Best-worst scaling, an alternative method to assess perceptual sound qualities

Victor Rosi, Aliette Ravillion, Olivier Houix, Patrick Susini

JASA EXPRESS LETTERS ARTICLE asa.scitation.org/journal/jel

Best-worst scaling, an alternative method to assess perceptual sound qualities

Victor Rosi, Aliette Ravillion, Olivier Houix, and Patrick Susini
Sound Perception and Design group, STMS, IRCAM, Sorbonne Université, CNRS, Ministère de la Culture, 75004 Paris, France
victor.rosi@ircam.fr, aliette.ravillion@gmail.com, olivier.houix@ircam.fr, patrick.susini@ircam.fr

Abstract: When designing sound evaluation experiments, researchers rely on listening test methods, such as rating scales (RS). This work aims to investigate the suitability of best-worst scaling (BWS) for the perceptual evaluation of sound qualities. To do so, 20 participants rated the "brightness" of a corpus of instrumental sounds (N = 100) with RS and BWS methods. The results show that BWS procedure is the fastest and that RS and BWS are equivalent in terms of performance. Interestingly, participants preferred BWS over RS. Therefore, BWS is an alternative method that reliably measures perceptual sound qualities and could be used in many-sounds paradigm.

[Editor: Qian-Jie Fu] https://doi.org/10.1121/10.0011752

Received: 28 February 2022 **Accepted:** 3 June 2022 **Published Online:** 24 June 2022

1. Introduction

Listening tests constitute an essential tool for research that aim at revealing perceptual qualities of sounds. For this kind of experiment, participants are often asked to evaluate sound stimuli that may vary in loudness, pitch, or timbre properties. These tests have had multiple applications for timbre semantic studies (Kendall and Carterette, 1993), sound quality evaluation (Jeon *et al.*, 2007), or human-in-the-loop audio annotation (Wang *et al.*, 2019). Therefore, the design of such tasks is crucial due to its impact on the difficulty and duration of the experiment, which may be conditioned by the needs of the study. For example, some studies investigating timbre perception have employed dissimilarity ratings based on a relative judgment format called pairwise comparison (Grey, 1977; McAdams *et al.*, 1995). Dissimilarity ratings are suitable to assess subtle differences between stimuli, as each stimulus serves as the standard in a series of relative judgments with the other stimuli. However, the pairwise comparison method imposes a small corpus of sounds as the number of trials $[N(N-1)/2]$ increases rapidly with the number of stimuli (N), which has negative consequences on participants' fatigue or motivation.

Apart from pairwise comparison, the most frequently used method to study the perception of timbre is the rating scale, whether it is for assessing the timbre of musical instruments (Kendall and Carterette, 1993; Pratt and Doak, 1976) or environmental sounds (Bjork, 1985; Jeon *et al.*, 2007; Zeitler and Hellbrück, 2001). The rating scale can be used either for a semantic differential procedure (Osgood, 1964), i.e., using bipolar scales with opposite terms (e.g., "dark–bright"), or for the verbal attribute magnitude estimation (VAME) method, i.e., using unipolar scales (e.g., "not bright–bright") (Kendall and Carterette, 1993). In a typical use of rating scales, participants are asked to rate stimuli with the scale. Then, the score of a sound corresponding to the studied dimension is calculated by averaging the ratings of all participants for this sound. Importantly, when the set of stimuli is presented beforehand, participants may have a relative use of the rating scales and adjust their use of the scale to the range of stimuli depending on the dimension being studied (Poulton, 1979). However, rating scales may fail to differentiate between stimuli with a similar value along an underlying dimension evaluation and may show multiple consistency biases (Baumgartner and Steenkamp, 2001; Schuman and Presser, 1996).

Recently, another relative judgment method called BWS (Louviere *et al.*, 2015) was proven to be a valuable alternative to the rating scale for consumer preference studies (Auger *et al.*, 2007), semantic judgments (Kiritchenko and Mohammad, 2017), and face perception (Burton *et al.*, 2019). A BWS procedure consists of asking participants at each trial to select the best and the worst item in a subset of k items ($k = 4$ in the conventional use of BWS), according to the studied dimension. By counting the number of best and worst judgments, the BWS procedure allows us to build a ranking of items from worst to best. Some studies adapted BWS in order to make it suitable in many-items context for a semantic judgment task (Hollis, 2018; Kiritchenko and Mohammad, 2017). Specifically, Hollis (2018) proposed to consider each trial as a tournament paradigm where the choice of best and worst made by participants brings additional inducted information on the pairs of sound within the subset of sound. For instance, in a trial with four items (A, B, C, D), if a participant chooses A as best and D as worst, then, in addition to the deducted information that $A > D$, we also consider that $A > B$, $A > C$, $C > D$, and $B > D$. Crucially, this paradigm allows us to disseminate the information between different sequences of trials using a scoring algorithm, e.g., the Rescorla–Wagner model (Rescorla, 1972), to build the ranking of the dataset of items. To our knowledge, this method has never been used to evaluate perceptual properties of sounds.

The present study aims to evaluate whether BWS is a suitable method for assessing the perceptual qualities of timbre. More specifically, we wish to test whether BWS is a valuable alternative to the rating scale (in this case, a VAME) for the evaluation of timbral brightness, one of the main dimensions of timbre (Kendall and Carterette, 1993; Pratt and Doak, 1976; Zacharakis *et al.*, 2014). To evaluate the performance of the two methods, we considered different questions: (i) How valid are the two methods considering the explicit definition of brightness? (ii) How reliable participants are? (iii) Is one of the two procedures faster than the other? (iv) What are participants' impressions of the two methods? To do so, participants evaluated the brightness of a musical instrument sound corpus using both methods. In this study, we considered and presented the brightness of a sound to participants as essentially defined through the quantity of high-frequency components, as it was demonstrated in some studies (Faure, 2000; Saitis and Siedenburg, 2020; Schubert and Wolfe, 2006). Thus, in addition to a comparison of BWS and RS brightness scores on the stimuli, validity of the two methods was also assessed through the correlation of their scores with spectral centroid values. However, as brightness does not solely depend on spectral centroid (Alluri and Toiviainen, 2010; Marozeau and de Cheveigné, 2007), participants were offered an explanation of this specific definition of brightness before the experiment.

2. Materials and methods

The experiment aimed to measure the performances of BWS and RS in evaluating the brightness of a corpus of musical instrument sounds.

2.1 Participants

Twenty volunteer participants (10 women and 10 men, mean age = 24.3 y, age range = 21–27 y) took part in the experiment. None reported having hearing problems. They gave their informed written consent before the experiment and were compensated for their participation. Participants had no sound or music education and were not familiar with either of the two methods.

2.2 Setup

Sounds were presented to listeners through a Beyerdynamic DT-770 PRO headset (Heilbronn, Germany) at an average level of 65 dB sound pressure level (SPL) (A-weighted). The sound level was measured with the sound level meter type 2250-S of Brüel & Kjær (Nærum, Danemark). Participants were tested in a double-walled Industrial Acoustics Company (IAC) sound-insulated booth. The test interface was coded with Max (v8) (https://cycling74.com/products/max) on a Mac Mini.

2.3 Stimuli

The corpus was made of N = 100 musical instruments sounds (i.e., woodwind, brass and string). The sounds were selected from the Studio-Online Library (Ballet *et al.*, 1999) and the Vienna Symphonic Library (Vienna Symphonic Library). Each sound was a recording of a musical instrument playing sustained notes for 5 s. All sounds were octaves of Cs ranging from C1 (32.70 Hz) to C7 (2093.00 Hz). We selected 100 sounds in a corpus of 200 sounds with respect to the constraint of a constant spacing between two consecutive sounds in terms of spectral centroid. Previous work measured the Just Noticeable Difference (JND) of spectral centroid between two sounds as being 5% bigger than the sound with the lowest spectral centroid (Allen and Oxenham, 2014). As a result, 98% of the pairs of stimuli presented in the four-sounds BWS trials had a difference equal to or greater than the spectral centroid JND. The spectral centroid of each sound was computed and averaged for each sound with the *Librosa* library (Klapuri and Davy, 2007). The loudness of each sound sample was equalized following the European Broadcast Union (EBU) norm on loudness (R-128) with the *ffmpeg* library (Python Package Index).

2.4 Procedure

At the beginning of the experiment, the concept of brightness was introduced to the participants using the following definition: "Brightness is relative to the amount of perceived high-frequency components in the sound. A very bright sound has a large amount of high-frequency components. A less bright sound has a small amount of high-frequency components; it can also be called muffled or dull." This definition was illustrated using four pairs of sounds of equal pitch and different nature (musical instrument, voice, and synthetic sounds). For each pair of sounds, the sound source was the same and the two sounds differed in brightness, i.e., the brighter sound had more high-frequency components than the other sound. Then, participants would proceed to the two methods in a randomized order with a training session on the interface using eight sounds for each method. We added retest trials for both methods to assess intra-participant consistency. Finally, participants completed a questionnaire asking for their impressions of both methods at the end of the experiment. Specifically, they were asked to rate the pleasantness and difficulty of each method on a 7-point Likert scale and to choose the preferred method for evaluating brightness.

During the RS procedure, sounds (N = 100) were presented to the participant one by one in a random order, and 20 sounds were repeated as retest trials at the end of the sequence for the intra-participant consistency measure. The

RS procedure was based on VAME procedures (Kendall and Carterette, 1993; Zacharakis *et al.*, 2014). After listening to each sound, participants were asked to evaluate its brightness on a 9-point Likert scale, going from "not bright at all" to "very bright." The scale was presented on the computer's screen and responses were given by selecting a point of the scale with the mouse. The brightness scores were obtained by averaging the ratings for each sound.

During the BWS procedure, sounds were presented to the participant in groups of four ($k = 4$). Thus, each participant evaluated the entire set of sounds through 25 trials of four sounds, with the addition of five retest trials at the end of the procedure for the measure of intra-participant consistency. At each trial, participants had to select the brightest and the least bright sound in each group of four sounds with the mouse. As provided in Hollis' design and to maximize the information propagated, we generated trial sequences for each participant so that each pair of sounds in a trial was presented only once over all sequences. Brightness scores were obtained, based on the information provided by the pairs of sounds in groups of four sounds, thanks to a scoring algorithm using the Rescorla–Wagner model used by Hollis (2018). See Hollis' personal page (Ualberta) for the generation of trial sequences and the implementation of the scoring algorithm.

Figure 1 presents the overall experiment (A) along with the criteria considered for evaluating the performances of both methods (B).

3. Results

We used four main criteria to compare RS and BWS: validity; reliability which can be broken down into inter-participant consistency and intra-participant consistency; participants' impression of the methods; and duration [see Fig. 1(B)]. Validity is the extent to which a test measures what it claims to measure, based on a "true" value. In this study, we evaluated validity by computing the correlation between the brightness scores obtained through the two methods with each other. Furthermore, since we explained to the participants that brightness is related to the amount of high-frequency components in the spectrum of a sound, we also assessed the validity of the methods through the correlations of the two sets of scores with the logarithm of spectral centroid. The inter-participant consistency is a type of reliability measure that indicates the extent to which participants agree with each other. The intra-participant consistency measures how consistent a participant is with himself by means of retests trials. Finally, participants' impression is assessed through a questionnaire asking them to rate the pleasantness and difficulty of the two methods, and to select the most adapted method for the task.

3.1 Validity

Figure 2 reports on the validity of the two methods, i.e., the correlation between the brightness scores of the two methods, and the correlations of these scores to the logarithm of the spectral centroids of the sounds. According to Fig. 2(A), there is a strong correlation between the scores obtained for the two methods [$r(98) = 0.94, p < 0.001$]. Moreover, for both methods, there are strong and comparable correlations with spectral centroid [$r_{BWS}(98) = 0.89$, $p_{BWS} < 0.001$; $r_{RS}(98) = 0.88$, $p_{RS} < 0.001$] [Fig. 2(B)]. Steiger's test (Steiger, 1980) for the comparison of correlations from dependent samples did not reveal a significant difference between the two correlations.

3.2 Reliability

Reliability was evaluated through the measure of inter-participant consistency and intra-participant consistency. Hollis' design imposed a unique presentation of each sound pair, and the fact that the BWS does not provide individual scores, well-known reliability metrics, such as the Krippendorff's alpha or the Cronbach's alpha, were not suitable here. Therefore, the compliance to mean scores (C) was used. Originally, this metric was introduced by Hollis (2018) to identify non-

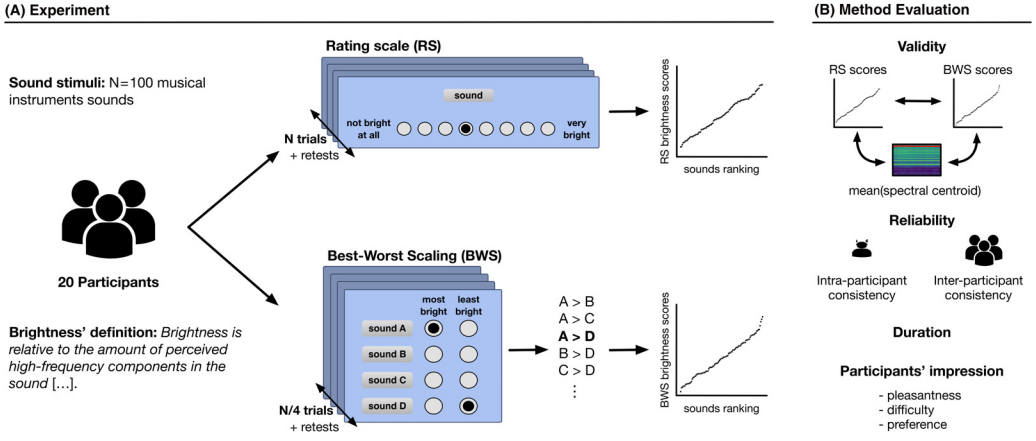

Fig. 1. Schema of the coupled BWS and RS experiment (A) with the evaluation criteria for the two methods (B).

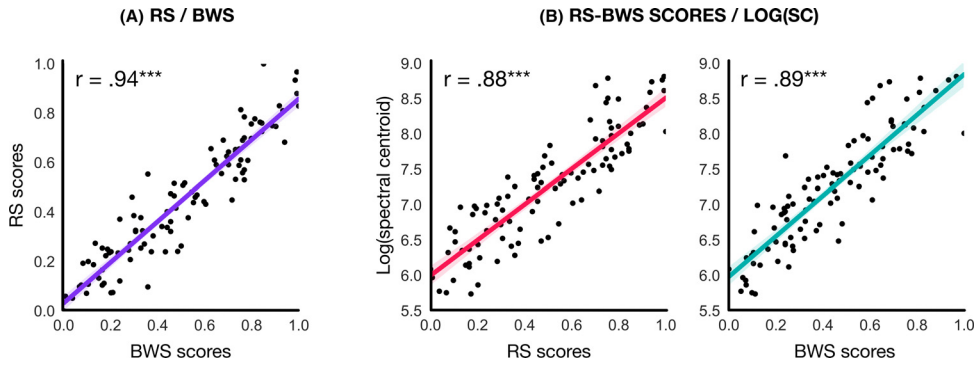

Fig. 2. Validity evaluation of the BWS and RS procedures. (A) Correlation between RS and BWS scores. (B) Correlations of the RS scores and BWS scores with the logarithm of the mean spectral centroids (log(SC)) of sounds.

compliant participants in a BWS experiment. Each participant's compliance with the group was calculated as the proportion of matching duels between the average scores calculated for the group and a participant's choices. In a trial, for example, if $A > B$, because a participant X chose A as the best and/or because B was chosen as the worst, we then compared whether this inequality is also true with the brightness scores obtained by the BWS scoring. Thus, a participant giving random responses would receive a compliance score of 50% with the group scores. To adapt this measure to RS ratings, we recreated the sound pairs considered for participants' BWS evaluation, and constructed inequalities based on their RS ratings. In the case of equal ratings for two sounds on a rating scale, the pair of sounds was not considered either for the RS or the BWS. As for the BWS, these inequalities were then compared to the averaged RS scores for all participants. The average compliance of the group thus provided a measure of inter-participant consistency. For both methods, compliance values were high ($C_{BWS} = C_{RS} = 83\%$) and did not differ significantly.

Intra-participant consistency is commonly measured by conducting tests and retests, i.e., presenting repeated trials in an experiment, and comparing their responses. For BWS, the test–retest measure was the proportion of duels answered similarly in the test and in the retest. Based on five repeated groups of sound, mean intra-participant consistency for BWS was 82% (random = 50%). For RS, test–retest was based on 20 repeated sounds. Here, intra-participant consistency was equal to 78% (random = 50%). Although we cannot compare the two measures, we nonetheless can conclude that on average, participants were able to do both tasks without trouble.

3.3 Duration

The BWS procedure and the RS procedure lasted, respectively, 9 min and 4 s and 9 min and 54 s on average. A paired t-test on the two normal distributions of the procedure times revealed that the BWS procedure was significantly faster than the RS procedure [$T(19) = 5.22$, $p = 0.03$, $d = 0.36$].

3.4 Participants' impression on the two methods

Participants were asked to give their impression of each method by rating them on pleasantness and difficulty, and by choosing the method most adapted for evaluating brightness. Evaluations of pleasantness and difficulty between RS and BWS were not significantly different. Considering the choice of the most adapted method, BWS was significantly preferred for evaluating brightness [$X^2(1, N = 20) = 5$, $p < 0.05$]. Participants also elaborated on their impressions of each method in writing. Thus, some of them argued that they struggled to calibrate their use of the scale during the RS task. Others found that the rating scale was not very accurate, and that they were not able to properly use extreme values. Additionally, some participants had the impression of being contradictory during the rating scale task. For BWS, some participants found it difficult to choose between similar sounds and that they had to listen to them several times. In addition, some participants felt more concerned about making a bad choice than they did with RS.

4. Discussion

In this study, we evaluated the performance of two methods for assessing timbral brightness based on validity, reliability, duration, and participants' impression. First, the scores of brightness obtained by both methods are highly correlated with each other and with spectral centroid values. This indicates that both BWS and RS methods provided good and similar evaluations of brightness, and thus, are comparable alternatives for studying perceptual qualities of sound. Second, the results on reliability for the BWS and RS tasks were equivalent, ensuring that participants could perform both tasks consistently at group and individual levels. Third, the BWS procedure was faster than the RS procedure. Because the time difference was only 50 s, we also evaluated the experimental times of the first half of the two procedures (i.e., the 15 first trials for BWS, and 30 first trials for RS). The BWS is still faster than the RS, but only by 26 s. This suggests that the time

difference between BWS and RS scales with the size of the sound corpus, and would tend to be greater in the context of annotating a larger amount of sounds. Finally, we found that participants' impressions on the two methods were the same, with a significant preference for BWS in terms of the method's suitability for the task. Thus, although showing similar performance, BWS may be a more satisfactory and comfortable method than RS for this type of task.

There is still some uncertainty about the extent to which brightness depends on spectral centroid or that it also interacts with other features like the fundamental frequency (F0) (Allen and Oxenham, 2014; Marozeau and de Cheveigné, 2007; Schubert and Wolfe, 2006). To avoid any possible confusion, we gave an essentially spectral definition of brightness to the participants before the experiment. Therefore, we were curious to report the ability of the two methods to distinguish brightness; in this case, understood as spectral centroid and the F0 in such an explicit context. Indeed, brightness scores of both methods are strongly correlated with the F0 of the sounds [$r_{RS}(98) = 0.90$, $p_{RS} < 0.001$; $r_{BWS}(98) = 0.82$, $p_{BWS} < 0.001$]. Interestingly, Steiger's test applied on both correlations of scores with the F0 revealed that BWS scores are less correlated with the F0 than RS scores [$Z = 4.75, p < 0.001$]. This suggests that in the BWS procedure, participants judgments were a little more faithful to the definition of brightness provided to participants. In particular, it may be due to the fact that the sound presentation format of the BWS favors the comparison of the brightness of equal F0 sounds.

Although the BWS procedure was comparable to the RS procedure for the measure of a perceptual property of sound, it showed specific disadvantages and advantages. On the one hand, the BWS procedure does not provide scores per individual—unlike the RS procedure, which makes any inter-participant analysis more challenging. In addition, the conditions of the BWS procedure (i.e., sequence generation, scoring algorithm) are more complex than those of the RS procedure. However, thanks to the contribution of Hollis (2018), the implementation of a BWS experiment is fast and straightforward. On the other hand, the BWS procedure was globally preferred by the participants, and took less time than the RS procedure. Thus, BWS seems to have crucial assets to consider for the design of online experiences, where it is important to spare the attention and time of the participants.

A growing number of sound perception studies rely on online crowdsourcing experiment designs to process larger quantities of sound stimuli evaluated by online participants (Cartwright *et al.*, 2016). One underlying reason for this trend is to provide more detailed analysis of perceptual sound qualities. In this context, and based on our results on duration and performance, BWS could be an interesting choice of experimental design. Moreover, according to other studies comparing RS and BWS, BWS can be conducted consistently on non-representative subsets of the entire sound dataset (Hollis and Westbury, 2018; Kiritchenko and Mohammad, 2017). Future works should therefore compare the two methods in a crowdsourcing context with the evaluation of a bigger dataset. In addition, future method comparison experiments should also involve other methods used to assess the perceptual properties of sound, such as the MUSHRA (Multiple Stimuli with Hidden Reference and Anchor) protocol which, based on a response format similar to that of BWS, has been shown to be suitable for assessing timbral brightness (Saitis and Siedenburg, 2020).

5. Conclusion

This work reports the suitability of BWS for accurately measuring the perception of timbral brightness. According to the criteria of performance (validity and reliability), duration, and preference, the coupled evaluation of a classic rating scale task and a BWS task attests for the equivalence of the two procedures, with a slight advantage in duration for BWS. Therefore, BWS, similarly to the RS, becomes a viable relative judgment method for assessing perceptual qualities of sounds when processing many sound stimuli.

Acknowledgments

This research was supported by the *Fonds K pour la musique*. The authors wish to thank Geoff Hollis (University of Alberta) and Svetlana Kiritchenko (National Research Council Canada) for key discussions when this work was initiated. They also thank the reviewers, Asteris Zacharakis (School of Music Studies) and the second anonymous reviewer for their thorough and comprehensive remarks on the manuscript.

References and links

Allen, E. J., and Oxenham, A. J. (**2014**). "Symmetric interactions and interference between pitch and timbre," J. Acoust. Soc. Am. **135**(3), 1371–1379.

Alluri, V., and Toiviainen, P. (**2010**). "Exploring perceptual and acoustical correlates of polyphonic timbre," Music Percept. **27**(3), 223–242.

Auger, P., Devinney, T. M., and Louviere, J. J. (**2007**). "Using best–worst scaling methodology to investigate consumer ethical beliefs across countries," J. Bus. Ethics **70**(3), 299–326.

Ballet, G., Borghesi, R., Hoffmann, P., and Lévy, F. (**1999**). "Studio online 3.0: An internet 'killer application,' for remote access to ircam sounds and processing tools," in Journées d'Informatique Musicale.

Baumgartner, H., and Steenkamp, J.-B. E. (**2001**). "Response styles in marketing research: A cross-national investigation," J. Mark. Res. **38**(2), 143–156.

Bjork, E. (**1985**). "The perceived quality of natural sounds," Acustica **58**, 185–188.

Burton, N., Burton, M., Rigby, D., Sutherland, C. A., and Rhodes, G. (**2019**). "Best-worst scaling improves measurement of first impressions," Cogn. Res. **4**(1), 36.

Cartwright, M., Pardo, B., Mysore, G. J., and Hoffman, M. (**2016**). "Fast and easy crowdsourced perceptual audio evaluation," in *2016 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, March 20–25, 2016, Shanghai, China, pp. 619–623.
Faure, A. (**2000**). "Des Sons Aux Mots, Comment Parle-t-on du Timbre Musical?," Ph.D. thesis, Ecole des Hautes Etudes en Sciences Sociales.
Grey, J. M. (**1977**). "Multidimensional perceptual scaling of musical timbres," J. Acoust. Soc. Am. **61**(5), 1270–1277.
Hollis, G. (**2018**). "Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments," Behav. Res. **50**(2), 711–729.
Hollis, G., and Westbury, C. (**2018**). "When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms," Behav. Res. **50**(1), 115–133.
Jeon, J. Y., You, J., and Chang, H. Y. (**2007**). "Sound radiation and sound quality characteristics of refrigerator noise in real living environments," Appl. Acoust. **68**(10), 1118–1134.
Kendall, R. A., and Carterette, E. C. (**1993**). "Verbal attributes of simultaneous wind instrument timbres: I. von Bismarck's adjectives," Music Percept. **10**(4), 445–467.
Kiritchenko, S., and Mohammad, S. M. (**2017**). "Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation," arXiv:1712.01765.
Klapuri, A., and Davy, M. (**2007**). *Signal Processing Methods for Music Transcription* (Springer New York, New York).
Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (**2015**). *Best-Worst Scaling: Theory, Methods and Applications* (Cambridge University Press, Cambridge, UK).
Marozeau, J., and de Cheveigné, A. (**2007**). "The effect of fundamental frequency on the brightness dimension of timbre," J. Acoust. Soc. Am. **121**(1), 383–387.
McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., and Krimphoff, J. (**1995**). "Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes," Psychol. Res. **58**(3), 177–192.
Osgood, C. E. (**1964**). "Semantic differential technique in the comparative study of cultures 1." Am. Anthropol. **66**(3), 171–200, available at https://www.jstor.org/stable/669329.
Poulton, E. C. (**1979**). "Models for biases in judging sensory magnitude," Psychol. Bull. **86**(4), 777–803.
Pratt, R., and Doak, P. E. (**1976**). "A subjective rating scale for timbre," J. Sound Vib. **45**(3), 317–328.
Python Package Index. https://pypi.org/project/ffmpeg-normalize/ (Last viewed June 17, 2022).
Rescorla, R. A. (**1972**). "A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement," in *Classical Conditioning II: Current Research and Theory,* edited by A. H. Black and W. F. Prokasy (Appleton Century Crofts, New York), pp. 64–99.
Saitis, C., and Siedenburg, K. (**2020**). "Brightness perception for musical instrument sounds: Relation to timbre dissimilarity and source-cause categories," J. Acoust. Soc. Am. **148**(4), 2256–2266.
Schubert, E., and Wolfe, J. (**2006**). "Does timbral brightness scale with frequency and spectral centroid?," Acta Acust. united Acust. **92**(5), 820–825.
Schuman, H., and Presser, S. (**1996**). *Questions and Answers in Attitude Surveys: Experiments on Question Form, Wording, and Context* (Sage, Thousand Oaks, CA).
Steiger, J. H. (**1980**). "Tests for comparing elements of a correlation matrix," Psychol. Bull. **87**(2), 245.
Ualberta. https://sites.ualberta.ca/~hollis/ (Last viewed June 17, 2022).
Vienna Symphonic Library. https://www.vsl.co.at (Last viewed June 17, 2022).
Wang, Y., Mendez, A. E. M., Cartwright, M., and Bello, J. P. (**2019**). "Active learning for efficient audio annotation and classification with a large amount of unlabeled data," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 880–884.
Zacharakis, A., Pastiadis, K., and Reiss, J. D. (**2014**). "An interlanguage study of musical timbre semantic dimensions and their acoustic correlates," Music Percept. **31**(4), 339–358.
Zeitler, A., and Hellbrück, J. (**2001**). "Semantic attributes of environmental sounds and their correlations with psychoacoustic magnitude," in *Proceedings of the 17th International Congress on Acoustics*, Rome, Italy, Vol. 28.