Check for updates

Original Article

QJEP

Quarterly Journal of Experimental Psychology
2020, Vol. 73(10) 1537–1545
© Experimental Psychology Society 2020

Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/1747021820938659
qjep.sagepub.com


Comparing unfamiliar voice and face identity perception using identity sorting tasks

Justine Johnson, Carolyn McGettigan and Nadine Lavan

Abstract

Identity sorting tasks, in which participants sort multiple naturally varying stimuli of usually two identities into perceived identities, have recently gained popularity in voice and face processing research. In both modalities, participants who are unfamiliar with the identities tend to perceive multiple stimuli of the same identity as different people and thus fail to "tell people together." These similarities across modalities suggest that modality-general mechanisms may underpin sorting behaviour. In this study, participants completed a voice sorting and a face sorting task. Taking an individual differences approach, we asked whether participants' performance on voice and face sorting of unfamiliar identities is correlated. Participants additionally completed a voice discrimination (Bangor Voice Matching Test) and a face discrimination task (Glasgow Face Matching Test). Using these tasks, we tested whether performance on sorting related to explicit identity discrimination. Performance on voice sorting and face sorting tasks was correlated, suggesting that common modality-general processes underpin these tasks. However, no significant correlations were found between sorting and discrimination performance, with the exception of significant relationships for performance on "same identity" trials with "telling people together" for voices and faces. Overall, any reported relationships were however relatively weak, suggesting the presence of additional modality-specific and task-specific processes.

Keywords

Identity processing; voices; faces; individual differences

Received: 15 November 2019; revised: 11 February 2020; accepted: 3 March 2020

Introduction

Many similarities have been described for voice and face processing (Yovel & Belin, 2013; see also Maurer & Werker, 2014, for comparisons of face and language processing). Although accuracy in voice processing tasks is generally lower (Barsics, 2014), humans can perceive a wealth of information from both a person's face and their voice, such as their emotional state or identity alongside any number of other inferred person characteristics (Belin et al., 2011; Bruce & Young, 1986). Furthermore, many of the classic effects described for face processing have been replicated for voices: Averaged faces and voices are perceived to be more attractive, and more distinctive faces and voices are better recognised (but see Papcun et al., 1989, for voice memory). Different face and voice identities are also both considered to be represented in relation to a face or voice prototype, respectively (see Yovel & Belin, 2013, for an overview).

Recent studies investigating the effects of within-person variability on voice and face perception have highlighted further similarities between voice and face identity processing (For voices: Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019; Lavan, Merriman, et al., 2019; Stevenage et al., 2020, For faces: Andrews et al., 2015; Balas & Saville, 2017; Jenkins et al., 2011; Laurence et al., 2016; Redfern & Benton, 2017; Short et al., 2017; Zhou & Mondloch, 2016). Images of faces and recordings of voices can vary considerably from instance to instance (Burton, 2013; Lavan, Burton, et al., 2019). In such naturally varying images of faces, the facial expression, hairstyle, lighting, posture, and type of camera, among

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

Corresponding author:
Nadine Lavan, Department of Speech, Hearing and Phonetic Sciences, University College London, 2 Wakefield Street, London WC1N 1PF, UK.
Email: n.lavan@ucl.ac.uk

other factors, vary substantially across different images of the same person (Burton, 2013; Jenkins et al., 2011). Similarly, the sound of a person's voice will change depending on the environment, the conversation partner, and speaking situation, among other factors. These factors lead to complex changes in the acoustic properties of the voices (Lavan, Burton, et al., 2019).

Identity sorting studies have used such naturally varying stimuli to examine how this within-person variability affects identity perception. In these identity sorting studies, participants are presented with sets of naturally varying stimuli, usually from two identities. Groups of participants who are either familiar or unfamiliar with the people represented in the stimuli are then asked to sort these stimuli by identity. For both the voice and face sorting tasks, a striking pattern of results emerges. Participants who do not know the identities tend to perceive there to be many more identities than are actually present. When looking at errors made by these participants, it becomes apparent that they fail to "tell people together," that is, participants unfamiliar with the presented identities perceive naturally varying images of the same person as different identities, confusing within-person variability with between-person variability. Notably, mixing errors—where participants perceive stimuli from two different people as the same person, i.e., fail to accurately tell people apart—rarely occur in either modality. In contrast to this, participants who are familiar with the identities can generally complete the task with good accuracy, most frequently arriving at the correct solution of two perceived identities.

These similarities across modalities are striking, but it remains unclear whether there is a relationship between performance in voice and face sorting tasks, indicating modality-general processes, or whether these similar findings derive from different modality-specific processes. This question can be addressed through an individual differences approach that tests whether participants who are good at voice sorting are also good at face sorting. A correlation across tasks would suggest that modality-general processes underpin sorting behaviour. Alternatively, there may be no relationship between modalities. In this study, therefore, we ran a voice sorting and a face sorting task with the same participants to investigate this question. Due to ceiling effects that are apparent for performance on both voice and face sorting tasks with familiar identities, we conducted the tasks using unfamiliar identities only.

Aside from whether there is a relationship between participants' performance across voice and face sorting, it is also unclear which perceptual processes or strategies may underpin sorting behaviour in either modality. Outside of sorting tasks, unfamiliar identity perception is often measured through matching or pairwise discrimination tasks. Identity sorting tasks differ from such explicit discrimination tasks in a number of ways but crucially do not dictate any specific strategy for how participants complete the task. Participants are thus relatively free to choose any strategy available to them. It has, however, been suggested that—despite the lack of clear instructions on how to complete a sorting task—discrimination strategies may still underpin how unfamiliar participants tackle an identity sorting task (Lavan, Burston, & Garrido, 2019). If this is the case, performance on a discrimination task should be correlated with performance on a sorting task (within modality). Alternatively, performance on identity sorting tasks may have no relationship with discrimination performance and would thus indicate that sorting tasks tap into other aspects of identity processing. To investigate this question, our participants completed two validated voice and face discrimination (or matching) tasks—The Bangor Voice Matching Task (BVMT; Mühl et al., 2018) and the Glasgow Face Matching Task (GFMT; Burton et al., 2010)—in addition to the identity sorting tasks.

Thus, in our experiment, participants completed two identity sorting tasks and two identity discrimination tasks, one each for voices and faces. We examined (1) whether there was a relationship in participants' performance across modalities in the two identity sorting tasks and (2) whether sorting behaviour could be linked to established tests of identity discrimination. We conducted all analyses based on an overall measure of performance (number of clusters for the sorting tasks and mean accuracy for the discrimination tasks). Furthermore, we conduct the same analyses for measures indexing participants' ability to "tell people together" and tell people apart separately: error rates for "telling people together" and telling people apart differ substantially in sorting tasks (Jenkins et al., 2011; Lavan, Burston, & Garrido, 2019) and accuracy on "same identity" trials (mapping onto "telling people together") and "different identity" trials (mapping onto telling people apart) in discrimination tasks is uncorrelated (for faces: Megreya & Burton, 2006), which suggest that these aspects of identity processing may be largely independent of one another. This study was preregistered on the Open Science Framework (osf.io/5gu3q).

Methods

Participants

Fifty participants (33 female) aged between 18 and 35 years were recruited from the Psychology Subject Pool at University College London. All participants were native speakers of English (34 British English, 8 American English, and 8 other English). None of the participants were familiar with the voices used in the study (as determined via a debrief questionnaire). All participants had corrected to normal vision and no reported hearing impairments. Ethical approval was given by the UCL Research Ethics Committee (Project ID number: SHaPS-2019-CM-030). Based on our preregistered exclusion criteria, four participants were

excluded. One participant did not accurately complete the catch trials in the sorting task. Another participant failed to move more than 80% of the icons in the voice sorting test. One participant's performance on the Glasgow Face Matching Test differed by more than 3 standard deviations from the group mean. For another participant, no data was recorded for the Glasgow Face Matching Test because of a technical error, so we discarded the whole data set. Thus, the final sample included 46 participants (mean age: 24.04, $SD = 3.77$, 33 female). This sample size was determined by the availability of funds for this project. Although the sample size is relatively low for studies of individual differences, similar sample sizes have been shown to produce replicable effects for individual difference studies in face perception (e.g., McCaffery et al., 2018 Study 1 and Study 2).

Materials

We created new sets of stimuli for the voice and the face sorting tasks, with the aim of including identities with whom participants in the United Kingdom would be unfamiliar. For this purpose, we identified two Canadian actors (Dillon Casey and Giacomo Gianniotti) who are largely unknown outside of Canada. We then gathered 15 stimuli of naturally varying stimuli per modality (voice recordings, face images) from these two identities, resulting in 60 stimuli in total. The voice recordings and pictures of the faces were sourced from Google image search, social media, YouTube videos, and Twitter.

Voice sorting materials. The 30 voice recordings were sampled from press interviews and social media posts, as well as from scenes from various television programmes. Stimuli thus include variability introduced by the use of different speaking styles reflecting the different intended audiences and speaking situations, different recording times, and different recording equipment and environments. Note that this approach for stimulus selection differs from previous voice sorting tasks where stimuli were selected from a single TV show, with the actor in question playing one specific character. The stimuli used here may therefore include more pronounced within-person variability, with stimuli being sampled from a wider range of sources. All stimuli included full meaningful utterances (e.g., "Do we have to go to this party?"; "Normally I would do it but I don't need it I'm not desperate") with as little background noise as possible and no other audible voices. The duration of the recordings ranged between 1 and 4 s ($M = 2.6$ s). The intensity of all stimuli was root-mean-square normalised to 67.7 dB using Praat (Boersma & Weenink, 2013). These stimuli were then added to a PowerPoint slide, represented by numbered boxes (see Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019; Lavan, Merriman, et al., 2019).

Face sorting materials. The 30 colour images included in this stimulus set were all broadly front facing with no part of the face being obscured, through for example sunglasses or hair. Like the voice stimuli, these images also included natural variability, such that they varied in lighting, type of the camera used, head position, and image backgrounds. Similarly, images were taken from different sources and occasions, thus including pictures of the two actors with different facial expressions, with different hairstyles, at different ages (mostly showing the actors as young adults; Giacomo Gianiotti is currently 30 years old, and Dillon Casey is currently 36 years old). The images were edited with Microsoft Photos (Microsoft Office 365 ProPlus) to 4:3 portrait ratio and cropped to show primarily the face (e.g., Jenkins et al., 2011). To better match the face sorting task to the inherently dynamic nature of the voice sorting task, we created short videos to control the duration of exposure to each of the images. These videos first showed a numbered box for 0.3 s (cf. the numbered boxes on the PowerPoint slide for the voice sorting task), followed by the static image for 2.6 s (mean duration of the auditory stimuli), followed again by the numbered box for 0.3 s. Crucially, when added to a PowerPoint slide, the images of the faces were thus not visible by default but instead numbered boxes were shown, with the images only appearing when participants played the video. All stimuli had a height of 3.12 cm and width of 2.78 cm on the PowerPoint slide. Participants were told not to change the size of the image or to pause the videos (which would have allowed them to keep the images on the screen).

Catch stimuli. In addition to these stimuli, two catch stimuli were added for each task. For the voice sorting task, a recording of a female voice created via the inbuilt text-to-speech function in an Apple Mac laptop, saying "Hello. My name is Laura," and for the face sorting task, these were two pictures of the cartoon character Bart Simpson.

Procedure

Each participant completed four tasks: a voice sorting task, a face sorting task, the BVMT (Mühl et al., 2018), and the GFMT (Burton et al., 2010). Up to 4 participants were tested simultaneously in a quiet room. All tasks were self-paced, and questions could be asked at any time. Participants completed the tasks on Hewlett Packer laptops with sounds being presented with Sennheiser headphones (e.g., Sennheiser HD 206) at a comfortable volume. The experiment lasted approximately 1 hr in total. Participants first completed the two sorting tasks (order counterbalanced) before completing the matching tasks (order also counterbalanced). This counterbalancing was chosen to avoid that participants would be biased towards using an explicit pairwise discrimination strategy in the sorting tasks if they completed one of the discrimination tasks first.

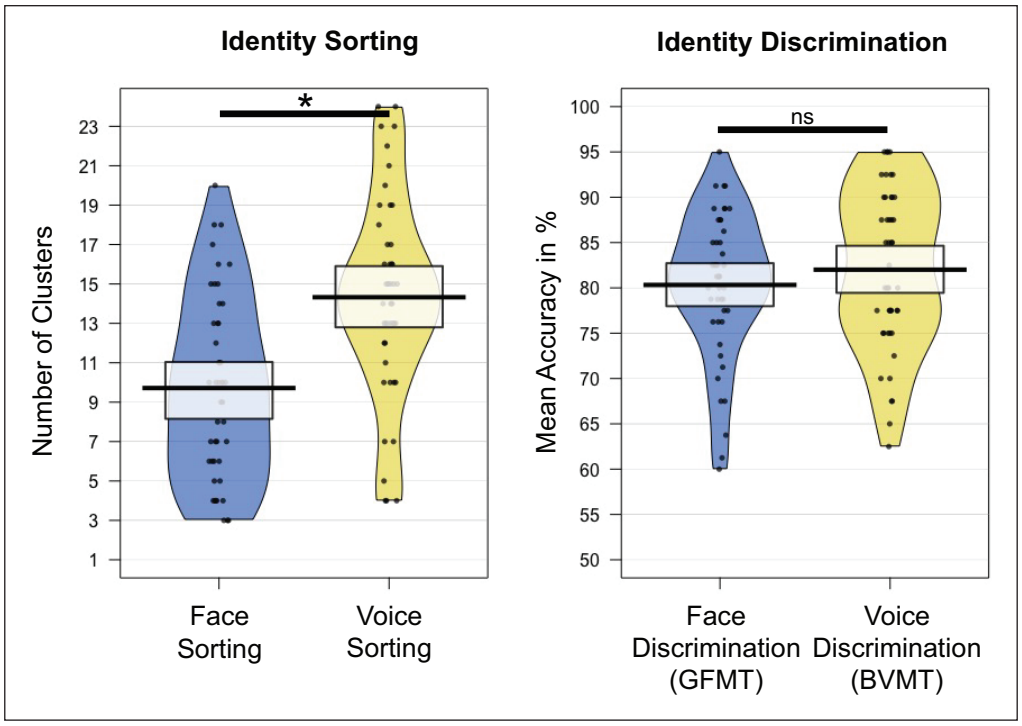

Figure 1. Performance on the identity sorting tasks (after exclusion of the catch trials; left panel) and identity discrimination tasks (right panel).
*$p < .05$. Boxes show 95% confidence intervals around the means. Dots show individual participants' performance.

The sorting tasks. For both the sorting tasks, participants were given a PowerPoint slide including the 32 stimuli (15 stimuli × 2 identities + 2 catch stimuli) represented by numbered boxes. Participants were instructed to sort these stimuli by identity, by dragging and dropping the different stimuli into distinct clusters to represent the different perceived identities. Participants were informed that there could be any number of identities represented (ranging from 1 to 32, which is the total number of stimuli). Stimuli could be replayed as many times as participants felt necessary.

The discrimination tasks. The short versions of the BVMT (Mühl et al., 2018) and the GFMT (Burton et al., 2010) were implemented on the Gorilla Experiment Builder (www.gorilla.sc; Anwyl-Irvine et al., 2020). In the BVMT (Mühl et al., 2018), participants were presented with 80 pairs of recordings of voices (40 male, 40 female) and were asked to decide whether the two recordings were from the same identity or two different identities in a two-way forced choice design (50% of the pairs were same-identity trials). The stimuli comprised read non-words (e.g., "hed," "hood," "aba," and "ibi"). Participants pressed a button to play each stimulus in turns. They could replay the stimuli as many times as they felt necessary. The GFMT (Burton et al., 2010) consists of 40 pairs of black and white images of faces (20 male, 20 female) presented simultaneously, next to each other. Here, participants were again asked to decide whether the two faces showed the same identity or were in fact two separate identities, in a two-way forced choice design (50% of the pairs were same-identity trials). The stimuli were full-face photographs (black and white) taken in the same session in good lighting conditions using two camera angles. For both tests, participants had to click the "same" or "different" button as each pair was presented before the test moved to the next pair. There was no time constraint to the tests.

Results

Exploratory analyses: overall performance on the sorting and matching tasks

Participants formed 14.32 clusters for the voice sorting task ($SD = 5.17$, $Range = 4–24$) and 9.72 clusters for the face sorting task ($SD = 4.65$, $Range = 3–20$) after we excluded the catch trials—note again that only two veridical identities were present in each of the sorting tasks (see Figure 1). An exploratory analysis confirmed that participants formed significantly fewer clusters and thus performed overall better in the face sorting task compared with the voice sorting task, $t(45) = 5.69$, $p < .001$.

The mean accuracy for the BVMT was 80.3% ($SD = 8.0\%$) and for the GFMT was 82.0% ($SD = 8.8\%$) (see Figure 1). An exploratory analysis showed that there was no difference in the overall accuracy in these two validated tasks, $t(45) = 1.12$, $p = .268$.

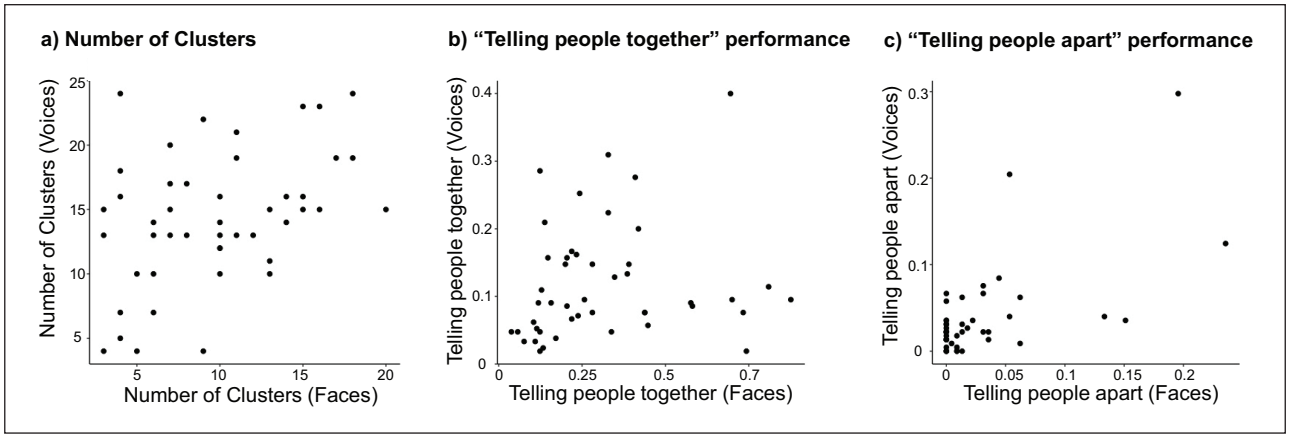

Figure 2. Scatterplots plotting measures from the voice sorting task against the face sorting task. For the number of clusters (Panel a), lower numbers indicate better performance (perfect performance = 2, as 2 identities were present in both the sorting tasks). For "Telling people together" performance (Panel b), higher numbers indicate better performance (perfect performance = 1, where all stimuli from the same identity were sorted into a single cluster). For "Telling people apart" performance (Panel c), lower numbers indicate better performance (0 = perfect performance, as this would indicate that no stimuli from different identities were put together in any of the clusters).

We note that we report a higher number in clusters for the voice sorting task than was previously reported in other voice sorting tasks (Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019). This is likely due to the broader range of materials sampled to create the stimulus sets for this study (interview footage, recordings sampled from different TV shows), compared with previous voice sorting studies (in-character voice recordings from a single TV show). The number of clusters reported for the face sorting task is similarly higher than that in previous reports (Jenkins et al., 2011; Zhou & Mondloch, 2016): We argue that this may be a result of our task design, in which faces were only visible when participants played the short video they were embedded in, increasing the task difficulty. Finding overall worse performance for voice sorting compared with face sorting aligns with other studies reporting worse performance for voice perception compared with face perception (e.g., Barsics, 2014). The matching tests do not show this difference as both tasks were normed and validated for a specific level of accuracy and designed for the purpose of detecting individual differences in the population (Burton et al., 2010; Mühl et al., 2018). Overall, the mean accuracy in our sample map well on the accuracies reported for the validated tests (BVMT: Mühl et al., 2018: 84.6%, current sample: 80.3%; GFMT: Burton et al., 2010: 81.3%, current sample: 82.0%).

Is there a relationship between performance on voice sorting and face sorting tasks?

To investigate whether there was a relationship between participants' performance across stimulus modality on the identity sorting tasks, we ran a number of correlation analyses. Shapiro–Wilk tests indicated that data were normally distributed for the total number of clusters for the sorting tasks but not for other dependent variables. Therefore, for consistency, we use Kendall's τ correlations throughout these confirmatory analyses in these sections. These were implemented in the R environment using the *Kendall* package (McLeod, 2011). We note that results remained the same when we analysed the normally distributed data with parametric tests.

There was a significant relationship between the voice and face sorting tasks for the total number of clusters, that is, the number of identities perceived (Kendall's $\tau = .27$, $p = .01$, see Figure 2a). Furthermore, we computed an index of each participant's ability of "telling people together" and telling people apart. These indices were computed in the same way as described for other voice sorting tasks (see Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019; Lavan, Merriman, et al., 2019). In brief, we created 30 × 30 item-wise response matrices for each participant (catch items were excluded), which are symmetrical around the diagonal. In these response matrices, each cell codes for whether the relevant pair of stimuli was placed within the same cluster (coded as 1) or placed in two separate clusters (coded as 0). The "telling people together" score is the average of all cells that code for pairs of stimuli that were veridically from the same identity. The closer to 1 this score is, the better participants were at correctly "telling people together," i.e., sorting different stimuli from the same person into the same cluster. The same process was implemented to compute the "telling people apart" indices, which are calculated by taking the average of all cells that code for pairs of stimuli that were veridically sampled from the two different identities. The closer the score to 0, the better participants were at telling people apart, i.e., not mixing stimuli

from different identity within a cluster. For "telling people together," we found no significant relationship, although there is a positive trend (Kendall's $\tau = .16$, $p = .127$, see Figure 2b). For telling people apart, we found a significant relationship across modalities (Kendall's $\tau = .37$, $p = .001$, see Figure 2c). We note that outliers may be driving the correlation for telling people apart. Therefore, we excluded three participants whose performance differed by more than 3 standard deviations from the mean on the respective "telling apart" measures and reran the correlation. Although the correlation got weaker, it remained significant (Kendall's $\tau = .27$, $p = .02$).

Overall, these results indicate that participants who performed well (as indicated by a smaller number of clusters) on aspects of the voice sorting task also performed well on the face sorting task and vice versa, although this relationship was not significant for "telling people together" indices. Results for "telling people apart" indices should be regarded with caution because of the limited variance in the measures due to near-perfect performance on "telling people apart" (i.e., participants only very rarely sorted stimuli from different identities into the same cluster, see Figure 2c).

Is there a relationship between performance on sorting tasks and discrimination tasks within modality?

To investigate whether there is a relationship between performance on discrimination tasks and the sorting tasks within modality, further correlation analyses were run. As the residuals for some variables were not normally distributed as determined via an inspection of Q-Q plots, we did not perform a linear regression analysis and thus diverge from our preregistered analysis plan. Instead, we again used Kendall's τ correlations to probe our research question.

To align these analyses with the analyses of sorting behaviour across modalities, we computed separate accuracy scores for the two matching tasks for all trials, "same identity" trials only and "different identity" trials only. These can then serve as counterparts to the "telling people together" (same trials) and telling people apart (different trials) indices for the sorting tasks. Both are measures of accuracy based on pairwise comparisons of either the same identity ("same identity" trials, "telling people together" index) or different identity ("different identity" trials, "telling people apart" index). For faces, no relationship between the total number of clusters created, and the mean accuracy on the GFMT was found, although there was a non-significant trend (Kendall's $\tau = .18$, $p = .100$, Figure 3a). Similarly, for voices, we found no significant relationship between the total number of clusters created and the mean accuracy on the BVMT (Kendall's $\tau = -.08$, $p = .434$, see Figure 3d).

We also correlated participants' "telling together" and "telling apart" indices in the sorting tasks with their accuracy for the "same" trials and "different" trials, respectively, in the modality-matched matching tasks. Here, we found a significant relationship between "telling people together" indices and accuracy on the "same" trials for both modalities (Voices: Kendall's $\tau = .23$, $p = .030$, Figure 3e; Faces: Kendall's $\tau = .23$, $p = .036$, Figure 3b). No relationship was found between "telling apart" indices and accuracy on the "different" trials, in either modalities (Voices: Kendall's $\tau = .01$, $p = .939$, Figure 3f; Faces: Kendall's $\tau = .04$, $p = .740$, Figure 3c).

The relationship between performance on a sorting task and a modality-matched matching task is thus less clear. Only "telling people together" indices significantly correlated with "same" trial accuracy, although this relationship is still relatively weak. We again note that the correlations with the "telling people apart" indices should be regarded with caution because participants performed well at "telling people apart."

Discussion

This study addressed two research questions to shed further light on the processes and strategies of identity processing in the context of identity sorting tasks: (1) Is there a relationship in participants' performance on identity sorting tasks across modalities? (2) Can sorting behaviour be linked to established tests of identity discrimination, thus suggesting common underlying processing mechanisms?

With regard to the first research question, we found significant correlations for voice and face sorting tasks across modalities. This was true for the number of clusters formed and for "telling people apart" indices, although we only found a non-significant trend for "telling people together" indices. Despite the modality differences, some overlap in the underlying processes is, therefore, apparent across for faces and voices.

What these modality-general processes or strategies might be remains unclear from the current experiment; we hypothesised that participants may use pairwise identity discrimination as a (modality-general) candidate strategy to complete sorting tasks when dealing with unfamiliar identities (cf. Kreiman & Sidtis, 2011; Lavan, Burston, & Garrido, 2019). However, we did not find a relationship between modality-matched performance on sorting identity tasks (as measured by perceived number of clusters) and the overall accuracy on the modality-matched matching tasks. Similarly, no relationship was apparent between "telling people apart" indices and accuracy on the trials including different identities. There were, however, weak but significant correlations between "telling people together" indices and accuracy on trials including the same identity, for both voices and faces.[1] The lack of a clear relationship suggests that neither voice sorting nor face sorting tasks tap into the same processes or strategies as discrimination tasks, making pure pairwise discrimination unlikely as a candidate strategy underpinning both face and voice sorting.

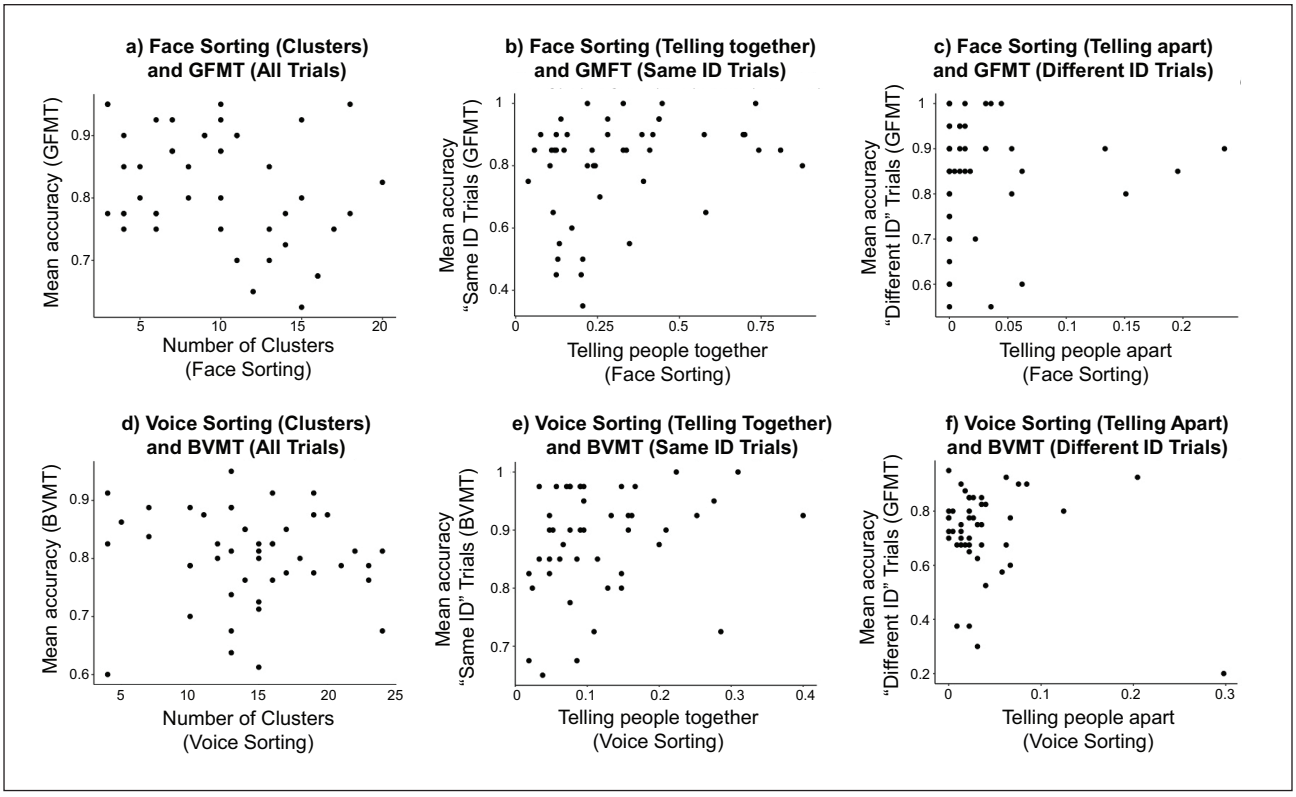

Figure 3. Scatterplots showing the measures from the sorting tasks against the measures from the modality-matched discrimination tasks. Panels a-c show scatterplots for face sorting and GFMT performance for all trials (Panel a), "same identity" trials (Panel b) and "different identity" trials (Panel c). Panels d-f show scatterplot for the voice sorting and BVMT accuracy, again for for all trials (Panel d), "same identity" trials (Panel e) and "different identity" trials (Panel f).

We note, however, the lack of a relationship could have alternatively arisen from the systematic differences in how much within-person variability is included in the sorting versus the discrimination tasks: the sorting tasks featured pronounced within-person variability across stimuli, while the discrimination tasks did not. This difference may affect difficulty and the strategies chosen by participants to complete the tasks and may therefore have obscured or changed any relationship that might have been observed with more closely matching. However, this interpretation does not fully fit our results. There are significant correlations for both voice and face tasks for "same trials" and "telling people together" performance. This is surprising because within-person variability has been shown to most dramatically affect participants' ability to "tell people together" (Jenkins et al., 2011; Lavan, Burston, & Garrido, 2019; Lavan, Burston, Ladwa, et al., 2019). Therefore, if the mismatch across tasks in within-person variability had obscured or changed the relationship between discrimination and sorting tasks, we should have been least likely to observe a relationship between "telling people together" indices and accuracy on the "same" trials. In the presence of the significant relationship, however, shared underlying mechanisms may be present for identity sorting and identity discrimination at least for "telling people together"—despite the differences in within-person variability and thus the differences in difficulty of these judgements across tasks.

Which perceptual strategies or processes may therefore underpin face and voice sorting tasks? Previous research has reported significant correlations between different measures of face sorting tasks (e.g., overall error rates, sensitivity) and the Cambridge Face Memory Test (CFMT; Balas & Saville, 2017; Short et al., 2017). These findings in conjunction with our findings could, therefore, suggest that good performance in face sorting tasks may be in fact more closely linked to the mechanisms underpinning good performance on a face learning or recognition task as opposed to a discrimination task. This is perhaps not too surprising, when considering that face sorting tasks have been successfully used as training tasks (Andrews et al., 2015; Murphy et al., 2015). Conceptualising sorting tasks as self-guided learning tasks instead of simple identity perception tasks is overall an intriguing possibility, and future work will need to determine whether and to what extent a relationship may be present for voice sorting and voice learning or recognition. If sorting tasks in both modalities could be linked to (perceptual) learning and recognition performance, individual differences in how readily participants can learn/recognise faces and voices may be a modality-general candidate mechanism underpinning individual differences in identity sorting ability.

Crucially, all significant relationships reported in this article are moderate to weak in strength (Kendall's $\tau < .38$). This corresponds well with previous reports of how

validated tests of voice identity processing correlate with each other and with tests of face identity processing (BVMT and GFMT: Pearson's $r = .24$; BVMT and the Glasgow Voice Memory Test, a voice learning and recognition test: Pearson's $r = .23$; Mühl et al., 2018). Correlations between different validated tests of face perception have, however, previously been shown to be slightly higher than what we find here (Pearson's r ranges between .2 and .53; McCaffery et al., 2018; Verhallen et al., 2017). Our reports of potential commonalities and shared mechanisms for voice and face processing, therefore, need to be contextualised by the strength of these relationships, indicating that modality- and task-specific mechanisms are additionally present. Therefore, all of our tasks seem to also tap into at least partially distinct aspects of identity perception—within modality, across task and across modality, within task (see also McCaffery et al., 2018 and Verhallen et al., 2017 for a discussion for faces).

This study is the first study to examine sorting tasks across modalities. As a starting point, we have opted to implement the methods of previously published sorting tasks from both modalities, by including 2 identities only and using highly variable stimuli. These design choices may have affected the findings of this study. For example, as in previous sorting studies, participants did not make many errors for telling people apart, leading to near-perfect performance in these measures. Future research may attempt to increase the difficulty of "telling people apart" by, for example, using less variable stimuli that sound less distinct from each other, both across and possibly also in within person (e.g., Stevenage et al., 2020). However, we note that changes to the stimuli will not only affect how well participants can tell people apart but will also affect "telling people together" (e.g., making this aspect of the task easier through decreased variability). Similarly, there is an argument to include more than two identities in sorting tasks. This would not only increase the generalisability of study but also make sorting more similar to other tasks used to measure identity perception. Most discrimination or recognition tasks include few exemplars of many identities. For sorting tasks, it would again be important to find a middle ground: the fewer items there are per identity in the context of including many identities, the less opportunity there is for participants to accurately "tell people together". Given the tendency of unfamiliar identities to be perceived as different identities, there is, therefore, a risk of participants not being able to "tell people together" at all. Overall, there is, however, much scope to explore how sorting behaviour is affected within and across modalities through changes to the stimuli, design, and task instructions.

More broadly, future research will also be required to further examine what underpins the tasks commonly used to probe identity perception, to map out how these tasks relate to each other, and crucially, to determine how closely they reflect aspects of identity processing from voices and faces outside of laboratory tasks. Overall, this study has further contextualised identity sorting paradigms within the set of tasks routinely applied to probe identity perception in voices and faces. We provide some evidence that identity sorting tasks in different modalities may tap into partially similar mechanisms, although the relationship to other tasks remains unclear.

Declaration of conflicting interests

The author(s) declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The author(s) disclosed receipt of the following financial support for the research, authorship, and/or publication of this article: This work was supported by a Research Leadership Award from the Leverhulme Trust (RL-2016-013) awarded to C.M.

ORCID iD

Nadine Lavan https://orcid.org/0000-0001-7569-0817

Note

1. We also note that the mean scores for the BVMT and the GFMT were correlated at Pearson's $r = .2$ ($p = .175$). Although this correlation is not significant, the strength of the correlation replicates Mühl et al.'s (2018) study, where a correlation of Pearson's $r = .24$ ($p = .004$ in their sample of 149 participants) is reported.

References

Andrews, S., Jenkins, R., Cursiter, H., & Burton, A. M. (2015). Telling faces together: Learning new faces through exposure to multiple instances. *Quarterly Journal of Experimental Psychology, 68*(10), 2041–2050. https://doi.org/10.1080/17470218.2014.1003949

Anwyl-Irvine, A., Massonnié, J., Flitton, A., Kirkham, N., & Evershed, J. (2020). Gorilla in our midst: An online behavioral experiment builder. *Behavioural Research Methods*, *52*, 388–407.

Balas, B., & Saville, A. (2017). Hometown size affects the processing of naturalistic face variability. *Vision Research*, *141*, 228–236.

Barsics, C. G. (2014). Person recognition is easier from faces than from voices. *Psychologica Belgica*, *54*(3), 244–254.

Belin, P., Bestelmeyer, P. E., Latinus, M., & Watson, R. (2011). Understanding voice perception. *British Journal of Psychology*, *102*(4), 711–725.

Boersma, P., & Weenink, D. (2013). Praat: Doing phonetics by computer [Computer program].

Bruce, V., & Young, A. (1986). Understanding face recognition. *British Journal of Psychology*, *77*(3), 305–327.

Burton, A. M. (2013). Why has research in face recognition progressed so slowly? The importance of variability. *Quarterly Journal of Experimental Psychology*, *66*(8), 1467–1485.

Burton, A. M., White, D., & McNeill, A. (2010). The Glasgow Face Matching Test. *Behavior Research Methods*, *42*(1), 286–291.

Jenkins, R., White, D., Van Montfort, X., & Burton, A. M. (2011). Variability in photos of the same face. *Cognition*, *121*(3), 313–323.

Kreiman, J., & Sidtis, D. (2011). *Foundations of voice studies: An interdisciplinary approach to voice production and perception*. John Wiley & Sons.

Laurence, S., Zhou, X., & Mondloch, C. J. (2016). The flip side of the other-race coin: They all look different to me. *British Journal of Psychology*, *107*(2), 374–388.

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*(3), 576–593.

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*(9), 2240–2248.

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.

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, 72(9),* 2240–2248. https://doi.org/10.1111/bjop.12416

Maurer, D., & Werker, J. F. (2014). Perceptual narrowing during infancy: A comparison of language and faces. *Developmental Psychobiology*, *56*(2), 154–178.

McCaffery, J. M., Robertson, D. J., Young, A. W., & Burton, A. M. (2018). Individual differences in face identity processing. *Cognitive Research: Principles and Implications*, *3*(1), e21.

McLeod, A. (2011). Kendall: Kendall rank correlation and Mann-Kendall trend test. R package version 2.2. https://CRAN.R-project.org/package=Kendall.

Megreya, A. M., & Burton, A. M. (2006). Unfamiliar faces are not faces: Evidence from a matching task. *Memory & Cognition*, *34*(4), 865–876.

Mühl, C., Sheil, O., Jarutytė, L., & Bestelmeyer, P. E. (2018). The Bangor Voice Matching Test: A standardized test for the assessment of voice perception ability. *Behavior Research Methods*, *50*(6), 2184–2192.

Murphy, J., Ipser, A., Gaigg, S. B., & Cook, R. (2015). Exemplar variance supports robust learning of facial identity. *Journal of Experimental Psychology: Human Perception and Performance*, *41*(3), 577.

Papcun, G., Kreiman, J., & Davis, A. (1989). Long-term memory for unfamiliar voices. *The Journal of the Acoustical Society of America*, *85*(2), 913–925.

Redfern, A. S., & Benton, C. P. (2017). Expressive faces confuse identity. *i-Perception*, *8*(5), 2041669517731115.

Short, L. A., Balas, B., & Wilson, C. (2017). The effect of educational environment on identity recognition and perceptions of within-person variability. *Visual Cognition*, *25*(9–10), 940–948.

Stevenage, S., Symons, A., Fletcher, A., & Coen, C. (2020). Sorting through the impact of familiarity when processing vocal identity: Results from a voice sorting task: Familiarity and voice sorting. *Quarterly Journal of Experimental Psychology*, *73*, 519–536. https://doi.org/10.1177/1747021819888064

Verhallen, R. J., Bosten, J. M., Goodbourn, P. T., Lawrance-Owen, A. J., Bargary, G., & Mollon, J. D. (2017). General and specific factors in the processing of faces. *Vision Research*, *141*, 217–227.

Yovel, G., & Belin, P. (2013). A unified coding strategy for processing faces and voices. *Trends in Cognitive Sciences*, *17*(6), 263–271.

Zhou, X., & Mondloch, C. J. (2016). Recognizing "Bella Swan" and "Hermione Granger": No own-race advantage in recognizing photos of famous faces. *Perception*, *45*(12), 1426–1429.