**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**

Human vocalizations are incredibly flexible and may sound different depending on their uses and functions – from interpersonal semantic communication, to expressing emotions and intentions, to cueing a speakers’ age and body size. Even though the voice works as such an important sociobiological signal, there are major gaps in our knowledge about how we process and perceive this auditory information. For instance: why do we like some speakers/singers more than others? Does it depend on the type of vocalization? Studies on the speaking voice suggest a clear link between vocal attractiveness and acoustic characteristics, but the relationship between acoustic features and singing voices preferences is not so straightforward, with recent findings suggesting that perceptual features (i.e., numerical sound descriptions by listeners) are more relevant than acoustic features (i.e., physical characteristics of the audio files).

We propose to investigate voice preferences with an integrative approach, encompassing contrasting types of vocalizations. To do that, we will use a newly recorded and validated stimulus set of contrasting vocalizations containing 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 out loud 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 as well as on scales for different perceptual features (e.g., breathiness, loudness, timbre). By measuring the consistency of these preferences across participants and over time (with two testing sessions), as well as modeling liking ratings based on perceptual features, we aim to characterize voice preferences in a wider framework – taking into account the variability both in vocalizations’ uses and functions and in participants’ aesthetic appreciation of them – to better understand a question central to human experience.

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

**1. Introduction**

The human voice serves several functions, and can sound very differently depending on its use. In the case of speaking, infant-directed vocalizations have special acoustic qualities (with high fundamental frequency, F0, high variability of F0, slow articulation rate and large vowel space area, see Cox et al., 2022; Hilton et al., 2022), and may be used to direct infants’ attention, express affect, communicate intent and facilitate language learning (Bryant & Barret, 2007; Fernald, 1989). Beyond supporting interpersonal communication and conveying semantic information, the speaking voice can also work as an important sociobiological signal, indicating emotional states (Banse & Scherer, 1996), personality traits (Goupil et al., 2021; McAleer et al., 2014; Scherer, 1978) and even cuing a speakers’ body size, health and age (Babel et al., 2014). When it comes to songs, there is high variability both between and (especially) within cultures in terms of voice quality, tempo, melodic and rhythmic complexity, pitch range and accent, but all studied cultures have some kind of singing (Mehr et al., 2019; Savage et al, 2015). Singing to infants is a widespread, cross-cultural practice, which may serve different functions, such as mood regulation and attention directing in the case of play songs and soothing an infant or encouraging sleep, in the case of lullabies (Mehr et al., 2018; Mehr et al., 2019; Rock at al., 1999; Trehub et al., 1993; Trehub & Schellenberg, 1995). Lullabies are characterized by simple, repetitive melodies, simple rhythm and preponderance of small melodic steps, allied to typical performance features like a unaccompanied, soft and quiet singing by a caregiver, often paired with movements such as rocking, swaying or patting (Mehr et al,. 2019; Trehub & Trainor, 1998; Unyk et al., 1992).

The studies on speaking and singing mentioned above glimpse at the diversity of human vocalizations from the point of view of their uses and functions. But why do we like some voices more than others? How does our enjoyment of voices vary across diverse types of vocalizations? Here we investigate the aesthetic appeal of a range of contrasting vocalizations – singing and speaking – in an integrative way.

In the case of spoken voices, the attractiveness of voices is believed to provide signals of the fitness of potential partners. Voice attractiveness has been shown to co-vary with sexually dimorphic traits. Individuals with more attractive voices also exhibit larger shoulder-to-hip ratios (for males) or smaller waist-to-hip ratios (for females) (Hughes et al., 2004). Accordingly, studies indicate that the acoustic bases of vocal attractiveness lie in the height of F0 and formant dispersion of a given voice, with higher F0 and more spread formants preferred for women‘s voices (Collins & Missing, 2003), and lower F0 preferred for men’s voices (Collins, 2000). Studies also report a general preference for voices with higher harmonic-to-noise ratios, which can be considered a measure of voice quality – it decreases with aging (Ferrand, 2002) or when being hoarse due to medical reasons (Yumoto et al., 1982). Familiarity and “averageness” were also linked to vocal attractiveness. Bruckert et al. (2010) used voice morphing software and observed that morphed, averaged voices (which are smoother and have higher harmonic-to-noise ratios) were considered more attractive than most of the individual voices presented to participants. Interestingly, Valentova et al. (2019) reported high correlations between attractiveness ratings of speaking and singing vocalizations, and suggested that they may work as “backup signals”, both shaped by sexual selection and cuing body size and fitness. However, further studies are needed to test this association with more varied singing material, and also scrutinizing acoustic, musical, and perceptual features (e.g., pitch accuracy, tempo, etc,) since they have been shown to influence the perception of singing and speaking abilities (e.g., Merril, 2022; Merril & Larrouy-Maestri, 2017).

To understand voice preferences, we take an interactionist approach (e.g., Wassiliwizky & Menninghaus, 2021), in the larger framework of empirical aesthetics – one that takes into account aspects of the stimulus as well as subjective, internal factors relating to the person making the aesthetic evaluation. This means that, in addition to examining mean liking ratings as indicative of average preferences, we will also examine the variability in these ratings across participants. One way of assessing the relative contributions to preferences of individual versus common factors is to measure agreement across participants (e.g., Vessel et al., 2010, 2014, 2018). We focus here on the variability (or consistency) of aesthetic judgements across vocalization styles, across participants and over time, as well as on the role of perceptual attributes of the voices. Note that we use the terms liking and aesthetic preferences in interchangeable ways.

A first step in this direction was taken in a previous study (Bruder et al., 2021a, 2021b/in preparation), in which participants were asked to rate pop singing performances in terms of perceptual attributes of the voices (i.e., articulation, breathiness, pitch accuracy, loudness, tempo, resonance, preciseness and softness of vocal onsets, amount of vibrato, timbre), as well as in terms of how much they liked them. Mixed linear models showed that liking ratings could be predicted by perceptual features (with about half of the variance of liking ratings explained by perceptual ratings), but not by acoustic features frequently used to describe voices such as jitter, shimmer, vibrato rate and extent, harmonics-to-noise ratio and tilt measures. In line with previous research on perceptual ratings of music (Lange & Frieler, 2018; Schedl et al., 2016) and voice (Merrill, 2022), we also observed that inter-rater agreement of perceptual and liking ratings was low. Interestingly though, mean liking ratings in this lab-based experiment with German participants correlated highly with mean liking ratings of a parallel online experiment presenting the same stimuli to US-based participants. This suggests the emergence of robust average preferences amidst high individual differences in how participants perceive and like singing voices. It is to be determined if this finding would generalize to other types of singing, that is, beyond pop singing.

Here we propose to expand the findings of Bruder et al. (2021a, 2021b/in preparation) and investigate lay listeners’ aesthetic preferences for voices in contrasting singing and speech styles. To do so, we use a newly recorded and validated stimulus set of naturalistic but controlled a capella singing and speech performances [the stimulus dataset, along with details about the validation experiment and acoustic analyses of the stimuli, will be, at the time of publication of this paper, available open access – currently it is work in progress]. Twenty-two female singers performed six different melody excerpts in three contrasting singing styles – as a lullaby, as a pop song and as opera aria; and read the corresponding lyrics out loud in two contrasting ways – as if speaking to an adult audience and as if speaking to an infant. Note that the term pop singing is used here in a broad and unspecific way. Though there are different schools and techniques associated with different aesthetic goals within pop music, pop singing is defined here as singing without any specific type of technique. Operatic or classical singing, on the other hand, is the result of a very specific technique, and is associated with a very clear acoustic profile. Larrouy-Maestri et al. (2014) compared acoustic features of operatic and non-operatic singing by asking the same singers to perform two melodies (“Happy Birthday” and a romantic song of free choice) both with and without operatic technique, and found wider vibrato extent, higher standard deviation of the F0, increased jitter and shimmer, as well as lower harmonics-to-noise ratio and lower energy ratio distribution in operatic singing.

These five contrasting “categories” of vocalizations allow us to draw an interesting comparison with findings from the visual domain. Using a correlational measure of agreement (“Mean Minus One”, MM1), 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 an even smaller for artworks (which reflected strong individual differences or idiosyncratic taste). They argue that the behavioral relevance of naturally occurring types of stimuli such as landscapes and human faces results in information processing, and hence aesthetic experience, that is highly conserved across individuals. 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. Applying this rationale to the auditory domain and our voice stimuli and quantifying the amount of shared taste for these five types of vocalizations should help us characterize voice preferences in an integrative way, focusing not only on average preferences, but also acknowledging individual differences in these preferences.

* 1. **Study aims and hypotheses**

This study aims to empirically investigate individuals’ voice preferences and consistency (between and within participants), as well as the perceptual grounds of these preferences across a varied but controlled stimulus set of contrasting vocalizations. Given the scarcity of previous empirical research on this subject, a part of this study is exploratory. We will nonetheless make (somewhat speculative) predictions based on available theories from the fields of empirical aesthetics and music cognition (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** |
| 1A) How much do people agree in terms of which voices they prefer across singing styles? | H0: no difference in agreement between lullaby, pop and opera  H1: agreement for lullaby higher than pop; and for pop higher than opera | 45 participants rating each stimulus in terms of liking | Comparison of  MM1 measures  for lullaby, opera  and pop: one-way  rANOVA/Kruskal-Wallis  (see 1.2.1) | Significance of difference in mean MM1 scores across 3 styles according to ANOVA/Kruskal-Wallis | The null hypothesis of no difference in MM1 across singing styles is rejected if *p* < .05 |
| 1B) Is inter-rater agreement (MM1) consistent over time (two testing sessions)? | H0: similar amount of agreement in sessions 1 and 2  H1: different amount of agreement in sessions 1 and 2 | 45 participants rating each stimulus in terms of liking | Comparison of MM1 values (pooled across styles) in sessions 1 and 2: paired t-/Wilcoxon test (see 1.2.1) | Significance of difference in MM1 scores between session 1 and 2 according to paired t-/Wilcoxon test | The null hypothesis of no difference in MM1 between sessions is rejected if  *p* < .05 |
| 2A) On average, will the same performers be preferred across styles? | Ho: no difference in rankings of performers across styles (“better” voices consistently ranked higher)  H1: rankings differ across styles | Average liking ratings by 45 participants | Based on mean liking ratings by performer, comparison of ranking of performers for each style: Friedman test comparing 5 rankings (see 1.2.2) | Significance of differences in rankings between styles as measured by Friedman’s test | The null hypothesis of no difference in rankings across styles is rejected if *p <* .05 (suggesting preferred voices vary across styles) |
| 2B) How consistent are (average) preferences for performers over time (two testing sessions)? | H0: similar rankings of performers in sessions 1 and 2  H1: different rankings of performers in sessions 1 and 2 | Average liking ratings by 45 participants | Based on mean liking ratings by performer (pooled across styles),  Spearman correlation score between sessions 1 and 2 (see 1.2.2) | A correlation higher than .75 indicates consistency of rankings of performers across testing sessions | The null hypothesis of no difference in rankings of performers between sessions is rejected if  *r test, retest <* .75 |

***1.1.1 Hypothesis regarding (in)consistency of preferences across styles (Question 1A) and over time (Question 1B)***

Individuals’ agreement in terms of their voice preferences may vary depending on the type of vocalization, as observed in the visual domain. Expanding on Vessel and colleagues’ (2010, 2014, 2018) findings, we expect a higher degree of shared taste (high inter-rater agreement) for aesthetic ratings of lullabies, a more “natural”, evolutionarily important (e.g., Mehr & Krasnow, 2020) kind of singing, than pop and operatic singing; and higher shared taste for pop than operatic singing (the latter representing a more technical and specific type of stimulus). We refrain from hypothesizing about agreement for the speaking performances. We will also assess how consistent are these agreement measures across two testing sessions.

***1.1.2 Hypothesis regarding average preferences for some singers (Question 2A) and over time (Question 2B)***

If some voices are “fundamentally” more likeable, this should happen consistently across styles, that is, participants’ rankings of favorite singers/speakers should not vary across styles. On the other hand, differences in rankings across styles would suggest that some performers and voice qualities were more “adequate” or “conformant” to some styles than to others. Again, the consistency on these rankings over time (two testing sessions) will inform us on the strength of the observed patterns.

***1.1.3 Hypothesis regarding the prediction of "liking" from perceptual features***

In the case of pop singing, recent findings (Bruder et al., 2021a, 2021b/in preparation) showed that about half of the variance in participants’ liking of singing performances could be predicted based on perceptual (but not acoustic) features of the voice. We expect to replicate and extend such observation across different vocalizations. Note that this analysis is not included in Table 1 because it is exploratory (i.e., no clear hypotheses other than stating that we expect to be able to predict liking ratings from perceptual features, and to find interactions of style and perceptual features). In any case, these models should provide a more integrative way to describe liking of vocalizations, taking into account dependencies coming from repeated measures in subjects and stimuli items.

[Please see accompanying scripts with R code for simulating a dataset and running all proposed analyses]. Whenever requirements are met, we plan to use parametric methods, aiming at higher power (otherwise adjusting to non-parametric alternatives; these changes are specified in our proposed analyses code).

* + 1. ***Consistency of preferences across styles and over time***

For question 1A, for each stimulus item, we will average ratings from sessions 1 and 2 to compare MM1 measures between the lullaby, pop and operatic singing styles with a one-way repeated measures ANOVA, followed by Tukey post-hoc test in case of significant difference between styles (or Kruskal-Wallis and Dunn post-hoc tests as nonparametric alternatives). Our directional hypothesis specifies that MM1 values should be higher for lullabies than pop performances, and higher for pop than operatic performances. For question 1B, we will compare participants’ MM1 values (pooled across styles) for sessions 1 and 2 with a paired t- (or Wilcoxon) test. We will also compute a Pearson correlation between MM1 values between sessions and stipulate that a high correlation (*r* > .75) indicates that MM1 values did not vary much between sessions, that is, that the amount of agreement remained consistent over time.

Power analysis with G\*Power (version 3.1) indicated that a sample of 45 participants would be enough to detect a small to moderate effect size of *f* = .2 with power = .8 and alpha set to 0.05 in a repeated-measures ANOVA with one group and 3 measurements (using default settings of correlation among repeated measures = .5, nonsphericity correction *ε* = 1, and effect size specification as in G\*Power 3.0). Accordingly, for a paired t-test (two-tailed), this sample size of 45 participants would be enough to detect a small to moderate effect size of *dz* = 0.43 with power = .8 and alpha set to 0.05. In case we need to run nonparametric tests instead, we will have less power, but should still be able to detect moderate effect sizes (considering the recommendation of using a sample size 15% bigger for nonparametric tests than one would use for a parametric test - Lehmann, 1998).

* + 1. ***Average preferences for some singers***

For question 2A, based on mean liking ratings per stimulus item across all participants and pooled (averaged) across sessions, we will compare the rankings of singers across the five styles of vocalization with a Friedman test *(*and *posthoc* tests, controlling for the FWER with the Holm method in case there is a difference). Note that in this case rejecting the null hypothesis of no difference in rankings across styles indicates that there is a difference in terms of which voices are preferred across styles. Since power analysis for nonparametrical tests is not so straightforward, we followed a recommendation of using a sample size 15% bigger than one would use for a parametric test (Lehmann, 1998). Power analysis indicated that for a repeated-measures ANOVA with one group with 5 measurements, a sample size of 32 participants would be enough to detect a small to moderate effect size of *f =*.2 with power = 0.8 and alpha = 0.05 (using default settings of correlation among repeated measures = .5, nonsphericity correction *ε* = 1, and effect size specification as in G\*Power 3.0). Multiplying this estimated necessary sample size by 1.15 to estimate power for a nonparametrical test leads to a required sample size of 36 participants for our proposed Friedmans’ test (fewer than our specified 45 participants).

For question 2B, based on mean liking ratings by performer (pooled over styles), we will compute a Spearman correlation between values for sessions 1 and 2. We will consider a high correlation (*ρ* > .75) an indication that preferences for certain performers did not vary much between sessions, that is, that these preferences remained consistent over time.

***1.2.3 Predicting liking from perceptual features***

Linear mixed models will be proposed to explore the level of prediction of liking ratings achieved based on perceptual features. We plan to fit one model for data from trials with the two styles of speaking and one model for data from trials with the three styles of singing (since they are partially based on different rating scales). Preliminary model specification is as follows: liking ratings predicted from perceptual ratings as fixed effects, including the interaction of style (lullaby, opera or pop for the singing model; adult-directed or infant-directed for the speaking model) with each predictor, and participants and stimuli items as random effects. [We will try to optimize this base model using standard model comparison techniques. As these are dependent on actual data, we did not include them in the analyses script yet].

**2. Method**

**2.1 Participants**

Participants will be recruited from the participant database of the Max Planck Institute for Empirical Aesthetics’s, in Frankfurt, Germany, which consists mostly of lay listeners, with a preponderance of students and retired subjects. 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), by examining participants with a large range of musical expertise, we hope to enhance representativity of the general population (compared to the alternative of recruiting only musically trained participants). Note that studies indicate that lay listeners are able to evaluate spoken (Bänziger et al, 2014) and singing voices (Merrill, 2022) if suitable scales are made available to them.

Participants will be rewarded for their participation at a 7€ per half hour rate. The only exclusion criterion for participation in data collection will be reported hearing impairments (announced in the invitation participants receive, prior to participation). Data from participants will be excluded from analyses if their responses pattern shows they were clearly not performing the task attentively (for instance, if they give the same rating through a whole block of trials).

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.

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

In addition to providing biographical data, participants will be asked to complete three questionnaires to be used in exploratory analyses.

1) the 18-items version of the general Music Sophistication subscale from the Goldsmiths Music Sophistication Index (Müllensiefen et al., 2014), as computed with the Gold-MSI configurator (https://shiny.gold-msi.org/gmsiconfigurator).

2) the Ten-Item Personality Inventory (TIPI), which is a short self-report measure of the Big-Five personality domains (Gosling et al., 2003), in the German version (Muck et al., 2007). Each of the five personality dimensions – Extraversion, Agreeableness, Conscientiousness, Emotional stability (or Neuroticism) and Openness to new experiences – is measured by two items, selected from the high and low poles of each domain. Each question presents two central descriptors, andparticipants to rate on a scale from 1 (disagreestrongly) to 7 (agree strongly) how much those two traits apply to them.

3) the reviewed Short Test of Music Preference (STOMP-R), a short self-report inventory for musical preferences (Rentfrow et al., 2011; see Fricke & Herzberg, 2017 for a German validation).

**2.2 Materials**

***2.2.1 Stimulus set***

The complete dataset of vocalizations consists of six melody excerpts (the first phrase of different songs) performed by 22 highly trained 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 studio conditions and performed each melody excerpt as a lullaby, as a pop song, as an opera aria (performed one fifth higher as pop and lullaby stimuli), and spoke the corresponding lyrics as if directed to an adult audience and as if directed to an infant. The exact instructions given to singers during the recording session were: 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. All productions were made both with the original lyrics and with a /lu/ sound instead of the original lyrics. In all cases, performances with a /lu/ sound were recorded directly after the performance with lyrics; specifically for speech performances, the resulting /lu/ performance followed the same rhythmic and prosodic contour of the performance with lyrics. Singing stimuli are on average 9 seconds long, and speech stimuli are on average 5 seconds long.

**Loudness normalization**. Stimuli were loudness normalized using the To Audio Converter (Version 1.0.16 – 1059) software. All speaking and pop singing stimuli were loudness normalized to -18 LUFS; all lullaby stimuli to -25 LUFS; and all opera stimuli to -14 LUFS. This was done to ensure stimuli within each style had the same perceptual level of intensity, while still keeping some variability within their general stylistic characteristics (that is, from the softness of lullabies to the higher intensities resulting from the use of operatic technique).

**Validation of the stimulus material.** The stimulus set was validated in one lab experiment where participants (lay listeners, total N = 75, divided into three groups; each stimulus was judged by 25 participants) were asked to indicate in each trial if a given singing performance sounded like a lullaby, a pop song, or an opera aria, or (with a different group of participants) if a given speaking performance was directed to an adult or to a baby/child. The overall average accuracy achieved was .79 for singing performances and .80 for speech performances. For the subset of /lu/ performances, the overall average accuracy reached .84 for opera, .79 for lullabies and .65 for pop performances; and .8 for adult-directed and .75 for infant-directed performances. Further details about the analyses plan for the validation experiment can be found in its preregistration: <https://osf.io/wuvb8>. The final validated dataset, along with details of the validation experiment and the description of extracted acoustic features of all stimuli, will be made available via open access once all analyses are done. Examples of the stimuli used in the present work are currently available at https://owncloud.gwdg.de/index.php/s/6IWIvTc828vB77R.

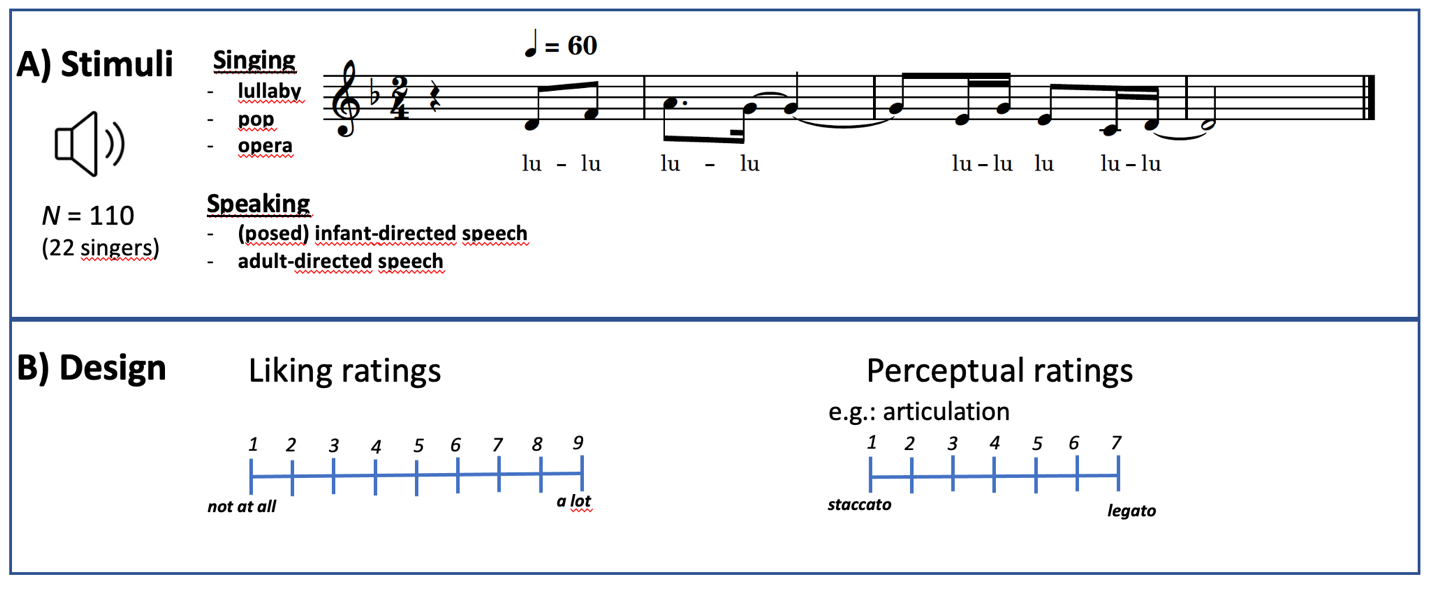
For the experiments proposed in the present work, we will only use one of the melody excerpts, the first phrase from “Chove Chuva” (by Brazilian artist Jorge Ben Jor), and only performances with a /lu/ sound. This leads to 110 performances (by 22 singers, each performing three styles of singing and two styles of speaking).

### **2.2.2 Acoustic analyses**

Each individual singing performance was segmented to individual notes using Tony (Mauch & Dixon, 2014; Mauch et al, 2015). After eventual note corrections (made manually upon visual inspection of individual files), the note data were exported as text files containing information about F0 and duration of each individual note. Each melody excerpt was then cut into individual chunks (one for each sung note) using a sox bash script. Individual notes (as wav files) were then entered into acoustic analysis. Using Praat (Boersma, 2001,Version 6.0.46) and default settings, except for: pitch\_floor = 75; pitch\_ceiling = 800, we extracted the following measures: F0, F0\_max and F0\_min and standard deviation of the F0; shimmer (perturbation in the amplitude of F0); jitter (perturbation in the periodicity of F0). To calculate pitch accuracy, we first converted F0 values from Herz to cents (100 cents corresponds to 1 semitone; the reference lowest note used was 261,626 Hz), then calculated the absolute difference between these values and reference (“correct”, according to sheet music) notes, also in cents; then averaged the pitch (in)accuracy per take. Using VoiceSauce (Shue, 2010; Shue et al., 2011) and with the same Praat settings of pitch\_floor = 75 pitch\_ceiling = 800), we extracted the following measures related to the spectral composition of the audio signal: harmonics-to-noise ratio in the 0–3.5 kHz band (HNR35), a ratio between periodic and nonperiodic components; cepstral peak prominence (CPP), a voice quality measure; energy, the amplitude of the sound wave. Note that even though acoustic features were extracted for individual notes, for subsequent analyses (statistical models using acoustic features as predictors) we calculated averaged values per performance.

**2.3 Procedure**

Following the procedure proposed by Bruder et al. (2021/in preparation) and designed to order to explore the role of perceptual features in listeners’ preferences, we will ask participants to rate singing and speech stimuli in terms of perceptual attributes shown to be relevant in the appreciation of singing and speech/speechsong (Merril, 2018, 2022; Merril & Larrouy-Maestri, 2017). These ratings will be made on bipolar scales ranging from 1 to 7 and displaying contrasting anchor words on each pole. Six scales will be used both for singing and speech stimuli (Table 1): attack (soft – hard), breathiness (not at all – very breathy), loudness (quiet – loud), resonance (thin – full), tempo (slow – fast), timbre (light – dark). The following four features will be collected only for singing stimuli: (perceived) pitch accuracy (out of tune – in tune), voice onset (imprecise – precise), articulation (staccato – legato), and amount of vibrato (not at all – a lot), whereas the following three features will be used only for speech performances: overall pitch (low – high), pitch range (narrow – wide), loudness range (narrow – wide). This means for blocks with singing stimuli there will be 10 perceptual scales to rate, and in blocks with speech stimuli there will be 9 perceptual scales to rate. Additionally, participants will be asked to rate how much they liked each stimulus on a scale of 1 (not at all) to 9 (a lot). Half-way through the experiment, participants will be asked to complete the three questionnaires mentioned above. Participants will complete two testing sessions (test-retest), not longer than 10 days apart from each other. The second session will be identical to the first one, with the exception that no questionnaires will have to be filled in the second session. The whole experiment will be conducted in German.



**Figure 1A)** Illustration of the stimulus material, consisting of performances of a short melody excerpt, performed by 22 singers a capella, with a /lu/ sound, in three different singing styles (as a lullaby, as a pop song, as an opera aria) and in two speaking styles (as if directed to an adult audience, as if directed to an infant). Note that operatic singing was performed one fifth higher than pop and lullaby (in A minor instead of D minor). **B)** Illustration of the experimental design: participants will be asked to rate each stimulus in terms of how much they liked it on a 9-point scale (not at all – a lot), as well as on different 7-point bipolar perceptual scales (see Table 2 for individual description of all scales).

**Table 2: Definition of perceptual ratings to be collected on bipolar scales, along with anchor words (inside parenthesis)**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Perceptual attribute** | | **Definition** | | |
| **All stimuli** | |  | | |
| *Attack* | | The way in which a note or syllable begins (soft – hard) | | |
| *Breathiness* | | The amount of air flow in the voice: how breathy does the voice sound? (not at all – a lot) | | |
| *Loudness* | | The magnitude of the auditory sensation (quiet – loud) | | |
| *Resonance* | | The fullness or reverberation of a voice (thin – full) | | |
| *Tempo* | | The speed or pace of the performance (slow – fast) | | |
| *Timbre* | | The perceived sound quality of the voice (dark – bright) | | |
| **Singing stimuli only**  *Pitch accuracy*  *Articulation*  *Vibrato*  **Speaking stimuli only**  *Overall pitch*  *Pitch range*  *Loudness range* | | | | How precise is each note along the melody (in-tune – out-of-tune)  How notes are connected to each other (staccato – legato)  A slight and periodic oscillation of the pitch of a sustained note: how much vibrato does the performer use? (none at all – a lot)  The mean pitch of the performance (low – high)  How much the pitch varies during the performance (narrow – wide range)  How much the loudness varies during the performance (narrow – wide range) |

The experiment will be divided in five blocks, one for each style of singing/speaking. Half

of the participants will start with singing stimuli, half with speech stimuli. Each block will comprise 22 trials, corresponding to one take by each of the 22 singers, presented in a randomized order. The order of these three or two blocks within each type of stimuli (singing or speech) will be counterbalanced across participants. Participants will complete the experiment at their own pace and are expected to take up to two hours to complete the experiment. Breaks will be proposed between blocks.

The experimental session will run as follows: after receiving general instructions, the definitions of the rating scales will be presented on the computer screen, along with three examples of singing performances or two examples of speech performances (one for each style), and participants will be asked to familiarize themselves with the rating scales. Then the actual experiment will begin. For each experimental stimulus, participants will be able to click on the “play” button as many times as they want to and listen to the stimulus again as they rate the bipolar scales and the scale for liking ratings. Additionally, there will be two open-end optional fields to fill: “This voice is especially… (write an adjective)” and “This performance is especially… (write an adjective)”. These were added to explore if similar “labels” might be given to the same voices by different participants, and to allow participants to express how they perceive the stimuli in a freer way. When all scales are completed, participants will be able to press the “next” button to proceed to the next page, where the next stimulus will be presented.

Stimuli will be presented and data will be recorded in the experimental platform Labvanced (Finger et al., 2017). Participants will be tested in the laboratories of the Max Planck Institute for Empirical Aesthetics, in Frankfurt, Germany.

**2.4 Data analyses**

**2.4.1 *Inter-rater agreement.*** Agreement will be assessed using a “mean-minus-one” (MM1) correlation measure as described by Vessel et al. (2014): a correlation is computed between a given participant’s liking ratings for each individual stimulus and the average ratings of all other participants. The across-observer average MM1 score is computed by 1) transforming individual r-values to z values, 2) computing a mean, and 3) transforming that score back to an r-value for easier interpretability. This method has been shown to result in less biased estimates than averaging raw correlations (Corey et al., 1998). Additionally, to allow for direct comparisons with other studies using perceptual ratings of voices, we will also report Krippendorff’s alpha (α*K*), as recommended by Lange and Frieler (2018) for perceptual ratings of music stimuli, using the kripp.alpha function in the irr R package (Gamer et al., 2019) and intraclass correlations (ICC2 or single random raters, absolute values), using the ICC function in the psych R package (Revelle, 2021).

**2.4.2 *Intra-rater agreement.*** In addition to the consistency measures mentioned in Section 1.2, we will also report, for each participant and based on his/her ratings of 110 stimuli in the first and second sessions, Pearson correlation scores as a measure of test-retest intra-rater agreement (separately for liking rating and each of the 10 perceptual feature ratings).

**2.4.3 *Modelling.*** Different linear mixed effects analyses will be proposed using the lmer function from the lme4 package (Bates et al., 2015) implemented in R (R Core Team, 2014). For all models reported, residual plots and QQ-curves will be visually inspected to make sure there was no deviation from normality or homoscedasticity. Estimates for degrees of freedom, F statistics, and p-values will be computed using Satterthwaite approximation with anova() function in the lmerTest package (Kuznetsova et al., 2015).Variance inflation factor (VIF) will be monitored with the goal of keeping it under 4 for all predictors.

**2.5 *Exploratory analyses*.** The non-exhaustive list of exploratory analyses includes investigating if participants’ characteristics collected with questionnaires (Section 2.1.1) can predict liking ratings and/or interact with ratings in the different perceptual scales. Further, we will test if findings from the field of voice attractiveness - higher voice attractiveness for higher mean F0 and more dispersed formants for female voices (Babel et al, 2014; Valentova et al, 2019) - replicate in our study: we will test models predicting liking ratings from mean F0, formant dispersion and indirect estimates of vocal tract length; as well as other acoustic measures such as jitter, shimmer, harmonics-to-noise ratio, cepstral peak prominence (CPP) and tilt measures H1H2, H1A1, H1A2, H1A3.

**References**

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, Mächler M, Bolker B, Walker S (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software*, 67(1), 1–48. doi:10.18637/jss.v067.i01.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International* 5(9/10), 341–345. https://hdl.handle.net/11245/1.200596

Bruckert, L., Bestelmeyer, P., Latinus, M., Rouger, J., Charest, I., Rousselet, G. A., Kawahara, H., & Belin, P. (2010). Vocal attractiveness increases by averaging. *Current biology,* *20*(2), 116–120. https://doi.org/10.1016/j.cub.2009.11.034

Bruder, C., Jacoby, N., Poeppel, D., & Larrouy-Maestri, P. (2021a, July). Predicting aesthetic ratings from the acoustics of sung melodies [Paper presentation]. 16th International Conference on Music Perception and Cognition and 11th Triennial Conference of the European Society for the Cognitive Sciences of Music (ICMPC-ESCOM2021), online.

Bruder, C., Jacoby, N., Poeppel, D., & Larrouy-Maestri, P. (2021b, November). What makes a singer your favorite one? [Paper presentation]. International Conference of Students of Systematic Musicology (SysMus21), online/Aarhus, Denmark.

Bryant, G. A., & Barrett, H. C. (2007). Recognizing intentions in infant-directed speech: evidence for universals. *Psychological science*, *18*(8), 746–751. https://doi.org/10.1111/j.1467-9280.2007.01970.x

Collins S. A. (2000). Men's voices and women's choices. *Animal behaviour*, *60*(6), 773–780. https://doi.org/10.1006/anbe.2000.1523

Collins, S. A., & Missing, C. (2003). Vocal and visual attractiveness are related in women. *Animal Behaviour*, *65(5)*, 997–1004. https://doi.org/10.1006/anbe.2003.2123

Corey, D. M., Dunlap, W. P., & Burke, M. J., 1998. Averaging Correlations: Expected Values and Bias in Combined Pearson rs and Fisher's z Transformations. *The Journal of General Psychology*, 125:3, 245-261. https://doi.org/10.1080/00221309809595548

Cox, C., Bergmann, C., Fowler, E. *et al.* A systematic review and Bayesian meta-analysis of the acoustic features of infant-directed speech. *Nature Human Behav*ior(2022). https://doi.org/10.1038/s41562-022-01452-1

Demetriou, A., Jansson, A., Kumar, A., & Bittner, R.M. (2018, September 23-27). *Vocals in music matter: the relevance of vocals in the minds of listeners*. 19th International Society for Music Information Retrieval Conference, ISMIR 2018, Paris, France. https://archives.ismir.net/ismir2018/paper/000098.pdf

Erich L. Lehmann, Nonparametrics : Statistical Methods Based on Ranks, Revised, 1998, ISBN=978-0139977350.

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

Ferrand C. T. (2002). Harmonics-to-noise ratio: an index of vocal aging. *Journal of voice*, *16*(4), 480–487. https://doi.org/10.1016/s0892-1997(02)00123-6

Finger, H., Goeke, C., Diekamp, D., Standvoß, K., & König, P. (2017). LabVanced: a unified JavaScript framework for online studies. In International Conference on Computational Social Science (Cologne).

Fricke, K. R., & Herzberg, P. Y. (2017). Personality and self-reported preference for music genres and attributes ina German-speaking sample. Journal of Research in Personality, 68, 114–123. https://doi.org/10.1016/j.jrp.2017.01.001

Gamer, M., Lemon, J., Fellows, I., & Singh, P. (2019). irr: Various Coefficients of Interrater Reliability and Agreement. R package version 0.84.1. https://CRAN.R-project.org/package=irr

Gosling, S. D., Rentfrow, P. J., & Swann, W. B., Jr. (2003). A Very Brief Measure of the Big Five Personality Domains. Journal of Research in Personality, 37, 504-528. https://doi.org/10.1016/S0092-6566(03)00046-1

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

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *The Behavioral and brain sciences*, *33*(2-3), 61–135. https://doi.org/10.1017/S0140525X0999152X

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*. https://doi.org/10.1038/s41562-022-01410-x

Hughes, S. M., Dispenza, F., & Gallup, G. G., Jr. (2004). Ratings of voice attractiveness predict sexual behavior and body configuration. Evolution and Human Behavior, 25(5), 295–304. https://doi.org/10.1016/j.evolhumbehav.2004.06.001

Kuznetsova, A., Brockhoff, P. B. ,& Christensen, R. H. B. (2017). "lmerTest Package: Tests in Linear Mixed Effects Models." *Journal of Statistical Software*, 82(13), 1–26.

Larrouy-Maestri, P., Magis, D., & Morsomme, D. (2014). Effects of Melody and Technique on Acoustical and Musical Features of Western Operatic Singing Voices. *Journal of Voice*, *28*(3), 332–340. https://doi.org/10.1016/j.jvoice.2013.10.019

Lange, E. B., & Frieler, K. (2018). Challenges and Opportunities of Predicting Musical Emotions with Perceptual and Automatized Features. *Music Perception*, *36*(2), 217–242. https://doi.org/10.1525/mp.2018.36.2.217

Lehmann, E.L. (1998). Nonparametrics: Statistical Methods Based on Ranks, Revised. Springer.

Mauch, M., & Dixon, S. (2014). PYIN: A fundamental frequency estimator using probabilistic threshold distributions. *2014 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. https://doi.org/10.1109/icassp.2014.6853678

Mauch, M., Cannam, C., Bittner, R., Fazekas, G., Salamon, J., Dai, J., Bello, J., & Dixon, S. (2015). Computer-aided Melody Note Transcription Using the Tony Software: Accuracy and Efficiency. *Proceedings of the First International Conference on Technologies for Music Notation and Representation*, https://code.soundsoftware.ac.uk/attachments/download/1423/tony-paper\_preprint.pdf

McAleer, P., Todorov, A. & Berlin, P. (2014). How do you say ’Hello’? Personality impressions from brief novel voices. *PLOS ONE 9*(3).

Merrill, J., & Larrouy-Maestri, P. (2017). Vocal Features of Song and Speech: Insights from Schoenberg's Pierrot Lunaire. *Frontiers in psychology*, *8*, 1108. https://doi.org/10.3389/fpsyg.2017.01108

Merrill, J. Auditory perceptual assessment of voices: Examining perceptual ratings as a function of voice experience. *Curr Psychol* (2022). https://doi.org/10.1007/s12144-022-02734-7

Mehr, S. A., Singh, M., York, H., Glowacki, L., & Krasnow, M. M. (2018). Form and Function in Human Song. *Current biology : Current Biology*, *28*(3), 356–368. https://doi.org/10.1016/j.cub.2017.12.042

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 (New York, N.Y.)*, *366*(6468), eaax0868. https://doi.org/10.1126/science.aax0868

Mehr, S. A., Krasnow, M. M., Bryant, G. A., & Hagen, E. H. (2020). Origins of music in credible signaling. *The Behavioral and brain sciences*, *44*, e60. https://doi.org/10.1017/S0140525X20000345

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

Muck, P. M., Hell, B., & Gosling, S. D. (2007). [Construct validation of a short Five-Factor Model instrument: A self-peer study on the German adaptation of the Ten-Item Personality Inventory (TIPI-G)](http://gosling.psy.utexas.edu/wp-content/uploads/2014/10/muck-et-al-07-German-TIPI.pdf). European Journal of Personality Assessment.

R Core Team (2021). R: A language and environment for statistical computing. R Foundation for

Statistical Computing, Vienna, Austria. URL https://www.R-project.org/.

Rentfrow, P. J., Goldberg, L. R., & Levitin, D. J. (2011). The structure of musical preferences: A five-factor model. *Journal of Personality and Social Psychology*, 100, 1139–1157. https://doi.org/10.1037/a0022406

Revelle, W. (2021). psych: Procedures for Personality and Psychological Research, Northwestern University, Evanston, Illinois, USA, https://CRAN.R-project.org/package=psych Version = 2.1.6.

Rock, A. M., 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

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 of the United States of America*, *112*(29), 8987–8992. https://doi.org/10.1073/pnas.1414495112

Schedl, M., Eghbal-zadeh, H., Gómez, E. & Tkalvcivc, M. (2016). An analysis of agreement in classical music perception and its relationship to listener characteristics. Proceedings of the 17th ISMIR Conference, New York City, USA,

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

Shue, Y.-L.(2010). *The voice source in speech production: Data, analysis and models* [Doctoral dissertation, UCLA]. http://www.seas.ucla.edu/spapl/paper/shue\_dissertation.pdf

Shue, Y.-L., P. Keating , C. Vicenik, K. Yu (2011) VoiceSauce: A program for voice analysis. *Proceedings of the ICPhS XVII*, 1846-1849. https://linguistics.ucla.edu/people/keating/Shue-etal\_ICPhS\_2011.pdf

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

Trehub, S. E. & Schellenberg, E. G. (1995). Music: It’s relevance to infants. *Annals of child development*, 11, 1-24.

Trehub, S. E. & Trainor, L. (1998). Singing to infants: lullabies and play songs. *Advances in Infancy Research*, 12, 43–78.

Trehub, S. E., Unyk, A. M. & Trainor, L. J. (1993). Adults identify infant-directed

music across cultures. *Infant Behavior and Development* ,16, 193–211. https://doi.org/10.1016/0163-6383(93)80017-3

Trehub, S.E., Becker, J, Morley, I. (2015). Cross-cultural perspectives on music and musicality. *Philosophical Transactions of the Royal Society B, 370:*20140096.

Unyk, A. M., Trehub, S. E., Trainor, L. J., & Schellenberg, E. G. (1992). Lullabies and simplicity: A cross-cultural perspective. Psychology of Music, 20(1), 15–28. https://doi.org/10.1177/0305735692201002

Valentova, J. V et al. (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

Vessel, E. A. (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. *SPIE Proceedings*. https://doi.org/10.1117/12.2043126

**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. http://doi.org/10.1016/j.cognition.2018.06.009

Wassiliwizky, E., & Menninghaus, W. (2021). Why and how should cognitive science care about aesthetics? Trends in Cognitive Sciences, 25(6), 437 - 449. https://doi.org/10.1016/j.tics.2021.03.008

Yumoto, E., Gould, W. J., & Baer, T. (1982). Harmonics-to-noise ratio as an index of the degree of hoarseness. *Journal of the Acoustical Society of America*, *71*(6), 1544–1550. https://doi.org/10.1121/1.387808