**Voice preferences across contrasting singing and speaking styles**

**Camila Bruder1, Klaus Frieler1 & Pauline Larrouy-Maestri1,2**

1 Max Planck Institute for Empirical Aesthetics, Frankfurt am Main, Germany

2 Max Planck-NYU Center for Language, Music, and Emotion (CLaME), USA & Germany

**Abstract**

Voice preferences are an integral part of interpersonal interactions and shape how people connect with each other. While a large number of studies has investigated the mechanisms behind (spoken) voice attractiveness, very little research was dedicated to other types of vocalizations. In this Registered Report, we propose to investigate voice preferences with an integrative approach. To this end, we will use a newly recorded and validated stimulus set of contrasting vocalizations by 22 highly trained female singers speaking and singing the same material in contrasting styles (sung as a lullaby, as a pop song or as an opera aria; and spoken aloud as if directed to an adult audience and as if directed to an infant). We will ask participants to rate these vocalizations in terms of how much they liked them; and we will compare the amount of shared taste (i.e., how much participants agree in their preferences) across styles. This approach will allow us to characterize voice preferences in a broader framework, taking into account the variability in both the uses and functions of vocalizations and in participants’ aesthetic appreciation of them, in order to better understand a question central to human experience.

**Keywords: aesthetics, liking, voice perception, voice attractiveness, vocalization**

1. **Introduction**

The voice is highly significant to human experience. Voice selective areas have been described in the human cortex (Belin et al., 2000), and there is evidence for neural populations that respond selectively to songs (Norman-Haignere et al., 2022). Melodies are easier to remember when presented vocally than when played on a piano, banjo or marimba (Weiss et al., 2012), even for trained pianists (Weiss et al., 2015). The voice is also incredibly flexible: it can serve a myriad of functions, and it sounds differently depending on its current use. Besides its obvious functions, that is, to express and exchange semantic meaning via speech, the voice conveys a wide range of non-verbal information. A person’s voice may cue the speakers’ body size and shape, health and age (Pisanski et al., 2014, 2016). During speech, fluctuations in voice intonation (generally known as speech prosody or “the melody of speech”) may convey intent (Hellbernd & Sammler, 2016, 2018), emotional states (Banse & Scherer, 1996; Larrouy-Maestri, et al., 2023; van Rijn & Larrouy-Maestri, 2023), and even personality traits (Goupil et al., 2021; McAleer et al., 2014; Scherer, 1978). Across cultures, certain features consistently distinguish song and speech (Albouy et al., 2023; Ozaki et al., 2022), and all studied cultures have some form of singing (Mehr et al., 2019; Savage et al., 2015). Both speech and song sound differently when directed to infants (Cox et al., 2022; Fernald, 1989; Hilton & Moser et al., 2022), and also in the case of singing, different uses of the voice are associated with different functions (e.g., loud, rhythmic singing in play songs to entertain, versus unaccompanied, soft and quiet singing of lullabies to soothe an infant – Rock et al., 1999; Trehub & Trainor, 1998). Given the voice’s multiple facets, how can we understand individuals’ enjoyment of voices in different contexts? How shared are our preferences across different types of vocalization?

In the case of spoken voices, voice attractiveness is thought to signal the speaker’s physical fitness of the speaker to potential mates. Voice attractiveness has been shown to covary with sexually dimorphic traits: individuals with more attractive voices also have larger shoulder-to-hip ratios (for males) or smaller waist-to-hip ratios (for females) (Hughes et al., 2004). Voice attractiveness is also related to certain acoustic characteristics (e.g., higher fundamental frequency and more spread formands preferred for female's voices, see Collins, 2000). More nuanced instances of voice attractiveness have also been described, with, for example, conformance to community speech norms increasing voice attractiveness ratings (Babel et al., 2014). A different (though related) line of research has proposed a role for averageness and typicality in voice preferences. Bruckert et al. (2010) found that morphed, averaged voices (which are smoother and have higher harmonics to noise ratios) were rated as more attractive than most of the individual voices presented to participants. Accordingly, average ratings of voice attractiveness have been reported to be highly correlated with ratings of stereotypicality (Babel & McGuire, 2015), or negatively correlated with ratings of atypicality/distinctiveness in relation to an average voice (Zäske et al., 2020) – though also see Mook & Mitchel (2019) for a study where the positive effect of averageness (via morphing) on voice attractiveness was not replicated.

In the case of the singing voice, fewer studies have investigated the mechanisms behind our preferences. Bruder et al. (2023) recently observed that participants’ ratings of ten different perceptual attributes of the voices (i.e., articulation, breathiness, pitch accuracy, loudness, tempo etc) were better predictors of listeners' liking of pop voices than computationally extracted acoustic features commonly used to describe voices such as jitter, shimmer, vibrato rate and extent and harmonics-to-noise ratio. Importantly, while preferences were highly idiosyncratic, as indicated by the low interrater agreement in liking ratings (Krippendorff’s alpha was .16), some average preferences emerged for some voices, as shown by highly correlated averaged liking ratings between the two experiments conducted (one with German and one with US participants). This suggests the emergence of robust average preferences amidst large individual differences in how participants perceive and like singing voices. Based on the literature on spoken voice attractiveness, it is difficult to say how much individual differences lie behind the typically reported average preferences. Valentova et al. (2019) reported high correlations between average attractiveness ratings of spoken and sung vocalizations (of “Happy Birthday” and national anthems) produced by the same subjects and argued that spoken and sung voice may work as “backup signals” that convey the same information about a subject’s physical fitness. Because they used Cronbach’s alpha as a measure of interater agreement, a reportedly problematic measure (it is inflated by larger sample sizes and ignores within-person variability – Hönekopp, 2006; Kramer et al., 2018), the relationship between voice attractiveness of singing and speaking (as well as the role of individual differences in this relationship) needs to be clarified.

Here we investigate the aesthetic appeal of a set of contrasting vocalizations – singing and speaking – in an integrative manner. We adopt an interactionist approach (e.g., Wassiliwizky & Menninghaus, 2021) within the larger framework of empirical aesthetics – an approach that takes into account aspects of the stimuli as well as subjective, internal factors related to the person making the aesthetic evaluation. This means that, in addition to examining mean liking ratings as an indication of average preferences, we will also examine the variability in these ratings across participants. One way to assess the relative contribution of individual versus shared factors to preferences is to measure the amount of shared taste across participants (e.g., Germine et al., 2015; Hönekopp, 2006; Leder et al., 2016; Vessel & Rubin, 2010). Here, we focus on the variability of aesthetic judgements across contrasting vocalization styles. We propose to use a newly recorded and validated stimulus set of naturalistic but controlled a cappella vocal performances. Twenty-two female classical singers performed different melody excerpts in three contrasting singing styles – as a lullaby, as a pop song and as opera aria; and read the corresponding lyrics aloud in two contrasting ways – as if speaking to an adult audience and as if speaking to an infant. The three singing styles (i.e., contrasting sounding vocalizations) were chosen as a pragmatic way to have the same singers produce contrasting performances in different styles without having to learn another specific singing technique (such as belting). The five proposed vocalization styles can be seen as a subset of possible categories of human vocalizations, sampled from a multidimensional continuum – for example, from the speech-music continuum described by Phillips (2023), or from the “musilanguage continuum” described by Brown (2000) and extended by Leongómez et al. (2022, Figure 1b).

Further, the five proposed styles of vocalization allow for an interesting comparison with findings from the visual domain. Using a correlational measure of agreement (“mean-minus-one”, MM1) and variance partitioning analysis, Vessel and colleagues (2014, 2018), found a higher degree of shared preferences for images of faces and landscapes than for images of exterior architecture and interior architecture, and little shared taste for artworks (which reflected strong individual differences or idiosyncratic taste). They argued that naturally occurring types of stimuli, such as landscapes and human faces, have uniform behavioral relevance, which results in shared semantic meaning that is highly conserved across individuals. This would lead to similar aesthetic experience: for instance, participants tend to agree in their higher liking of an image of an oriental garden (associated with leisure) and in their lower liking of an image of a parking lot (associated with work), even when images are controlled for low-level visual features (example taken from Vessel et al., 2018). On the other hand, artifacts of human culture, such as architecture and artwork, lack this uniform behavioral relevance, and allow for the expression of individual subjects’ idiosyncratic taste. In fact, the authors suggest that this opposition between artificial (human-made) and natural (non-human-made) categories may be a fundamental organizational principle for how humans aesthetically evaluate objects (Vessel et al., 2018). If this is indeed the case, then natural types of stimuli should also elicit more shared taste in the auditory domain.

Applying this rationale to the auditory domain, and to voices in particular, and quantifying the amount of shared taste for these five types of vocalizations should help us characterize voice preferences in an integrative way while also acknowledging individual differences in these preferences. We posit that, even though all these vocalization categories are natural (in the sense that they are produced by the human vocal apparatus), and behaviorally relevant, their behavioral relevance is not uniform across individuals. Drawing a parallel with the visual domain, we argue that lullabies constitute a more “natural” (in the sense of universal) kind of singing than the pop and operatic styles; and thus predict more shared taste (that is, that participants will agree more in terms of which voices they prefer) for lullabies than for both other styles of singing. The rationale behind this is based on findings that lullabies are ubiquitous and cross-culturally recognizable (Mehr et al., 2019; Trehub et al., 1993; Yurdum et al., 2023) and arguably evolutionarily important (e.g., Dissanayake, 2000; Mehr & Krasnow, 2017), hence a more “natural” (and universal) kind of singing than the pop and operatic styles. On the other hand, we expect to find more idiosyncratic taste for operatic singing, as the least “natural” kind of singing of the three (i.e., related to a very specific technique, and appreciated by a very specific audience). Concerning the two speaking styles, we argue that adult- and infant-directed speech are both “natural” and highly behaviorally relevant in their own way. We thus expect to observe equivalent amounts of shared taste for both speaking styles.

* 1. **Study aims and significance statement**

This study aims to empirically investigate and characterize voice preferences across a varied but controlled stimulus set of contrasting vocalizations. Given the scarcity of previous empirical research on singing voice preferences, our approach is partially and inevitably exploratory. We base our theoretical framework in a parallel with the visual domain, where enough pioneering work has been done to inform our predictions (Table 1).

**Table 1. Registered Report Design Planner**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Question** | **Hypothesis** | **Sampling**  **plan** | **Analysis Plan** | **Rationale for**  **deciding the**  **sensitivity of the**  **test for confirming**  **or disconfirming**  **the hypothesis** | **Interpretation**  **given different**  **outcomes** | **Theory that could**  **be shown wrong by**  **the outcomes** |
| 1) Is there a difference in the amount of shared taste across contrasting vocalization styles?  (see 1.2.1). | H1A): there will be more shared taste for lullaby than for pop; and for pop than for operatic singing (MM1  lullaby > pop > opera). | 60 participants rating each stimulus in terms of liking. | Comparison of agreement  (MM1 measures) between the  three singing styles with three  pairwise comparisons (paired  t-tests, one-tailed). | Sample size determined by power analysis assuming a SESOI of .5, power of .95 and adjusting alpha for three comparisons (please see 1.3.1 for details). | Only if all three planned (and directional) pairwise comparisons are significant, the results will support our hypothesis of higher shared taste for more “natural”/ universal (lullabies) than for more “artificial” (operatic) kinds of singing, with pop in an intermediary position. | Outcomes would not falsify an established theory, but would suggest that predictions grounded on findings in the visual domains do (not) generalize to the auditory one, namely for vocalizations. |
| H1B) there will be equivalent amounts of shared taste for adult- and infant -directed speech  (MM1 AD = ID) | Equivalence testing of  agreement (MM1 measures)  between the two speaking  styles to assess if the effect is  statistically equivalent to  zero (|effect size| < .5). | The sample size of 60 participants is enough for power =.99 in the equivalence test with equivalence bounds between -.1 and .1 (our SESOI) (please see 1.3.1 for details). | If the equivalence test is significant (and the null-hypothesis test is not), the result will support that there is no meaningful difference in agreement between both styles. If both the NHST and the equivalence test are not significant, the result will be inconclusive. |
| 2) On average, will the same performers be preferred across styles?  (see 1.2.2) | H2: Preferred performers will differ across vocalization styles. | 60 participants rating each stimulus in terms of liking | Based on mean liking ratings for vocalizations by each performer in each style, we will use MM1 also as a measure of interstyle agreement: if preferences are highly consistent across styles, interstyle agreement should be high (see 1.3.2). | We adopt a threshold of .8 (including the confidence interval) to consider preferences highly consistent across styles. | We consider interstyle agreement high (i.e., preferences highly consistent across styles) if the lower bound of the confidence interval of interstyle agreement is equal to or higher than .8 | The finding that preferences are not highly consistent across styles (i.e., that different voices are preferred for different styles) would contradict the idea that singing and speaking voice work as “backup” signals, conveying the same information about a person’s physical fitness. |

* 1. **Hypotheses**
     1. ***Hypothesis regarding the amount of shared taste across styles (Question 1)***

Does interrater agreement (as measured by “mean-minus-one”, MM1; see Methods for details) in liking ratings vary depending on the type of vocalization? Concerning the singing performances (Hypothesis 1A), expanding on Vessel and colleagues’ (2010, 2014, 2018) findings in the visual domain, we expect a higher degree of shared taste (higher interrater agreement) for aesthetic ratings of lullabies, a more “natural” kind of singing; and lower shared taste for ratings of operatic singing (a more technical and specific type of singing), with intermediary values for pop singing. Concerning the speech performances, since both adult- and infant-directed speech seem highly behaviorally important and “natural”, we predict equal amounts of shared taste for both of them.

**1.2.2 *Hypothesis regarding average preferences for some singers (Question 2)***

If some voices are “fundamentally” more likeable, this should happen consistently across styles, that is, the same singers/speakers should be liked the most (or the least) across styles. That is to say, if sexual selection accounts of voice attractiveness – suggesting that singing and spoken voice work as “backup” signals, displaying the same (i.e., redundant) information about an individual’s physical fitness – are correct, then the rankings of favorite voices should be the same across all styles, with the “best” voices consistently preferred. On the other hand, differences in which singers are preferred across styles would suggest that some performers and/or voice qualities were more adequate or conformant to some styles than to others, that is, that style-specific influences are determinant of voice preferences.

* 1. **Analysis plan and sample size justification**

Our dependent variable is the participants' liking rating. Our independent variable corresponds to vocalization style (with five levels). Note that we use the terms liking and aesthetic preferences in interchangeable ways; we interpret higher liking ratings for a certain singer as indication that she was “preferred”, even though we don’t have an explicit pairwise comparison design. Whenever requirements are met, we plan to use parametric tests, aiming at higher power (otherwise adjusting to non-parametric alternatives; these changes are specified in our proposed analyses code).

* + 1. ***Comparing the amount of shared taste across styles***

To test hypothesis 1A, we propose to compare MM1 measures between the three singing styles with three pairwise comparisons (paired t-tests, one tailed, adjusting p-values for multiple comparisons with the Holm method; or using Wilcoxon test if a nonparametric alternative if necessary). We expect MM1 values to be higher for lullabies than for pop performances, and higher for pop than for operatic performances (and, logically, also higher for lullabies than for operatic singing). Concerning the speech performances, we expect to find the same amount of agreement in adult- and infant- directed performances. To test hypothesis 1B, we propose to run equivalence testing of MM1 values for these two speech styles using the two one-sided tests (TOST) procedure (Caldwell, 2022; Lakens, 2017).

Power analysis was informed by data from previous experiments. First, to illustrate how much MM1 values vary across contrasting categories in the visual domain, Vessel et al. (2018) reported MM1 values of 0.31 (SD = 0.17) for (images of) artwork, 0.38 (SD = 0.18) for exterior architecture, 0.40 (SD = 0.12) for interior architecture, 0.6 (SD = 0.15) for landscapes and 0.85 (SD = 0.12) for faces (we recalculated these MM1 and standard deviation estimates ourselves based on their openly available data, since the original paper reports confidence intervals instead of SD). Second, we calculated MM1 for the liking ratings of two pop melodies described in Bruder et al. (2023) (Supplementary Information, Supplementary Figure S1). In the first experiment, where participants were tested online, MM1 was 0.46 (SD = 0.22) for 146 “consistent” participants (with test-retest Pearson correlation scores equal or superior to .5 in a subset of 16 repeated trials). In Experiment 2, where 42 participants were tested in the lab, MM1 was 0.42 (SD = 0.17). So for our calculations we used as reference an intermediary MM1 value of 0.44 (SD = 0.2). To estimate our sample size, we first stipulated the minimum difference in MM1 values we should be able to statistically detect when comparing styles, that is, our smallest effect size of interest (SESOI – e.g., Lakens, 2014). We stipulated our SESOI to be a .1 difference in overall MM1 values per style. This corresponds to an effect size of dz = 0.5 [calculated using the esc\_mean\_sd function from the *esc* R package (Lüdecke, 2019) and the values: mean group 1 = 0.44, mean group 2 = 0.34, SD group 1 = 0.2, SD group 2 = 0.2, correlation for within-subject designs = .5; see accompanying R script for power analyses]. Concerning the correlations between repeated measures, we set it to .5 as a conservative estimate, since we have no grounded indication of a more appropriate value to use. But note that studies with reaction times and rating scales reportedly have high intercorrelations between the levels of a repeated-measures factor (Brysbaert, 2019), and it does make sense to expect participants to be consistently more or less “generous” in their use of the rating scale for liking (for instance, in the mentioned previous study, participants scoring higher on the personality trait Agreableness systematically gave out higher liking ratings – Bruder et al, 2023).

To test hypothesis 1A, the power analysis showed that, considering our SESOI of d = 0.5 and an alpha of .017 (adjusted for three comparisons), we will reach power = .95 with a sample size of 60 participants (paired, two-sided t-tests, calculated with the pwr.t.test function from the pwr R package – Champely, 2020). Note that resorting to the nonparametric alternative of Wilcoxon tests should not lead to great loss of power. Using the MKpower function from the MKpower R package (Kohl, 2023), we estimated that, based on a sample size of 60 participants, we would have power of approximately .9 to detect the difference in MM1 values mentioned above (i.e., the difference between 0.34 and 0.44, with SD = 0.2, and stipulating the same conservative alpha of.017). To test hypothesis 1B, the sample size of 60 participants ensures very high power for the equivalence test. Using the power\_t\_TOST function from the TOST package (Caldwell, 2022; Lakens, 2017) and setting the equivalence bound between -.1 and .1(based on our SESOI), with SD of .2 and conventional alpha of .05, a sample size of 60 participants would ensure power of .99. Please see accompanying R file for code for the power analysis. Please refer to the Methods Section (2.4.1) for details on how to compute MM1 and please see the Supplementary Information (Supplementary Figure S2) for an illustration of this analysis conducted on simulated datasets with increasing amounts of interrater agreement.

**1.3.2** ***Assessing the consistency of average preferences across styles***

For question 2, analyses are based on mean liking ratings across all participants and pooling (averaging) values of the two testing sessions. We will compute a grand average of liking ratings for each singer in each vocalization style. If the same voices are preferred consistently across all styles, all pairwise correlations between styles should be high. We propose to also use MM1 to measure this agreement across the five styles (hereafter referred to as “interstyle agreement”). A threshold will be set at .8 (including 95% confidence intervals; see Methods for details), in which case average preferences will be considered highly consistent across styles. Alternatively, an interstyle agreement value (including the confidence interval) inferior to .8 will indicate that preferences were not highly consistent (i.e., preferences varied depending on the style). The rationale for choosing the value of .8 as our threshold is based on a general recommendation of this value as a minimally acceptable level of reliability (Krippendorff, 2004, p. 241) and on our own data simulations, which allowed us to observe that this value indeed corresponds to a high level of consistency in preferences across styles. Please see the Methods section (2.4.2) for details on how to compute interstyle agreement and please see the Supplementary Information (Supplementary Figures S3 and S4) for a simulation-based demonstration of this solution. Note that this is a descriptive approach that does not fit into conventional hypothesis testing based on p-values, nor does it allow for power analysis. However, it does allow us to test our prediction: we will conclude that preferences varied depending on the style if the upper bound of the confidence interval is equal or superior to .8).

**2. Method**

**2.1 Participants**

Participants will be recruited from the participant database of the Max Planck Institute for Empirical Aesthetics, in Frankfurt, Germany, which consists of adults, mostly lay listeners, with a preponderance of students and retired persons. While we acknowledge that this convenience sample shares the generalizability limitations of most studies sampling from “WEIRD” populations (White, Educated, Industrialized, Rich, and Democratic - Henrich et al, 2010), we attempt to enhance representativity of the sample by examining participants with a large range of musical expertise (i.e., not recruiting only musically trained participants) and keeping balanced genders in the recruited sample. Participants will be rewarded for their participation at a rate of 7€ per half hour. The only exclusion criterion for participation in data collection will be reported hearing impairments. We will exclude from analyses data from participants whose scores are the same for more than 85% of trials. This is specified in our analysis code. The experimental procedure was ethically approved by the Ethics Council of the Max Planck Society (No 2017\_12), and will be undertaken with written informed consent of each participant.

**2.2 Materials**

**2.2.1 *Questionnaires for collection of participant-related data****.*

In the end of the first testing session, participants will be asked the following information, to be used in exploratory analyses:

a) demographic questions about age, languages spoken, gender (female / male / non-binary / prefer not to disclose / prefer to self-describe), and sexual orientation (heterosexual or straight / gay or lesbian / bisexual / prefer not to disclose / prefer to self-describe).

b) questions about their experience while doing the experimental task : 1) Did you perform the task conscientiously? 2) Did you recognize the language spoken and sung in the stimuli (if yes, which was it)? 3) Do you have any comments about your experience while doing the task? 4) During the experiment, each block of trials contained different types of vocalization. How would you label the five types of vocalization you listened to?

c) the 18-items version of the general Music Sophistication subscale from the Goldsmiths Music Sophistication Index (Gold-MSI; Müllensiefen et al., 2014), as computed with the Gold-MSI configurator (https://shiny.gold-msi.org/gmsiconfigurator). The Gold-MSI is a self-report measurement instrument to assess musical skills and behaviors in the general population.

d) a short questionnaire about music preferences, asking participants how much they like to listen to certain styles of music (pop, opera, rock, world music), and on average how many hours they spend per week listening to that style of music.

***2.2.2 Stimulus set***

The stimuli proposed for this study come from a newly recorded stimulus set comprising singing and speech performances. Detailed information about the singing performances is presented in Bruder and Larrouy-Maestri (2023) [and a preprint thoroughly describing the whole stimulus set will be available soon]. In what follows, we summarize the findings that are relevant to the current study. The stimulus set consists of vocalization by 22 highly trained Brazilian female classical singers (16 sopranos, 6 mezzo-sopranos, aged from 22 to 45 years old, M = 32.5, SD = 7.1), with vocal training ranging from 4.5 to 27 years (M = 12.9 years, SD = 6). Singers were recorded in a professional music recording studio in Sao Paulo, Brazil, and performed the same melody excerpts (the first phrase of different songs) as a lullaby, as a pop song, or as an opera aria, and spoke the corresponding lyrics aloud as if directed to an adult audience and as if directed to an infant. Singing stimuli are on average 9 seconds long, and speech stimuli are on average 5 seconds long. The exact instructions given to singers during the recording session were as follows. For lullaby singing: imagine you have a baby on your chest and you want to make it sleep. For pop singing: imagine you are performing a pop song on a microphone. For operatic singing: imagine you are on stage performing an opera aria. For speaking the text aloud: imagine you are reading out loud the translation of the lyrics from something you have just performed on stage. For posed infant-directed speech: read the same text out loud but this time imagine you are talking to a baby or a small child. Operatic singing was performed with higher pitch than pop and lullaby (one fourth or one fifth higher, depending on the range of the melody), aiming at naturalistic performances and considering that operatic singing typically has higher pitch than both other styles. The singing stimuli were validated in lab experiments (two forced-choice tasks, N = 25 participants per stimulus or higher) where participants were asked to indicate, in each trial, if a given singing performance sounded like a lullaby, a pop song, or an opera aria; and if a given speech performance was directed to an adult or to a baby/child. For the subset of stimuli to be used in the current study, the proportion of correct recognition was higher than 67% for all styles. The proposed subset consists of three melody excerpts: “Nana Nenê”, originally a lullaby; “Chove Chuva”, originally a MPB (Música Popular Brasileira, a genre of Brazilian popular music) song by Brazilian artist Jorge Ben Jor (b. 1939–); and Melodia Sentimental, originally an art song by Brazilian classical composer Heitor Villa-Lobos (1887–1959). This leads to 330 performances (22 singers performing three melody excerpts in five vocalization styles). We chose to use performances with lyrics in Brazilian Portuguese (a version of each performance with /lu/ sound is also available in the dataset) to preserve the phonetic variability of speech. Also, for the present study, we plan to loudness normalize all stimuli to -23 LUFS, thus controlling for possible influences of loudness in liking ratings. Please Supplementary Information (Supplementary Figure S4) for sheet music of the melody excerpts. The stimuli used in the present work are currently available at https://osf.io/8k4af/?view\_only=506d243a6e7a4d3680c81e696ca81025.

**2.3 Procedure**

The experimental session will run as follows: after general instructions, the experiment will start with three training trials to familiarize participants with the task. Participants will be asked to rate how much they liked each stimulus on a scale of 1 (not at all) to 9 (a lot), by clicking with the mouse on the corresponding number on the scale presented on the computer screen. In each trial, a “Next” button will become visible only after the stimulus ends. Clicking on this button will prompt the next trial and playing of the next stimulus. The experiment will be divided in five blocks, one for each style of vocalization. Each block will comprise 66 trials, corresponding to one performance by each of the 22 singers for each of the three melodies, presented in a randomized order. The order of these blocks will be counterbalanced across participants. Participants will complete the experiment at their own pace and are expected to need around one hour in total. Breaks will be proposed between blocks. At the end of the experiment, participants will be asked to complete the questionnaires mentioned above. Participants will complete two testing sessions (test-retest), no longer than 14 days apart from each other, and preferably one week apart. The rationale behind stipulation of this time interval is as follows. According to Allen & Yen (1979), two aspects need to be considered when testing reliability with the test-retest method: the possibility of learning, carry-over, or recall effects (i.e., that the first testing may influence the second); and the possibility of a change in status of the measured trait between sessions (e.g., change in a cognitive ability in children). None of these aspects is of particular concern in our paradigm. Given the high number of stimuli (330), the possibility of participants remembering their answers from one session to the next is probably negligible and music abilities and engagement seem to be relatively stable among adults. Müllensiefen et al., (2014) report a test-retest correlation of *r* = .86 or higher for all subscales of the Gold-MSI self-report inventory, with participants tested on average 23 days apart (*SD* = 9.2); and George & Ilavarasu (2021) report test–retest reliability of *r* = .87 for a 15 day interval and *r* = .91 for a one month interval in the validation of their Music Receptivity Scale. We thus privilege pragmatic aspects of data collection in our decision to propose the test-retest interval. The second session will be identical to the first one, with the exception that no questionnaires will have to be filled. Stimuli will be presented and data will be recorded in the experimental platform Labvanced (Finger et al., 2017). The whole experiment will be conducted in German. Participants will be tested in the laboratories of the Max Planck Institute for Empirical Aesthetics, in Frankfurt, Germany.

**2.4 Data analyses**

All analyses will be performed using R Statistical Software (R Core Team, 2021) and R Studio (RStudio Team, 2022). Please see accompanying .Rmd scripts for code to run all of the proposed analyses.

**2.4.1 *Shared taste or interrater agreement***

We will measure interrater agreement (or shared taste) by computing the “mean-minus-one” (MM1) measure, a leave-one-out type of correlational agreement measure (Vessel et al., 2014, 2018). To compute MM1, a Pearson correlation is computed between a given participant’s liking ratings for the stimulus set and the average ratings of all other participants. This is done for all participants in the sample. The resulting individual correlations are then converted to z scores (Fisher’s r-to-z transform), averaged, and converted back into an r score (z-to-r transform) for easier interpretation of the final MM1 measure. This method has been shown to result in less biased estimates than averaging raw correlations (Corey et al., 1998).

**2.4.2 *Consistency of preferences for some singers (interstyle agreement)***

As outlined in our analysis plan, we propose to use MM1 also to measure interstyle agreement, that is, to assess how consistent were average preferences for some singers across the different styles of vocalization. Based on grand averages of liking ratings for each singer in each vocalization style, we will compute interstyle agreement using the same code used to compute MM1 interrater agreement, but with the five styles as the “raters” who “judge” . That, is, a Pearson correlation will be computed between a given style’s (grand average) ratings of singers and the average of the (grand average) ratings of the four other styles; the same will be done for all styles; the resulting five individual correlations will be converted to z scores, averaged, and converted back into an r score, which will be our “interstyle MM1”, used to assess how much styles “agree” with each other. Interstyle agreement will be considered high if it is equl or superior to .8 (including it’s 95% confidence interval).

**2.4.3 *Exploratory Analyses***

Our design with two testing sessions allows us to conduct interesting complementary analyses: we will assess intrarater agreement, and we will compute the beholder index (Hönekopp, 2006). Aditionally,methodological discussions about , we will also report the more widely-known measures of Krippendorff’s alpha and Intraclass Correlations. For completeness, we plan to compute these two last measures both for interrater agreement in Question 1 and for interstyle agreement in Question 2.

**2.4.3.1 *Intrarater agreement.*** For each participant, Pearson correlation scores will be computed based on the ratings of 330 stimuli in the first and second sessions, as a measure of test-retest intra-rater agreement. Measuring how self-consistent participants are is vital to understand how much participants can agree with each other in the first place. If the interrater agreement is low but the intrarater agreement is high (that is, participants’ ratings are consistent between test and retest), one can be confident that ratings were not random, but instead indicate a preponderance of private or idyosincratic taste. A similar pattern was reported by Bruder et al. (2023), where test-retest agreement was high for about half of the online participants (*r*test-retest ≥ .5), but very low interrater agreement indicated highly idiosyncratic preferences for pop singing.

***2.4.3.2 Variance component analysis and beholder index.*** While we will focus our hypothesis testing on MM1, we will also report and jointly discuss the beholder index (Hönekopp, 2006) as a complementary measure of agreement. Based on generalisability theory (Brennan, 2001), Hönekopp (2006) proposed the beholder index as a measure of the amount of private taste in ratings of attractiveness of face stimuli. To estimate the beholder index (“*bi* ”), one needs at least two sets of ratings by each rater. Variance components are computed and *bi* is estimated as a ratio between the amount of private taste and the total meaningful (i.e., accounted for or non-residual) variance. Beholder index estimates should thus mirror MM1 estimates (that is, when MM1 is high, *bi* should be low, and vice-versa). To estimate *bi,* one first needs to conduct a variance component analysis (VCA). One convenient way of conducting VCA is to compute a multilevel model (using the lmer function from the lme4 package in R – Bates et al., 2015) with random intercepts for stimuli, participants, blocks, and all two-way interactions between these terms (Martinez et al., 2020). VCA allows one to compare the variance in different clusters, which are components that are similar across measurements, such as raters, stimuli or occasions, and are treated as if they are sampled from a random population (Martinez et al., 2020).In this context, the variance in the Stimulus cluster is related to shared taste; the variance in the Rater × Stimulus cluster is related to idiosyncratic or private taste (and would allow inferences about differences in ranking preferences across participants); and the interpretation of the Rater cluster is controversial: while it seems to be related to individual differences (e.g., personality, mood), it is not clear whether it should count as a source of idiosyncratic contribution for judgment (Hönekopp, 2006). For example, if ratings of three stimuli by multiple raters lead to average ratings of 3, 4 and 5, and one particular rater gives out the ratings of 1, 2 and 3, respectively: this 2-point difference compared to the average ratings may be interpreted as meaningless differences in scale use, since they are in the same direction as the average ratings, thus indicating agreement with the average ratings and with the overall ranking of stimuli. Alternatively, this difference may reflect genuine differences in perception, in the sense that this rater disagrees with the average taste (thus indicating private taste). Hönekopp (2006) proposes two different versions of the beholder index: *bi1*, that disregards the Rater Cluster, and *bi*2, that takes the Rater Cluster into account. Please refer to Hönekopp (2006, p. 2) for formulae; to our accompanying R files for code to compute these indices; and to the Supplementary Information for an example of these analyses conducted on previous data from Bruder et al. (2023; Suppementary Figure S6). Note that we will explore if it makes sense to adapt the structure of clusters specified above to include a Singer cluster. We suspect that that would lead to too complex models (with too many terms and interactions, thus singular or not converging). We will test the viability of this by comparing model fit for models with and without the Singer cluster with standard model comparison techniques (likelihood ratio tests).

Note that since there is no straightforward way to summarize the precision of variance components (Bates, 2010), and *bi* is a ratio between variance components, comparisons between estimated *bi* across conditions must remain descriptive, that is, no confidence intervals or standard deviations can be reported. This is why we are focusing Hypothesis 1 in MM1 measures, though we hope to gain considerable insight into participants’ preferences based on *bi* estimates.

***2.4.3.2 Krippendorff’s alpha.*** To allow for direct comparison with other studies, we will also report Krippendorff’s alpha as an alternative measure of interrater agreement. Krippendorff’s alpha is a generalization of several known reliability indices, and widely applicable (Krippendorff, 2004, 2011). We will use the kripp.boot function from the kripp.boot R package (Proutskova & Gruszczynski, 2023), which implements Krippendorff’s algorithm (Krippendorff, 2011) for bootstrapping the α*K* coefficient and 95% confidence intervals. We will run 100 iterations and take the resulting mean value of all of bootstrapped replicates. Note that in our simulations, this value was very similar to the output of the kripp.alpha function from the irr R package (Gamer et al., 2019); and 10 or 100 iterations produced very similar results.

***2.4.3.2 Intraclass Correlations (ICC).*** A more widely-known interrater agreement measure, ICC, is also reported to allow for comparison with other studies. We will use the ICC function in the psych R package (Revelle, 2021) to compute ICC2 (single random raters, absolute values).

**Data/code/stimuli availability statement:**

Please find analysis code (R markdown files), a Supplementary Information file and stimuli at: https://osf.io/8k4af/?view\_only=506d243a6e7a4d3680c81e696ca81025

(Note the Supplementary Information is now merged at the end of the present file).

**References**

Albouy, P., Mehr, S. A., Hoyer, R. S., Ginzburg, J., & Zatorre, R. J. (2023). *Spectro-temporal acoustical markers differentiate speech from song across cultures* [Preprint]. Neuroscience. https://doi.org/10.1101/2023.01.29.526133

Babel, M., & McGuire, G. (2015). Perceptual fluency and judgments of vocal aesthetics and stereotypicality. *Cognitive Science*, *39*(4), 766–787. https://doi.org/10.1111/cogs.12179

Babel, M., McGuire, G., & King, J. (2014). Towards a more nuanced view of vocal attractiveness. *PLoS ONE*, *9*(2), e88616. https://doi.org/10.1371/journal.pone.0088616

Banse, R., & Scherer, K. R. (1996). Acoustic profiles in vocal emotion expression. *Journal of Personality and Social Psychology*, *70*(3), 614–636. https://doi.org/10.1037/0022-3514.70.3.614

Bates, D. (2010). *Mixed-effects modeling with R*. Springer. http://lme4.r-forge.rproject.org/book

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1). https://doi.org/10.18637/jss.v067.i01

Belin, P., Zatorre, R. J., Lafaille, P., Ahad, P., & Pike, B. (2000). Voice-selective areas in human auditory cortex. *Nature*, *403*(6767), 309–312. https://doi.org/10.1038/35002078

Brennan, R. L. (2001). *Generalizability Theory*. Springer New York. https://doi.org/10.1007/978-1-4757-3456-0

Brown, S. (2000). The ‘Musilanguage’ model of language evolution. In *The Origins of Music* (eds S. Brown, B. Merker, and N. L. Wallin, pp. 271–300). MIT Press.

Bruder, C., & Larrouy-Maestri, P. (2023). Classical singers are also proficient in non-classical singing. *Frontiers in Psychology*, *14*. https://doi.org/10.3389/fpsyg.2023.1215370

Bruder, C., Poeppel, D., & Larrouy-Maestri, P. (2023). *Perceptual (but not acoustic) features predict singing voice preferences* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/qvp8t

Brysbaert, M. (2019). How many participants do we have to include in properly powered experiments? A tutorial of power analysis with reference tables. *Journal of Cognition*, *2*(1), 16. https://doi.org/10.5334/joc.72

Caldwell, A. R. (2022). *Exploring Equivalence Testing with the Updated TOSTER R Package* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/ty8de

Champely, S. (2020). *pwr: Basic functions for power analysis (Version 1.3-0).* https://CRAN.R-project.org/package=pwr

Cox, C., Bergmann, C., Fowler, E., Keren-Portnoy, T., Roepstorff, A., Bryant, G., & Fusaroli, R. (2022). A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behaviour*, *7*(1), 114–133. https://doi.org/10.1038/s41562-022-01452-1

Dissanayake, E. (2000). Antecedents of the temporal arts in early mother–infant interaction. In *The origins of music* (N. L. Wallin, B. Merker, S. Brown (Eds.), pp. 389–410). The MIT Press.

Fernald, A. (1989). Intonation and communicative intent in mothers’ speech to infants: Is the melody the message? *Child Development*, *6*(60), 1497–1510.

George, M., & Ilavarasu, J. (2021). Development and psychometric validation of the Music Receptivity Scale. *Frontiers in Psychology*, *11*, 585891. https://doi.org/10.3389/fpsyg.2020.585891

Germine, L., Russell, R., Bronstad, P. M., Blokland, G. A. M., Smoller, J. W., Kwok, H., Anthony, S. E., Nakayama, K., Rhodes, G., & Wilmer, J. B. (2015). Individual aesthetic preferences for faces are shaped mostly by environments, not genes. *Current Biology*, *25*(20), 2684–2689. https://doi.org/10.1016/j.cub.2015.08.048

Goupil, L., Ponsot, E., Richardson, D., Reyes, G., & Aucouturier, J.-J. (2021). Listeners’ perceptions of the certainty and honesty of a speaker are associated with a common prosodic signature. *Nature Communications*, *12*(1), 861. https://doi.org/10.1038/s41467-020-20649-4

Hellbernd, N., & Sammler, D. (2016). Prosody conveys speaker’s intentions: Acoustic cues for speech act perception. *Journal of Memory and Language*, *88*, 70–86. https://doi.org/10.1016/j.jml.2016.01.001

Hellbernd, N., & Sammler, D. (2018). Neural bases of social communicative intentions in speech. *Social Cognitive and Affective Neuroscience*, *13*(6), 604–615. https://doi.org/10.1093/scan/nsy034

Hilton, C. B., Moser, C. J., Bertolo, M., Lee-Rubin, H., Amir, D., Bainbridge, C. M., Simson, J., Knox, D., Glowacki, L., Alemu, E., Galbarczyk, A., Jasienska, G., Ross, C. T., Neff, M. B., Martin, A., Cirelli, L. K., Trehub, S. E., Song, J., Kim, M., … Mehr, S. A. (2022). Acoustic regularities in infant-directed speech and song across cultures. *Nature Human Behaviour*, *6*(11), 1545–1556. https://doi.org/10.1038/s41562-022-01410-x

Hönekopp, J. (2006). Once more: Is beauty in the eye of the beholder? Relative contributions of private and shared taste to judgments of facial attractiveness. *Journal of Experimental Psychology: Human Perception and Performance*, *32*(2), 199–209. https://doi.org/10.1037/0096-1523.32.2.199

Kohl, M. (2023). *MKpower: Power analysis and sample size calculation.* R package version 0.7. https://github.com/stamats/MKpower

Kramer, R. S. S., Mileva, M., & Ritchie, K. L. (2018). Inter-rater agreement in trait judgements from faces. *PLOS ONE*, *13*(8), e0202655. https://doi.org/10.1371/journal.pone.0202655

Krippendorff, K. (2004). *Content analysis: An introduction to its methodology* (Second). Sage.

Krippendorff, K. (2011). *Computing Krippendorff’s Alpha-Reliability*. https://repository.upenn.edu/asc\_papers/43

Lakens, D. (2014). Performing high‐powered studies efficiently with sequential analyses. *European Journal of Social Psychology*, *44*(7), 701–710. https://doi.org/10.1002/ejsp.2023

Lakens, D. (2017). Equivalence Tests: A Practical Primer for *t* Tests, Correlations, and Meta-Analyses. *Social Psychological and Personality Science*, *8*(4), 355–362. https://doi.org/10.1177/1948550617697177

Larrouy-Maestri, P., Poeppel, D., & Pell, M. D. (2023). The sound of emotional prosody: Nearly three decades of research and future directions. *Perspective on Psychological Sciences.*, *(in press)*.

Leder, H., Goller, J., Rigotti, T., & Forster, M. (2016). Private and shared taste in art and face appreciation. *Frontiers in Human Neuroscience*, *10*. https://doi.org/10.3389/fnhum.2016.00155

Leongómez, J. D., Havlíček, J., & Roberts, S. C. (2022). Musicality in human vocal communication: An evolutionary perspective. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *377*(1841), 20200391. https://doi.org/10.1098/rstb.2020.0391

Lüdecke, D. (2019). *Esc: Effect size computation for meta analysis (Version 0.5.1)*. https://CRAN.R-project.org/package=esc. https://doi.org/10.5281/zenodo.1249218.

Martinez, J. E., Funk, F., & Todorov, A. (2020). Quantifying idiosyncratic and shared contributions to judgment. *Behavior Research Methods*, *52*(4), 1428–1444. https://doi.org/10.3758/s13428-019-01323-0

McAleer, P., Todorov, A., & Belin, P. (2014). How do you say ‘hello’? Personality impressions from brief novel voices. *PLoS ONE*, *9*(3), e90779. https://doi.org/10.1371/journal.pone.0090779

Mehr, S. A., & Krasnow, M. M. (2017). Parent-offspring conflict and the evolution of infant-directed song. *Evolution and Human Behavior*, *38*(5), 674–684. https://doi.org/10.1016/j.evolhumbehav.2016.12.005

Mehr, S. A., Singh, M., Knox, D., Ketter, D. M., Pickens-Jones, D., Atwood, S., Lucas, C., Jacoby, N., Egner, A. A., Hopkins, E. J., Howard, R. M., Hartshorne, J. K., Jennings, M. V., Simson, J., Bainbridge, C. M., Pinker, S., O’Donnell, T. J., Krasnow, M. M., & Glowacki, L. (2019). Universality and diversity in human song. *Science*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mook, A. T., & Mitchel, A. D. (2019). The role of audiovisual integration in the perception of attractiveness. *Evolutionary Behavioral Sciences*, *13*(1), 1–15. https://doi.org/10.1037/ebs0000128

Müllensiefen, D., Gingras, B., Musil, J., & Stewart, L. (2014). The musicality of non-musicians: An index for assessing musical sophistication in the general population. *PLoS ONE*, *9*(2), e89642. https://doi.org/10.1371/journal.pone.0089642

Norman-Haignere, S. V., Feather, J., Boebinger, D., Brunner, P., Ritaccio, A., McDermott, J. H., Schalk, G., & Kanwisher, N. (2022). A neural population selective for song in human auditory cortex. *Current Biology*, *32*(7), 1470-1484.e12. https://doi.org/10.1016/j.cub.2022.01.069

Ozaki, Y., Tierney, A., Pfordresher, P., Mcbride, J., Benetos, E., Proutskova, P., Chiba, G., Liu, F., Jacoby, N., Purdy, S., Opondo, P., Fitch, T., Hegde, S., Rocamora, M., Thorne, R., Nweke, F. E., Sadaphal, D., Sadaphal, P., Hadavi, S., … Savage, P. E. (2022). *Globally, songs and instrumental melodies are slower, higher, and use more stable pitches than speech [Stage 2 Registered Report]* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/jr9x7

Phillips, E. (2023). *A review of the speech-music continuum and its categorization: Evolution, form, and function* [Preprint]. PsyArXiv. https://doi.org/10.31234/osf.io/b2tkf

Pisanski, K., Fraccaro, P. J., Tigue, C. C., O’Connor, J. J. M., Röder, S., Andrews, P. W., Fink, B., DeBruine, L. M., Jones, B. C., & Feinberg, D. R. (2014). Vocal indicators of body size in men and women: A meta-analysis. *Animal Behaviour*, *95*, 89–99. https://doi.org/10.1016/j.anbehav.2014.06.011

Pisanski, K., Jones, B. C., Fink, B., O’Connor, J. J. M., DeBruine, L. M., Röder, S., & Feinberg, D. R. (2016). Voice parameters predict sex-specific body morphology in men and women. *Animal Behaviour*, *112*, 13–22. https://doi.org/10.1016/j.anbehav.2015.11.008

Proutskova, P., & Gruszczynski, M. (2023). *kripp.boot: Bootstrap Krippendorff’s alpha intercoder reliability statistic.* R package version 1.0.0.

R Core Team. (2021). *R: A language and environment for statistical computing* (Vienna, Austria). https://www.R-project.org/

Rock, A. M. L., Trainor, L. J., & Addison, T. L. (1999). Distinctive messages in infant-directed lullabies and play songs. *Developmental Psychology*, *35*(2), 527–534. https://doi.org/10.1037/ 0012-1649.35.2.527

RStudio Team. (2022). *RStudio: Integrated development environment for R*. Boston, MA. http://www.rstudio.com

Savage, P. E., Brown, S., Sakai, E., & Currie, T. E. (2015). Statistical universals reveal the structures and functions of human music. *Proceedings of the National Academy of Sciences*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Scherer, K. R. (1978). Personality inference from voice quality: The loud voice of extroversion. *European Journal of Social Psychology*, *8*(4), 467–487. https://doi.org/10.1002/ejsp.2420080405

Trehub, S. E., & Trainor, L. J. (1998). Singing to infants: Lullabies and playsongs. *Advances in Infancy Research*, *12*, 43–77.

Trehub, S. E., Unyk, A. M., & Trainor, L. J. (1993). Adults identify infant-directed music across cultures. *Infant Behavior and Development*, *16*(2), 193–211. https://doi.org/10.1016/0163-6383(93)80017-3

Valentova, J. V., Tureček, P., Varella, M. A. C., Šebesta, P., Mendes, F. D. C., Pereira, K. J., Kubicová, L., Stolařová, P., & Havlíček, J. (2019). Vocal parameters of speech and singing covary and are related to vocal attractiveness, body measures, and sociosexuality: A cross-cultural study. *Frontiers in Psychology*, *10*, 2029. https://doi.org/10.3389/fpsyg.2019.02029

van Rijn, P., & Larrouy-Maestri, P. (2023). Modelling individual and cross-cultural variation in the mapping of emotions to speech prosody. *Nature Human Behaviour*, *7*(3), 386–396. https://doi.org/10.1038/s41562-022-01505-5

Vessel, E. A., Maurer, N., Denker, A. H., & Starr, G. G. (2018). Stronger shared taste for natural aesthetic domains than for artifacts of human culture. *Cognition*, *179*, 121–131. https://doi.org/10.1016/j.cognition.2018.06.009

Vessel, E. A., & Rubin, N. (2010). Beauty and the beholder: Highly individual taste for abstract, but not real-world images. *Journal of Vision*, *10*(2), 1–14. https://doi.org/10.1167/10.2.18

Vessel, E. A., Stahl, J., Maurer, N., Denker, A., & Starr, G. G. (2014). *Personalized visual aesthetics* (B. E. Rogowitz, T. N. Pappas, & H. De Ridder, Eds.; p. 90140S). https://doi.org/10.1117/12.2043126

Weiss, M. W., Trehub, S. E., & Schellenberg, E. G. (2012). Something in the way she sings: Enhanced memory for vocal melodies. *Psychological Science*, *23*(10), 1074–1078. https://doi.org/10.1177/0956797612442552

Weiss, M. W., Vanzella, P., Schellenberg, E. G., & Trehub, S. E. (2015). Rapid Communication: Pianists exhibit enhanced memory for vocal melodies but not piano melodies. *Quarterly Journal of Experimental Psychology*, *68*(5), 866–877. https://doi.org/10.1080/17470218.2015.1020818

Yurdum, L., Singh, M., Glowacki, L., Vardy, T., Atkinson, Q. D., Hilton, C. B., Sauter, D., Krasnow, M. M., & Mehr, S. A. (2023). Universal interpretations of vocal music. *Proceedings of the National Academy of Sciences*, *120*(37), e2218593120. https://doi.org/10.1073/pnas.2218593120

Zäske, R., Skuk, V. G., & Schweinberger, S. R. (2020). Attractiveness and distinctiveness between speakers’ voices in naturalistic speech and their faces are uncorrelated. *Royal Society Open Science*, *7*(12), 201244. https://doi.org/10.1098/rsos.201244

**Supplementary information**

**Voice preferences across contrasting singing and speaking styles**

**Camila Bruder, Klaus Frieler & Pauline Larrouy-Maestri**

Please find additional information, including simulated data and analyses code as R markdown files in the project folder at osf:

https://osf.io/8k4af/?view\_only=506d243a6e7a4d3680c81e696ca81025



**Supplementary Figure S1**: Reanalysis of data from Bruder et al. (2023) to illustrate proposed analyses of interrater agreement with MM1 (“Mean-minus-one”). For Experiment 1, values are also shown separately for the whole sample (*N* = 326) and a subsample of 146 “consistent” participants (with rtest-retest ≥.5). Experiment 2 did not include repeated trials, so results are show only for the whole sample (*N* = 42). Error bars depict 95% confidence intervals.

**Interrater agreement**



**Supplementary Figure S2**: Illustration of proposed analysis for Question 1 with MM1 (“Mean-minus-one”) based on simulated datasets with increasing levels of interrater agreement or shared taste. For simplicity, simulations were made for 60 raters rating 22 items, and for only one vocalization style. In the low (null) agreement scenario, all ratings were random. In the intermediary scenario, the ratings given by 30 raters were nearly identical to each other (that is, the same sequence of integer ratings was given for all stimuli, with minimal variability introduced in the data generation process to avoid issues related to perfect correlation), and the ratings given by 30 raters were random. In the high agreement scenario, the ratings given by 50 raters were nearly identical to each other, and the ratings given by 10 raters were random. In the very high (perfect) agreement scenario, the ratings given by all 60 raters were nearly identical to each other. Error bars represent 95% confidence intervals based on the computed MM1 values.



**Supplementary Figure S3**: Illustration of consistency in preferences for certain singers across the five styles of vocalization in simulated data with increasing levels of consistency. Plotted dots are grand averages of liking ratings by singer for each style. Different colors depict the 22 singers and lines connect data for each singer across styles. The three simulated scenarios represent: varied preferences (**left)**; somewhat consistent preferences (**middle)**; and highy consistent preferences (**right)** across styles. AD: adult-directed; ID: infant-directed.

**Interstyle agreement**



**Supplementary Figure S4**: Illustration of proposed analysis for Question 2: “interstyle MM1” is used to measure interstyle agreement, that is, how consistent are average singer preferences across styles of vocalization. The plot shows computed interstyle MM1 values for the three simulated datasets displayed in Supplementary Figure S3, that is, for data with varied, somewhat consistent, and highly consistent average ratings by singer across the five styles. Error bars depict 95% confidence intervals. The dashed red line represents the stipulated threshold of .8 to consider preferences as “highly consistent”. Note we purposedly show one example of “somewhat consistent” interstyle agreement where the confidence interval crosses the threshold value, in which case preferences would not be considered “highly consistent”.

****

**Supplementary Figure S5**: Sheet music for the melody excerpts proposed as stimuli. The operatic style was performed with higher pitch than the pop and lullaby styles: for operatic singing, melodies were transposed one fifth (for “Nana Nenê” and “Chove Chuva”), or one fourth (for “Melodia Sensitmental”) higher.



**Supplementary Figure S6**: Reanalysis of data from Bruder et al. (2023) to illustrate proposed supporting analyses of variance component analysis (**top**) and beholder index (“bi”; **bottom**).