'). Listeners were asked to listen carefully and try to memorize the voice and

the name (cf. Lavan, Knight, Hazan & McGettigan, 2019; Lavan, Knight & McGettigan, 2019, for studies using a similar training paradigm). For the valenced training groups, an additional vignette was presented as text alongside the name on the screen during playback (e.g., ‘She helped an elderly man cross the road’ and ‘She is very patient when dealing with other people’s problems’ for positive valence; and ‘She is lying about her age in her dating profile’ and ‘She didn’t apologise even though she knew she was in the wrong’ for negative valence). The content of these vignettes was chosen to cover mildly valenced everyday situations as opposed to more extreme behaviours. This was done to avoid eliciting ceiling or floor effects in the mean ratings of trustworthiness, which would in turn affect the overall variability in judgements and would thus potentially produce misleading results.

In the second part of the training, listeners heard the same 100 stimuli again but were now asked to identify the voice for each stimulus in a 4-way forced-choice paradigm (‘Whose voice did you hear?’). Trialwise audio-visual feedback was provided indicating whether the response was correct or not, followed by a screen displaying the name of the correct voice identity (for both correct and incorrect responses). In the case of the valenced training, additional novel vignettes were presented alongside the name on the feedback screen (5 per voice identity, repeated 5 times across the training).

Following the training task, participants completed a brief recognition task, such that we could assess whether listeners had successfully learned to recognize the different voice identities. The task was the same as the one used in the second training (4-way forced-choice recognition), but without feedback. There were 20 trials (5 stimuli \times 4 speakers, 1 stimulus per speaking style) using previously unheard stimuli sampled from across the same speaking styles as the training stimuli. Listeners who were less than 50% correct on this task (chance = 25%) were excluded from the final sample ($N = 5$) as these listeners did not show sufficient familiarity with the voices. After these exclusions, participants correctly identified the voices in 79.8% of the trials ($SD = 14.5\%$). Listeners thus learned to recognize the voice with good accuracy by the end of the training.

Finally, listeners completed a trait rating task where they were asked to rate 100 previously unheard stimuli (25 stimuli \times 4 voice identities, sampled in the same way as described for the training sessions) for perceived trustworthiness (‘How trustworthy does this voice sound?’; 1 – not trustworthy at all, 7 – very trustworthy). For this task, we also included a number of vigilance trials for which a computer-generated male voice instructed listeners to give a certain rating for this trial (e.g., ‘Please click on 7’). Listeners who failed to accurately respond to more than 20% of these trials were excluded from the sample ($N = 6$).

The order of stimuli was randomized across all tasks. For the valenced training, we furthermore counterbalanced the vignettes presented alongside the voices across participants.

Results

Cronbach’s α for the ratings of all listeners groups was high (Positive: $\alpha = .91$; Neutral: $\alpha = .92$; Negative: $\alpha = .89$; No Training: $\alpha = .84$). The following analyses differ from our preregistered analyses: We preregistered all analyses as ANOVAs. It, however, became apparent that participant effects were present in the data that needed to be accounted for. We therefore opted to use linear mixed models instead and include participant as a random factor. Analyses that explore the same effects as described in the preregistration using linear mixed models are labelled here as confirmatory analyses. Analyses not

considered in the preregistration are labelled as exploratory analyses. All post-hoc tests were Bonferroni-corrected for multiple comparisons. We will first present the analysis of mean trustworthiness ratings, followed by the analyses of the variability of these ratings.

Effect of familiarity on trustworthiness ratings

In a confirmatory analysis, we assessed how familiarity affects overall trustworthiness ratings. Since the raw data were ordinal, we averaged these data across items to create quasi-continuous data following normal distributions.

To assess the effect of the different kinds of training on trustworthiness ratings, we first created an intercept-only linear mixed model (LMM) with training group (3 levels: Negative, Neutral, and Positive Training) and speaker (4 levels) as fixed effects and participant as a random effect using *lme4* in the R environment. Significance of effects was determined via log likelihood tests. There was a significant effect of training group on trustworthiness ratings ($\chi^2[2] = 7.95, p = .019$). A planned post-hoc contrast conducted using the *emmeans* package (Lenth, 2017) in R confirmed our prediction that ratings were indeed modulated in the expected pattern (Positive > Neutral > Negative Training; $t_{96,1} = 2.81, p = .006$; see Figure 1a).

We ran a separate LMM including the data from all 4 training groups to compare the data for the unfamiliar listeners who received no training to the Neutral Training group (confirmatory analysis) and the valenced training groups (exploratory analyses). This model again included training group and speaker as fixed effects (now with 4 levels: Negative, Neutral, Positive, and No Training) and participant as a random effect. There was again a significant effect of training group on trustworthiness ratings ($\chi^2[3] = 17.45, p < .001$). Three pairwise post-hoc tests conducted using the *emmeans* package (Lenth,

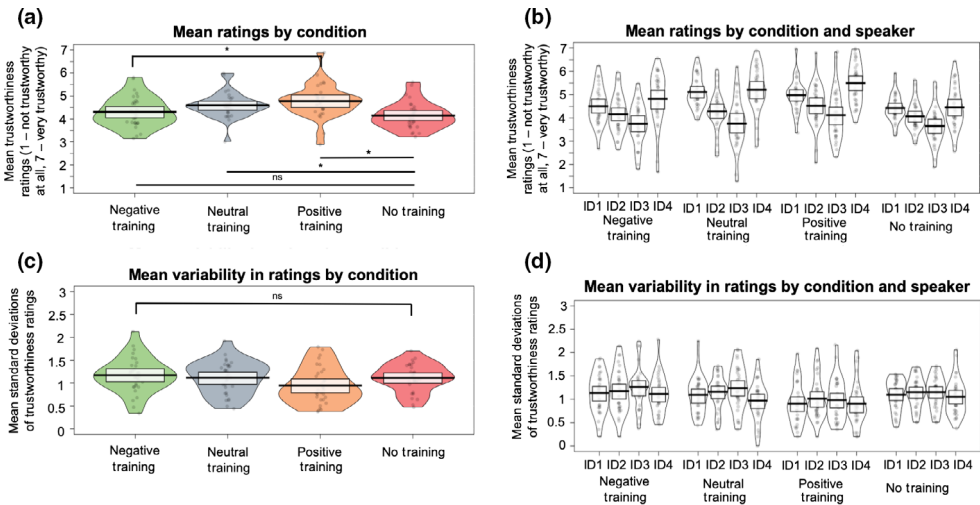


Figure 1. Results of the ratings task for Experiment 1. (a) Mean trustworthiness ratings by training group. (b) Mean trustworthiness ratings plotted by training group and speaker. (c) Mean standard deviations of trustworthiness ratings per participant by training group. (d) Mean standard deviations of trustworthiness ratings per participant plotted by training group and speaker. * indicates significant differences after Bonferroni correction. Boxes show the 95% confidence intervals.

2017) in R (alpha corrected to .017 for multiple comparisons) showed that ratings were significantly lower for the No Training group compared to the Neutral Training group ($t_{216} = 2.76, p = .007$). Ratings were also lower for the No Training group compared to the positively valenced training group ($t_{216} = 3.872, p < .001$). There was no significant difference between the ratings for the No Training group and the Negative Training group ($t_{216} = .988, p = .325$; see Figure 1a).

Mean trustworthiness ratings by speaker and training group (see Figure 1b) showed that there were significant speaker effects on mean trustworthiness ratings ($\chi^2[3] = 172.56, p < .001$). Crucially, however, no clear ceiling (or floor) effects were present for any of the individual speakers, which is essential for a valid assessment of the variability of ratings (see below).

Effect of familiarity on variability in trustworthiness ratings

In another set of confirmatory analyses, we assessed whether familiarity affects variability in trustworthiness ratings. For this purpose, we calculated the standard deviations of trustworthiness ratings per participant and speaker as an index of the variability in ratings. For a confirmatory analysis, we created a model including the 4 training groups with the same structure as the models described above.

Against predictions, this model showed no significant effect of training ($\chi^2[3] = 6.44, p = .092$; see Figure 1c). Planned post-hoc pairwise comparisons also showed no significant effects between the No Training group and any of the three training groups ($ts[127] < 1.76, ps > .080$). Similarly, there was no difference between the Neutral Training and Positive and Negative Training groups ($ts[127] = 1.79, ps > .075$). We also note that no speaker effects were apparent, with standard deviations being similar across all speakers and groups (Figure 1d).

Discussion

This experiment sought to investigate how ratings of perceived trustworthiness from variable stimuli differ for first impressions upon hearing a voice for the first time and second or lasting impressions after having learned to recognize a voice. We showed that mean ratings of trustworthiness are indeed affected by the different kinds of training: Listeners were exposed to positively or negatively valenced vignettes providing them with information about the behaviour or character traits of the people whose voices they were learning to recognize. Overall trustworthiness ratings were shifted in line with the valence of the vignettes compared to listeners who simply learned the names in the absence of any information about the person. This finding shows that additional information provided during learning of a voice identity is likely encoded and can affect the listener's evaluation of the speaker. It further confirms that listeners were sensitive to the different types of training in our experiment. It is worth noting that these training-induced shifts in ratings were relatively subtle. In fact, the differences in how trustworthy one voice sounds compared to another (e.g., ID3 vs. ID4 in this experiment) are more pronounced. Similarly, the overall pattern of trustworthiness ratings per identity (ID4 > ID1 > ID2 > ID3) is preserved despite the different training paradigms. These observations suggest that overall trustworthiness ratings – at least in this task format and with the selected stimuli – are not primarily driven by the training, but by stimulus-specific properties that shape the mean ratings.

Against predictions, trustworthiness ratings provided by unfamiliar listeners who had received no training were significantly lower than trustworthiness ratings from listeners who had only learned to recognize the voice by name without any additional (valenced) information. We, however, note that this effect was not replicated in Experiment 2 or Experiment 3 and should thus not be overinterpreted.

Also against predictions, we did not find any effect of familiarity on the variability of trait judgements: Trustworthiness ratings were similarly variable across all 4 training groups, no matter whether and how listeners had been familiarized with the voices (valenced training vs neutral training). Our results therefore suggest that neither familiarity nor semantic knowledge reduces the variability in explicit trait ratings from voices. This result is surprising in the context of the previous literature: Face perception research has shown that variability in ratings is reduced when judging social traits from familiar (famous) faces (Mileva, *et al.*, 2019). Similarly, in the voice identity perception literature, a body of work shows that within-person variability affects listeners differentially depending on whether they are familiar or unfamiliar with a voice, producing large behavioural differences as a function of familiarity (Lavan, Burston & Garrido, 2019; Lavan, Burston, Ladwa, *et al.*, 2019; Lavan, Merriman, *et al.*, 2019).

Experiment 2

Since the findings of Experiment 1 were unexpected, we sought to replicate these findings in Experiment 2 and extend them beyond trustworthiness ratings to another trait: dominance. We opted for dominance ratings as a second trait as this is the other most frequently described – and orthogonal – dimension in vocal trait space alongside the trustworthiness dimension (e.g., McAleer *et al.*, 2014). We also streamlined our design compared to Experiment 1: We only included the Neutral and No Training groups, as no differences in variability of ratings were apparent between the Positive and Negative versus Neutral Training groups in Experiment 1 (see Figure 1c).

Method

Participants

Sixty-two participants (*mean* age = 27.6 years, *SD* = 6.0 years, 42 female) were included in the final sample for this study (31 participants × 2 training groups [Neutral Training, No Training]). This sample size was matched to the one used in Experiment 1. Before arriving at this final sample, 7 participants were excluded based on preregistered exclusion criteria: 4 failed the vigilance trials (see *Materials and Procedure*), and 3 did not learn to recognize the voices well enough to pass our set cut-off of 50% correct (chance = 25%; see *Materials and Procedure*). Participants were recruited via Prolific (Prolific.co) and tested online using the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine *et al.*, 2019). All participants were again native speakers of English, aged between 18 and 40 years, had no reported hearing difficulties, a high acceptance rate (>90%) on Prolific, and had not taken part in any studies using similar stimulus materials in the laboratory. Ethical approval for this study was obtained from the departmental ethics committee.

Materials and procedure

The stimuli were identical to the ones used in Experiment 1. The procedure was also identical to Experiment 1 with two exceptions: 1) Instead of 3 training groups, we only retained the ‘Neutral Training’ group, and 2) listeners completed two rating blocks, one for perceived trustworthiness (‘How trustworthy does this voice sound?’; 1 – not trustworthy at all, 7 – very trustworthy) and another for perceived dominance (‘How dominant does this voice sound?’; 1 – not dominant at all, 7 – very dominant), with block order being counterbalanced across participants to ensure the relative independence of the trustworthiness and dominance ratings. In the post-training recognition test, listeners in the (neutral) training group were able to recognize the 4 voice identities in 75.3% ($SD = 13.1\%$) of trials (see Experiment 1).

Results

As in Experiment 1, Cronbach’s α for the ratings of all listener groups was high (for dominance ratings: Neutral: $\alpha = .94$, No Training: $\alpha = .92$; for trustworthiness ratings: Neutral: $\alpha = .90$, No Training: $\alpha = .85$).

Effect of familiarity on trustworthiness and dominance ratings

In a confirmatory analysis, we assessed whether familiarity affects overall ratings of trustworthiness and dominance. We again averaged the rating data per scale across items to create quasi-continuous data that follow a normal distribution. We then created intercept-only linear mixed models (LMMs) with training group (2 levels: Neutral Training, No Training) and speaker as fixed effects and participant as a random effect using *lme4* in the R environment. Significance of effects was again determined via log likelihood tests. There was no significant effect of training on trustworthiness ratings ($\chi^2[1] = .46$, $p = .496$; Figure 2a), thus not replicating the change in ratings between the Neutral and No Training groups from Experiment 1. There was, however, a significant effect of training on dominance ratings, with dominance ratings being higher after training ($\chi^2[1] = 8.15$, $p = .004$; Figure 2e).

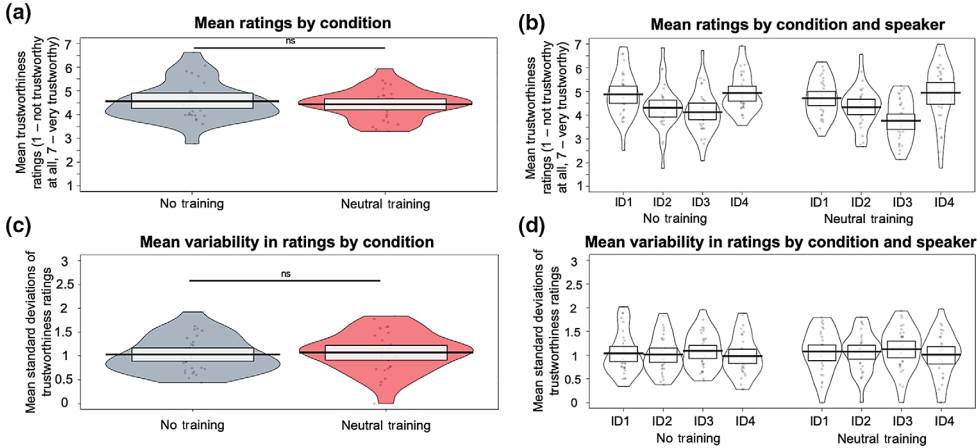
There were again significant speaker effects for mean trustworthiness ratings ($\chi^2[3] = 63.24$, $p < .001$; Figure 2b) and dominance ratings ($\chi^2[3] = 268.92$, $p < .001$; Figure 2f). However, again no clear ceiling (or floor) effects were present for any of the individual speakers.

Effect of familiarity on variability in trustworthiness and dominance ratings

As in Experiment 1, we assessed whether familiarity affects variability in trustworthiness and dominance ratings in another set of confirmatory analyses. For this purpose, we again calculated the standard deviations of trustworthiness and dominance ratings for each speaker and participant separately for each social trait as an index of variability in ratings. Our LMMs included speaker and training group as fixed effects and participant as a random effect. There was no significant effect of training on variability in trustworthiness ratings ($\chi^2[1] = .18$, $p = .672$; Figure 2c), replicating our finding from Experiment 1. There was also no effect of training on variability in dominance ratings ($\chi^2[1] = .30$, $p = .587$; Figure 2g).

We considered that our preregistered measure of variability per scale may not be sensitive enough to detect potential effects. To combine the data from the two rating scales within the same analysis for greater sensitivity (cf. Mileva *et al.*, 2019), we therefore

TRUSTWORTHINESS



DOMINANCE

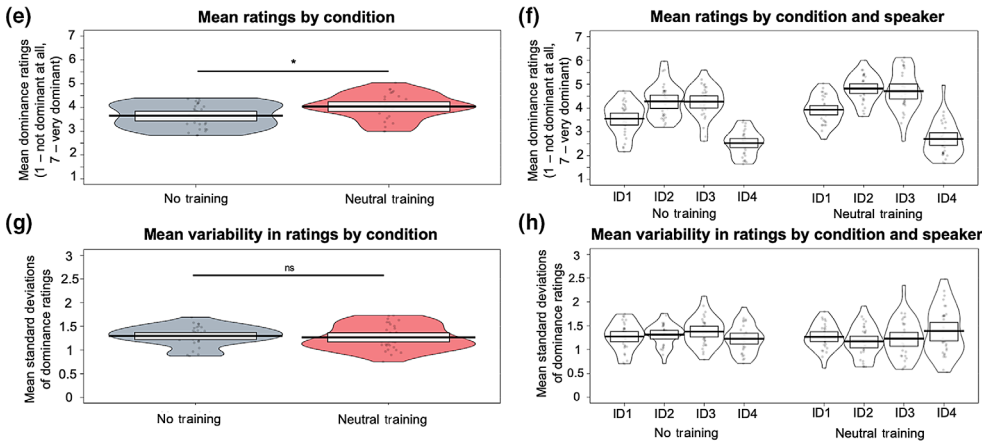


Figure 2. Results of the ratings task for Experiment 2. (a) Mean trustworthiness ratings by training group. (b) Mean trustworthiness ratings plotted by training group and speaker. (c) Mean standard deviations of trustworthiness ratings per participant by training group. (d) Mean standard deviations of trustworthiness ratings per participant plotted by training group and speaker. (e) Mean dominance ratings by training group. (f) Mean dominance ratings plotted by training group and speaker. (g) Mean standard deviations of dominance ratings per participant by training group. (h) Mean standard deviations of dominance ratings per participant plotted by training group and speaker. * indicates $p < .05$. Boxes show the 95% confidence intervals.

calculated trialwise 2D Euclidean distances relative to each participant’s mean ratings for each speaker. To align this exploratory analysis with the analysis of standard deviations above, we then averaged Euclidean distances across all items per speaker per participant to arrive at the same number of observations entered into the models. Using these participant- and speaker-wise averages, we again built LMMs with training and speaker as a fixed effect and participant as a random effect. This additional analysis did not find any effect of training on trait rating variability ($\chi^2[1] = .01, p = .935$). The spread of ratings in this 2D trait space thus appears to be similar for both participant groups, aligning with our original analyses using standard deviations as a measure of variability.

Discussion

There was again no evidence in Experiment 2 that familiarity – in this case in the absence of any semantic knowledge of the learned voice identities – affects variability in either trustworthiness or dominance judgements across two measures of variability. While trustworthiness ratings in Experiment 1 were significantly lower in the No Training group compared to the group that received neutral or familiarity only training, there was no difference in mean ratings for trustworthiness across groups in this experiment. There was, however, a significant difference in dominance ratings, with ratings being higher for listeners who received training. In order to determine whether these results are due to the way participants were artificially familiarized with the voice identities, Experiment 3 compares trait ratings attributed to naturally familiar and unfamiliar identities as well as their overall variability.

Experiment 3

It has often been discussed that familiarity established through laboratory-based training may differ from other kinds of familiarity (e.g., Fontaine, Love & Latinus, 2017, for differential effects for different types of familiarity). Notably, and in contrast to our experiments so far, Mileva *et al.* (2019) used images of familiar celebrities in their study to show a reduction in variability in ratings for familiar faces. We therefore hypothesized that the null findings in our laboratory-based training experiments could arise from listeners not being familiar enough with the voices. In a final experiment, we therefore opted to test listeners who had become familiar with the voices outside of laboratory-based tasks. We did this by measuring trait perception from 3 voices from a popular TV show (*Breaking Bad*) for groups of participants who were either unfamiliar or familiar with the show. We thus tested one group of listeners who had watched *Breaking Bad* and were familiar with the voices of the main characters, and a group of listeners who were unfamiliar with the show and furthermore could not recognize the actors included by their voices.

Method

Participants

Sixty-two participants (mean age = 27.3 years, $SD = 6.1$ years, 32 female) were included in the final sample for this study (31 participants \times 2 familiarity status [familiar, unfamiliar]). This sample size was matched to the one used in Experiments 1 and 2. Before arriving at this final sample, 21 participants were excluded based on preregistered exclusion criteria: 8 familiar participants reported to have seen less than a full season of the show, and 4 participants reported to be familiar with the show but were not able to recognize the characters in question with the desirable accuracy (50% correct; chance = 33%; see *Materials and Procedure*). For the listeners who reported to be unfamiliar (i.e., to have not watched *Breaking Bad*), 9 listeners were excluded as they reported having recognized the voice of one or more of the actors from elsewhere (e.g., Bryan Cranston as the father in *Malcolm in the Middle*). Participants were recruited via Prolific (Prolific.co) and tested online using the Gorilla Experiment Builder (www.gorilla.sc, Anwyl-Irvine et al., 2019). As in Experiments 1 and 2, all participants were native speakers of English, aged between 18 and 40 years, had no reported hearing difficulties, a high acceptance rate (>90%) on Prolific, and had not taken part in any studies using similar

stimulus materials in the laboratory. Ethical approval for this study was obtained from the departmental ethics committee.

Materials and procedure

We created new sets of stimuli for this experiment. Twenty-five brief, naturally varying recordings of voices of 3 of the main characters from the TV show *Breaking Bad* (Walter White, Hank Schrader, and Mike Ehrmantraut) were extracted from different scenes of the TV show. Stimuli included a full meaningful utterance with minimal background noise and included natural within-person variability. Catchphrases or linguistic content that could help identify the characters was avoided (see also Lavan, Merriman *et al.*, 2019). On average, these stimuli were 1.70 seconds ($SD = 0.56$ seconds) in duration.

The procedure was comparable to the one used in Experiments 1 and 2, although there was no training component in this Experiment: If listeners reported to have watched *Breaking Bad*, they first completed a brief 3-way forced-choice recognition task with 12 trials (3 identities \times 4 stimuli; stimuli were independent of those used in the trait ratings tasks) and then went on to complete the trait rating blocks. This recognition check confirmed that listeners who reported to have watched the show were indeed familiar with the voices, as they were able to correctly recognize the three voices in 78.0% ($SD = 26.2\%$) of the trials at the recognition test. Unfamiliar listeners completed the 2 trait rating blocks only. These rating blocks were identical in their design to the ones described in Experiment 2: Listeners rated the 75 stimuli (3 identities \times 25 stimuli) for perceived trustworthiness and perceived dominance. Block order was counterbalanced across participants.

Results

As in the previous experiments, Cronbach's α for the ratings of all listener groups was high (for dominance ratings: Familiar: $\alpha = .93$, Unfamiliar: $\alpha = .94$; for trustworthiness ratings: Familiar: $\alpha = .97$, Unfamiliar: $\alpha = .92$).

Effect of familiarity on trustworthiness and dominance ratings

In a confirmatory analysis, we again assessed whether familiarity affects overall ratings of trustworthiness and dominance. Using data that were averaged across stimuli (see Experiments 1 and 2), we created intercept-only LMMs with familiarity (2 levels: familiar, unfamiliar) and speaker as fixed effects and participant as a random effect using *lme4* in the R environment. There was no effect of familiarity on mean trustworthiness ($\chi^2[1] = .10, p = .752$; Figure 3a) or mean dominance ratings ($\chi^2[1] = .54, p = .462$; Figure 3e).

Speaker effects were again apparent for mean trustworthiness (Figure 3b; $\chi^2[3] = 7.79, p = .020$) and dominance ratings (Figure 3f; $\chi^2[3] = 102.87, p < .001$) with no clear ceiling (or floor) effects being present for any of the individual speakers.

Effect of familiarity on variability in trustworthiness and dominance ratings

As in the previous experiments, we assessed whether familiarity affects variability in trait ratings in another set of confirmatory analyses using the standard deviations of trustworthiness and dominance ratings. There was no significant effect of familiarity on variability for trustworthiness ratings ($\chi^2[1] = .345, p = .504$; Figure 3c), nor was there a

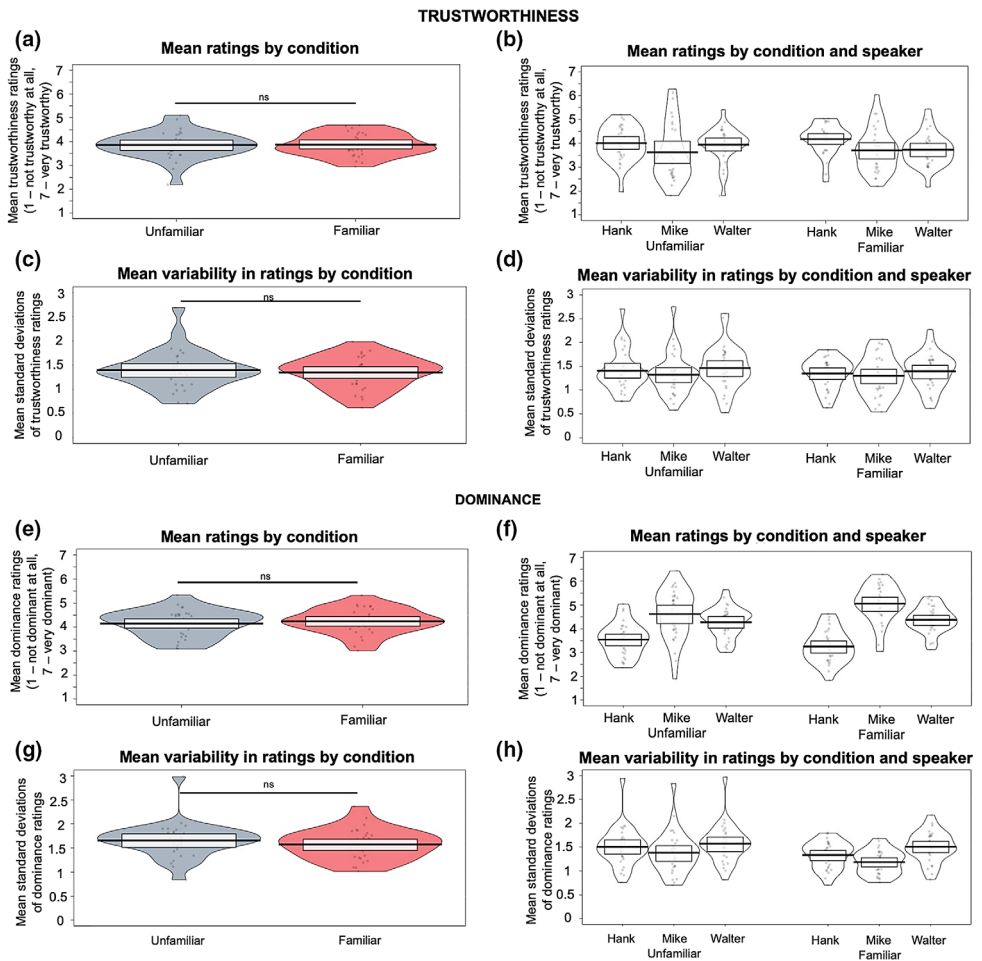


Figure 3. Results of the rating task for Experiment 3. (a) Mean trustworthiness ratings by training group. (b) Mean trustworthiness ratings plotted by training group and speaker. (c) Mean standard deviations of trustworthiness ratings by training group. (d) Mean standard deviations of trustworthiness ratings plotted by training group and speaker. (e) Mean dominance ratings by training group. (f) Mean dominance ratings plotted by training group and speaker. (g) Mean standard deviations of dominance ratings by training group. (h) Mean standard deviations of dominance ratings plotted by training group and speaker. * indicates $p < .05$. Boxes show the 95% confidence intervals.

significant effect for dominance ratings ($\chi^2[1] = 3.17, p = .075$; Figure 3g). We again ran an analysis measuring variability using 2D Euclidean distances to combine both trait ratings within the same analysis (see Experiment 2). As in Experiment 2, no effect of familiarity on 2D Euclidean distances was found ($\chi^2[1] = 1.69, p = .194$).

Discussion

In Experiment 3, we manipulated the kind of familiarity that listeners had with the voices, by stepping away from laboratory-based training towards the more naturalistic learning

context of watching a TV show. However, again we observed no evidence for effects of familiarity on trait judgements, in terms of the mean judgements of trustworthiness and dominance, or in terms of the variability in these judgements.

General Discussion

In a series of three experiments, we set out to explore the effect of familiarity on the variability in trait ratings attributed to voices. Overall, we found no compelling evidence that familiarity had an effect on how listeners rate social traits based on variable recordings of the same voices. We found some evidence in Experiment 1 that overall trait ratings could be affected via semantic knowledge, but some of these effects did not appear to be consistent across the different experiments. There was furthermore no evidence in any of the experiments to support our prediction that variability in trait judgements is reduced when listeners are familiar with the voices. These results are unexpected, and we will discuss possible explanations for the findings in the following paragraphs.

To rule out that design differences may have obscured any effects in our study, we will first map out differences between our study and the study of Mileva *et al.* (2019), which reports reduced variability in ratings for images of famous faces compared to unfamiliar faces. Mileva *et al.* (2019) employed a between-subjects design where two groups of participants rated 4 images from either 40 famous (and thus familiar) or 40 unfamiliar faces on 5 social traits (dominance, trustworthiness, attractiveness, distinctiveness, and extraversion) on a 9-point scale. A reduction in variability was detected using Procrustes analyses. Our experiments used between-subjects designs where groups of participants rated 25 voice recordings of 3 (Experiment 3) or 4 (Experiments 1 and 2) different identities on two different traits (trustworthiness and dominance). No reduction in variability was detected using standard deviations per trait ratings and 2D Euclidean distances derived from identity-specific mean ratings per participant across the two trait ratings.

There is a clear difference in how the studies weighted the number of identities against the number of items per identity. Since identity recognition is generally less reliable and more difficult for voices compared to faces (Barsics, 2014), we opted for a smaller number of identities to ensure that listeners could readily learn (Experiments 1 and 2) and recognize the voice identities with good accuracy. Additionally, we assumed that a reduction in variability should have been apparent for any identity used, as long as no ceiling or floor effects would be apparent. The degree of variability exhibited for the three (or four) different voice identities may differ (see Figures 1–3) and may not be a representative estimate of the absolute variability in human voices at large. We were, however, interested in relative reductions of the existing variability for familiar (relative to unfamiliar) listeners, reflecting our prediction that familiarity should be associated with more consistently rating the individual (familiar) *people* rather than individual stimuli. Similarly, we decided to select more items per identity to widely sample each voice's within-person variability. Thus, while the differences in stimulus sets give a different focus in the two studies, this difference is unlikely to affect our results.

Another difference is the type of analyses performed to quantify differences in variability, alongside the number of trait ratings collected for each image. Our study design was not suitable for running the same Procrustes analyses used by Mileva *et al.* (2019): Due to our design choices (using few identities and many items as opposed to many identities and fewer items), the Procrustes analyses were underpowered leading to highly variable

fits. We therefore reanalysed Mileva *et al.* (2019) trustworthiness and dominance ratings only, using standard deviations and 2D Euclidean distances as measures of variability, thus replicating the analyses reported in the current study. Using these analyses, we find significant reductions in the variability of ratings for familiar (vs. unfamiliar) viewers for both types of variability measures, and for both trustworthiness and dominance ratings (see Appendix S1). Neither the increased multidimensionality of the data nor the type of analyses should therefore have affected our results.

Finally, we used a reduced 7-point scale compared to the 9-point scale used in Mileva *et al.* (2019). Given the magnitude of the effects in our reanalysis of Mileva *et al.*'s (2019) data (see Appendix S1), however, we would expect that a 7-point scale should have been sufficient to detect similar effects in voices. Nevertheless, using the wider 9-point scale or even a visual analogue scale could increase sensitivity and therefore should be considered in future research.

We thus argue that it is unlikely that our specific experimental design may have obscured any results. If this is the case, our null findings thus differ from reports of reductions in variability of trait perceptions from familiar faces compared with unfamiliar faces (Mileva *et al.*, 2019). Neither do our results mirror the substantial behavioural effects that familiarity has on (voice) identity judgements in the context of within-person variability (e.g., Jenkins *et al.*, 2011; Lavan, Burston & Garrido, 2019; Lavan, Burston, Ladwa, *et al.*, 2019). What could explain these differences?

The differences between Mileva *et al.* (2019) findings for faces and our null effect for voices could stem from basic differences in face and voice processing, which may be interacting with the nature of the task. Voice identity perception is usually seen as being more difficult and less reliable than face identity perception (e.g., Barsics, 2014). We would first, however, argue that it is unlikely that broad differences in familiarity with our set of voices and Mileva *et al.*'s (2019) set of faces are driving the differences in results across studies: Recognition accuracy, at least within a 3- or 4-way forced-choice recognition task, was good across all studies, which indicates that listeners were familiar with the voices. We furthermore note that differences in the degree of familiarity across participants do not seem to be related to variability in trait judgements: When correlating familiar listeners' recognition accuracy – an index of the degree of familiarity – with the variability of their trait ratings, no significant relationships were found (see supplementary analyses 2). We can thus assume that the lack of effects for voices does not arise based on categorically lower familiarity with the identities for voices than for faces.

However, the fact that voice perception is in general less reliable and more difficult implies that familiarity with a person – and thus being able to recognize this person, even in the context of variable stimuli – could be more salient for faces. A clearer percept of the identity and associated semantic knowledge of a person may be harder to suppress and may thus lead to interference in trait ratings for familiar faces. For voices, the percept of identity may in general be less readily and reliably perceived and thus weaker. This may allow listeners to be able to judge individual stimuli of familiar voices for social traits without much interference from what they may know about the identities (at least when prompted explicitly).

A similar line of argument, invoking differences in the saliency or immediacy of familiarity in making trait judgements from faces and voices, could also underpin the differences for identity perception and for trait perception in voices. The relationship between trait evaluations and familiarity is more complex than the relationship between familiarity and identity perception. For example, there is a ground truth to an identity

percept that remains stable: A person is unlikely to apparently change their identity except for a few exceptional situations, for example, disguise or dramatic physiological changes. However, there is no such simple ground truth to trait evaluations: A person may be trustworthy in one context but not in another. It could therefore be seen as adaptive to be able to rapidly update trait evaluations, even for familiar others, and rely less – or not at all, as our data may suggest – on fixed trait impressions for familiar identities. Notably, this explanation does, however, not explain the differences in findings between face perception (Mileva *et al.*, 2019) and voice perception (the current study).

Additionally, it could be argued that identity perception and trait perception tap into different stages of the processing of other people, in a way that exposes differences across modalities: Trait perception from unfamiliar faces or voices can be achieved rapidly, with raters being able to provide trait judgements with high agreement after being only very briefly presented with a face (< 200ms; Todorov, Pakrashi, & Oosterhof, 2009) or a voice (< 400ms; McAleer *et al.*, 2014). For familiarity to affect trait judgements, participants need to recognize the face or voice to then access the person-specific information associated with this face/voice (Belin, Bestelmeyer, Latinus & Watson, 2011; Bruce & Young, 1986). Identity recognition is rapid for faces (e.g., 100 ms stimulus presentation time, Besson *et al.*, 2017, for identification; < 400 ms, Ramon, Caharel & Rossion, 2011, for familiarity judgements) but slower – at least if good accuracy is expected – for voices (e.g., > 1 second, Schweinberger, Herholz & Sommer, 1997, for familiarity judgements; Bricker & Pruzansky, 1966, for identification). The relative difference in when trait judgements for familiar people can be accessed between modalities could thus explain our results. We note, however, that stimulus materials all exceeded 1 second in duration and participants were required to listen to the entire voice recording before providing their judgement, without any time limit being imposed. Our task thus did not involve speeded responses and should have enabled listeners to fully evaluate the identity (if familiar) and access associated trait ratings before making a judgement. A differential time course of trait judgements for familiar and unfamiliar voices is therefore unlikely to have influenced our results.

Based on our findings, a number of open questions remain that should be tackled in future work: While we can offer speculative explanations for the differences between modalities, future studies specifically designed to probe potential differences across modalities will be required to better understand the source of the observed differences in face and voice processing. We also note that explicit trait ratings of people based on short recordings of voices are, after all, a highly artificial task, and we do not yet have a good grasp of how such ratings correspond to how people perceive social traits outside of experimental tasks. Future efforts should also examine how explicit trait ratings may map onto other measures, and how these in turn map onto trait perception as it may happen in naturalistic settings.

What do our results mean in the context of theoretical frameworks in the field? Finding overall no compelling differences in trait ratings for familiar and unfamiliar listeners suggests that all listeners were able to perceive that multiple recordings of the same voice varied in terms of how trustworthy or dominant that voice sounded for this particular recording. From this, it would thus follow that stimulus-based first impressions may not irretrievably fade away to be quickly replaced by second and lasting impressions when listeners become familiar with a person. Instead – at least for voices and within explicit rating tasks – familiar listeners still seem to be able to rate a stimulus, without rating the person.

Acknowledgements

This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to Carolyn McGettigan. The authors are grateful to Mike Burton and Andy Young for helpful discussions of the results.

Author Contributions

Nadine Lavan (Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Project administration; Visualization; Writing – original draft; Writing – review & editing) Mila Mileva (Writing – review & editing) Carolyn McGettigan (Conceptualization; Funding acquisition; Supervision; Writing – original draft; Writing – review & editing).

Data Availability Statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

References

- Anwyl-Irvine, A. L., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. K. (2019). Gorilla in our Midst: An online behavioral experiment builder. *Behavior Research Methods*, *52*(1), 388–407. <https://doi.org/10.3758/s13428-019-01237-x>
- Baker, R., & Hazan, V. (2011). DiapixUK: Task materials for the elicitation of multiple spontaneous speech dialogs. *Behavior Research Methods*, *43*, 761–770. <https://doi.org/10.3758/s13428-011-0075-y>
- Ballem, C. C., & Todorov, A. (2007). Predicting political elections from rapid and unreflective face judgments. *Proceedings of the National Academy of Sciences*, *104*(46), 17948–17953. <https://doi.org/10.1073/pnas.0705435104>
- Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica*, *54*, 244–254.
- Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*, 711–725. <https://doi.org/10.1111/j.2044-8295.2011.02041.x>
- Besson, G., Barragan-Jason, G., Thorpe, S. J., Fabre-Thorpe, M., Puma, S., Ceccaldi, M., & Barbeau, E. J. (2017). From face processing to face recognition: Comparing three different processing levels. *Cognition*, *158*, 33–43. <https://doi.org/10.1016/j.cognition.2016.10.004>
- Bricker, P. D., & Pruzansky, S. (1966). Effects of stimulus content and duration on talker identification. *The Journal of the Acoustical Society of America*, *40*(6), 1441–1449.
- Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*, 305–327. <https://doi.org/10.1111/j.2044-8295.1986.tb02199.x>
- Burton, A. M., Kramer, R. S., Ritchie, K. L., & Jenkins, R. (2016). Identity from variation: Representations of faces derived from multiple instances. *Cognitive Science*, *40*(1), 202–223. <https://doi.org/10.1111/cogs.12231>
- Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and competence as universal dimensions of social perception: The stereotype content model and the BIAS map. *Advances in Experimental Social Psychology*, *40*, 61–149. [https://doi.org/10.1016/S0065-2601\(07\)00002-0](https://doi.org/10.1016/S0065-2601(07)00002-0)
- Fontaine, M., Love, S. A., & Latinus, M. (2017). Familiarity and voice representation: From acoustic-based representation to voice averages. *Frontiers in Psychology*, *8*, 1180. <https://doi.org/10.3389/fpsyg.2017.01180>

- Harris, M. J., & Garris, C. P. (2008). You never get a second chance to make a first impression: Behavioral consequences of first impressions. In N. Ambady & J. J. Skowronski (Eds.), *First impressions* (pp. 147–68). New York, NY: Guilford Publications.
- Jenkin, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*, 313–323. <https://doi.org/10.1016/j.cognition.2011.08.001>
- Klofstad, C. A. (2016). Candidate voice pitch influences election outcomes. *Political Psychology*, *37*, 725–738. <https://doi.org/10.1111/pops.12280>
- Klofstad, C. A., & Anderson, R. C. (2018). Voice pitch predicts electability, but does not signal leadership ability. *Evolution and Human Behavior*, *39*, 349–354. <https://doi.org/10.1016/j.evolhumbehav.2018.02.007>
- Klofstad, C. A., Anderson, R. C., & Peters, S. (2012). Sounds like a winner: Voice pitch influences perception of leadership capacity in both men and women. *Proceedings of the Royal Society B: Biological Sciences*, *279*(1738), 2698–2704. <https://doi.org/10.1098/rspb.2012.0311>
- Lavan, N., Burston, L. F., & Garrido, L. (2019). How many voices did you hear? Natural variability disrupts identity perception from unfamiliar voices. *British Journal of Psychology*, *110*, 576–593. <https://doi.org/10.1111/bjop.12348>
- Lavan, N., Burston, L. F., Ladwa, P., Merriman, S. E., Knight, S., & McGettigan, C. (2019). Breaking voice identity perception: Expressive voices are more confusable for listeners. *Quarterly Journal of Experimental Psychology*, *72*, 2240–2248. <https://doi.org/10.1177/1747021819836890>
- Lavan, N., Burton, A. M., Scott, S. K., & McGettigan, C. (2019). Flexible voices: Identity perception from variable vocal signals. *Psychonomic Bulletin & Review*, *26*(1), 90–102. <https://doi.org/10.3758/s13423-018-1497-7>
- Lavan, N., Knight, S., Hazan, V., & McGettigan, C. (2019). The effects of high variability training on voice identity learning. *Cognition*, *193*, 104026. <https://doi.org/10.1016/j.cognition.2019.104026>
- Lavan, N., Knight, S., & McGettigan, C. (2019). Listeners form average-based representations of individual voice identities. *Nature Communications*, *10*(1), 2404. <https://doi.org/10.1038/s41467-019-10295-w>
- Lavan, N., Merriman, S. E., Ladwa, P., Burston, L. F., Knight, S., & McGettigan, C. (2019). ‘Please sort these voice recordings into 2 identities’: Effects of task instructions on performance in voice sorting studies. *British Journal of Psychology*. <https://doi.org/10.1111/bjop.12416>
- Lenth, R. V. (2017). Using lsmeans. *The Journal of Statistical Software*, *69*, 1–33. <https://mran.microsoft.com/snapshot/2018-04-11/web/packages/lsmeans/vignettes/using-lsmeans.pdf>
- Mahrholz, G., Belin, P., & McAleer, P. (2018). Judgements of a speaker’s personality are correlated across differing content and stimulus type. *PLoS One*, *13*(10), e0204991. <https://doi.org/10.1371/journal.pone.0204991>
- McAleer, P., Todorov, A., & Belin, P. (2014). How do you say ‘Hello’? Personality impressions from brief novel voices. *PLoS One*, *9*, e90779. <https://doi.org/10.1371/journal.pone.0090779>
- Mileva, M., Kramer, R. S., & Burton, A. M. (2019). Social evaluation of faces across gender and familiarity. *Perception*, *48*, 471–486. <https://doi.org/10.1177/0301006619848996>
- Mileva, M., Tompkinson, J., Watt, D., & Burton, A. M. (2020). The role of face and voice cues in predicting the outcome of student representative elections. *Personality and Social Psychology Bulletin*, *46*, 617–625. <https://doi.org/10.1177/0146167219867965>
- Olivola, C. Y., Funk, F., & Todorov, A. (2014). Social attributions from faces bias human choices. *Trends in Cognitive Sciences*, *18*(11), 566–570. <https://doi.org/10.1016/j.tics.2014.09.007>
- Olivola, C. Y., & Todorov, A. (2010). Fooled by first impressions? Reexamining the diagnostic value of appearance-based inferences. *Journal of Experimental Social Psychology*, *46*, 315–324. <https://doi.org/10.1016/j.jesp.2009.12.002>
- Oosterhof, N. N., & Todorov, A. (2008). The functional basis of face evaluation. *Proceedings of the National Academy of Sciences*, *105*(32), 11087–11092.
- Ramon, M., Caharel, S., & Rossion, B. (2011). The speed of recognition of personally familiar faces. *Perception*, *40*, 437–449. <https://doi.org/10.1068/p6794>

- Schweinberger, S. R., Herholz, A., & Sommer, W. (1997). Recognizing famous voices: Influence of stimulus duration and different types of retrieval cues. *Journal of Speech, Language, and Hearing Research*, *40*, 453–463. <https://doi.org/10.1044/jslhr.4002.453>
- Sussman, A. B., Petkova, K., & Todorov, A. (2013). Competence ratings in US predict presidential election outcomes in Bulgaria. *Journal of Experimental Social Psychology*, *49*, 771–775. <https://doi.org/10.1016/j.jesp.2013.02.003>
- Sutherland, C. A., Rhodes, G., Burton, N. S., & Young, A. W. (2019). Do facial first impressions reflect a shared social reality? *British Journal of Psychology*, *111*, 215–232. <https://doi.org/10.1111/bjop.12390>
- Sutherland, C. A., Young, A. W., & Rhodes, G. (2017). Facial first impressions from another angle: How social judgements are influenced by changeable and invariant facial properties. *British Journal of Psychology*, *108*, 397–415. <https://doi.org/10.1111/bjop.12206>
- Todorov, A., Mandisodza, A. N., Goren, A., & Hall, C. C. (2005). Inferences of competence from faces predict election outcomes. *Science*, *308*, 1623–1626. <https://doi.org/10.1126/science.1110589>
- Todorov, A., Olivola, C. Y., Dotsch, R., & Mende-Siedlecki, P. (2015). Social attributions from faces: Determinants, consequences, accuracy, and functional significance. *Annual Review of Psychology*, *66*, 519–545. <https://doi.org/10.1146/annurev-psych-113011-143831>
- Todorov, A., Pakrashi, M., & Oosterhof, N. N. (2009). Evaluating faces on trustworthiness after minimal time exposure. *Social Cognition*, *27*(6), 813–833.
- Todorov, A., & Porter, J. M. (2014). Misleading first impressions: Different for different facial images of the same person. *Psychological Science*, *25*, 1404–1417. <https://doi.org/10.1177/0956797614532474>
- Todorov, A., Said, C. P., Engell, A. D., & Oosterhof, N. N. (2008). Understanding evaluation of faces on social dimensions. *Trends in Cognitive Sciences*, *12*, 455–460. <https://doi.org/10.1016/j.tics.2008.10.001>
- Wilson, J. P., & Rule, N. O. (2015). Facial trustworthiness predicts extreme criminal-sentencing outcomes. *Psychological Science*, *26*, 1325–1331. <https://doi.org/10.1177/0956797615590992>
- Woods, K. J., Siegel, M. H., Traer, J., & McDermott, J. H. (2017). Headphone screening to facilitate web-based auditory experiments. *Attention, Perception, & Psychophysics*, *79*, 2064–2072. <https://doi.org/10.3758/s13414-017-1361-2>
- Zebrowitz, L. A., & McDonald, S. M. (1991). The impact of litigants' baby-facedness and attractiveness on adjudications in small claims courts. *Law and Human Behavior*, *15*, 603–623. <https://doi.org/10.1007/BF01065855>

Received 18 December 2019; revised version received 5 May 2020

Supporting Information

The following supporting information may be found in the online edition of the article:

Appendix S1. Supplementary Analyses.