POLITECNICO DI MILANO
M.Sc. Music and Acoustic Engineering
School of Industrial and Information Engineering

A comparison of Best-Worst Scaling and Rating Scale for timbre characterisation

Équipe Perception et Design Sonores

Supervisors
IRCAM : Victor Rosi, Olivier Houix, Patrick Susini
Politecnico : Prof. Fabio Antonacci

Master graduation thesis
Aliette Ravillion
Matricola 935848

Academic Year 2020-2021

Abstract

The subjective annotation of sounds is a key area of focus in the field of timbre. Rating scales (RS) are the most used method for annotating data, although they present several limitations related to scale biases. The Best-Worst Scaling method (BWS) has proven to be a reliable alternative to Rating Scales on semantic and visual elements, but the two methods have not been compared on an audio corpus yet. This work focuses on the comparison of RS and BWS methods applied to timbre characterisation of instrumental sounds. Our results firstly show that both methods are comparable on performance (i.e. validity and reliability), although this highly depends on the number of participants on the task. Secondly, results convey that both methods are comparable on ergonomy, and thirdly, that BWS is more robust to the complexity of the task than RS. This study also reveals that the data obtained with the RS and the BWS carry different information. Finally, this work calls for a wider use of Best-Worst-Scaling in subjective sounds annotation tasks.

Sommario

L'annotazione soggettiva dei suoni è un'area chiave di attenzione nel campo del timbro. Le scale di valutazione (RS) sono il metodo più usato per l'annotazione dei dati, anche se presentano gravi limiti legati alle distorsioni della scala. Il metodo Best-Worst Scaling (BWS) ha dimostrato di essere un'alternativa affidabile alle scale di valutazione su elementi semantici e visivi, ma il confronto non è mai stato applicato a un corpus audio. Questo lavoro si concentra sul confronto dei metodi RS e BWS applicati alla caratterizzazione del timbro dei suoni strumentali. I nostri risultati mostrano in primo luogo che entrambi i metodi sono comparabili in termini di prestazioni (cioè validità e affidabilità), anche se questo dipende fortemente dal numero di partecipanti al compito. In secondo luogo, i risultati indicano che entrambi i metodi sono comparabili in termini di ergonomia e, in terzo luogo, che il BWS è più robusto rispetto alla complessità del compito rispetto al RS. Questo studio rivela anche che i dati ottenuti con la RS e il BWS portano informazioni diverse. Infine, questo lavoro richiede un uso più ampio del Best-Worst-Scaling nei compiti di annotazione dei suoni soggettivi.

ACKNOWLEDGMENT

My deepest thanks go to Victor Rosi, without whom I wouldn't have got very far on this. I am truly grateful for his solid and caring management, for all the time he invested on pulling my work up and for the countless things I got to learn with him.

Thank you as well to Patrick Susini and Oliver Houix who supervised this master thesis at IRCAM, and to Nicolas Misdariis for making this internship possible. Thanks for their trust, their guidance and for giving me an inspiring insight into the world of research.

I also wish to address my warmest thanks to all the PhDs, postdoc and interns that are members of the PDS department or of its coffee machine clientele. Daily life at IRCAM in your company was the best surprise, thank you for your wonderful welcome and your reachability whenever I had a question, for the mezzanine talks and the flawed uno games. I wish you all the best in your respective works.

Thank you to my supervisor from Politecnico, Prof.Antonacci, for his support, and thank you to all the professors of the MAE of Politecnico di Milano for their rich and dedicated teaching.

Thank you to my friends Clément and Andriana who enlightened this master and its groups projects, and thank you to my classmates from Centrale Nantes without whom I would have never achieved any administrative steps of this cursus.

Thank you to all the people who helped me in this study and its redaction.

Finally, I address my most tender thanks to my n°1 supporters, mum and dad, for always bringing me your serene and wise perspective in this adventure like in all the others.

Contents

Chapter 1

Introduction

Once a sound has been described with basic audio characteristics such as pitch, duration and loudness, timbral attributes are needed to further characterise it. However, describing the timbre of a sound with words is a tricky task, as timbre is a multidimensional perceptual quality and usually rely on a non-consensual and metaphorical vocabulary. Multiple studies focused on the vocabulary related to sound qualities (Faure, 2000; Kendall and Carterette, 1993). In a work investigating the vocabulary most used by sound professionals (e.g. sound engineers, composers, sound designers), Carron (2017) created a sound lexicon of 35 terms. However, some uncertainty remains for several of them as to their understanding and definition, such as for "warm" or "round".

Previous works explored the perceptual dimensions of timbre (McAdams et al., 1995; Zacharakis et al., 2012; Grey, 1977), and tried to establish correlations between these perceptual dimensions and acoustic features (Grey and Gordon, 1978). In the following, timbral attributes refer to the semantic characterisation of those perceptual dimension of timbre. To refine the definition of a timbral attribute, a possible approach is to annotate a corpus of sounds according to the attribute of interest (Wallmark, 2019; Kendall and Carterette, 1993; Zacharakis et al., 2012). Then, one can analyse the corpus with audio features and find out which appear to be relevant for the acoustic definition of this attribute (Zacharakis et al., 2012; Disley and Howard, 2004). However, previous timbre studies rarely work with a sound corpus containing more than a dozen of stimuli. Evaluating hundreds of varied sounds could help to better refine the characterisation of these attributes with subtlety and richness. This issue of subjective evaluation of big corpus can be addressed in different ways.

In psychophysics and more broadly in social and cognitive sciences, there are two well-known experimental methods for subjective evaluation: the semantic scale (or rating scale) and the pairwise comparison. The rating scale is a psychometric tool frequently used for the subjective annotation of sounds (Kendall and Carterette, 1993; Wallmark, 2019). In a rating scale annotation task, participants rate a stimulus on a scale representing the dimension to be studied. In contrast, the pairwise comparison procedure is a preference judgement method where each participant must choose between two stimuli according to a dimension under study. Yet, both methods are not optimal. On the one hand, the rating scale has some shortcomings such as the inter-participant consistency (Schuman and Presser, 1996). On the other hand, the pairwise comparison offers a good reliability of results, but requires a large

number of annotations of order N^2, N being the number of stimuli in the corpus. Thus, pairwise comparison imposes a reduced corpus of stimuli. One could vary the paradigm slightly with no-forced choice pairwise comparison but still facing the same limitations (Parizet et al., 2005). Therefore, we considered the Best-Worst scaling (BWS), a novel subjective annotation method introduced by Louviere and Woodworth (1991). In a BWS procedure, participants have to make a judgement on subgroups of k objects, and choose the "best" and the "worst" object within this group according to the dimension studied (e.g. preference, valence, etc.). The application of Best-Worst Scaling to sound judgement was recently inaugurated in a work on timbre conducted by Victor Rosi at IRCAM. In this study investigating the meaning of some timbre attributes and the relationships between them, 520 instrumental sounds were evaluated regarding these attributes with the BWS method.

In studies on semantics in many-item contexts (Hollis (2018b) and Kiritchenko and Mohammad (2017a)), BWS has shown a quality of annotation superior to semantic scales. However, it is not trivial that a BWS procedure would give similar results when used for sound annotation, since listening to sounds gives a new temporal dimension to the participant's task. Yet, no comparison between rating scale and BWS has been carried out in the audio field so far.

The present study addresses the question of which annotation method is the most relevant to collect the judgements of sounds timbres. We aim to consider the multiple aspects of an annotation method, namely its ergonomy and the quality of the collected data.

Chapter 2

State of the art

In the literature of timbre characterisation, we noted various experimental paradigms for sound evaluation, with the two main ones being the rating scale and the pairwise comparison.

2.1 Subjective annotation of sounds

2.1.1 Semantic differential and VAME scales

Rating scales have been widely used for timbre characterisation experiments. In a study focusing on the perception of timbre, von Bismarck (1974) proposed a method based on semantic differentials (Osgood, 1964), and created bipolar scales such as "bright"/"dark" to rate 35 sounds. The collected ratings were analysed with a dimension reduction technique and used to extract salient dimensions for the timbre. Later on, this method was adapted for a similar task by Kendall and Carterette (1993), who judged that the opposite terms at the end of each scale should not be arbitrarily determined by the researcher. Hence, they employed unipolar scales, also called Verbal Attribute Magnitude Estimation (VAME), to investigate the association between sounds and sound attributes. One example of VAMEs is a scale ranging from 'not nasal' to 'nasal'. They evaluated 15 sounds with VAMEs and observed a better differentiation of sounds than the one obtained with semantic differentials. Zacharakis et al. (2012) also used VAMEs to extract three semantic dimensions of timbre present in the greek and english language, which he called *luminance*, *texture* and *mass*. Darke (2005) studied the level of agreement that can be obtained from a collective view of 12 attributes, and conducted timbre judgement experiments on 5-point unipolar scales. In an other study investigating the agreement on the perception of polyphonic timbres, Alluri and Toiviainen (2010) used VAMEs to evaluate 8 attributes on a hundred of musical excerpts. Both Darke (2005) and Alluri and Toiviainen (2010) concluded that the reliability of the VAME strongly varied from one attribute to another. Morevoer, Alluri measured the duration of the annotation task which was about one hour per participant for a hundred of 1.5s musical excerpts. In a research treating crosstalks between timbre semantics and perception, Wallmark (2019) investigated the perception of brightness and smoothness on 93 instrumental and synthesised sounds of 1.5s. In this study, the duration of the task was recorded to gauge the ergonomy of the method. The experiment lasted 20 minutes in average, demonstrating that a rating scale judgement can be done relatively fast if one evaluates

few attributes on short sounds.
Although a VAME procedure is relative to a simple rating scale task, the design can be varied in order to give a better overview of the corpus of stimuli. In a work investigating the definition of timbral hardness in the freesound[1] database, Pearce et al. (2019) asked the participants to evaluate 206 sounds in batches of 8 sounds with rating scales. The most and the least hard sounds, chosen by an expert, were present in each batch as anchors. The 6 other sounds were chosen by ear by the expert so that they were approximately evenly spaced in hardness between the two anchors. In order to prevent scale compression effects, participants were asked to use the full range within the 8 presented scales. Even though there is a lack of 2 information on the method's consistency, the experiment was rather fast, with an average duration of 50 minutes per participant.

2.1.2 Pairwise comparison

A second experimental paradigm for sound evaluation is the pairwise comparison, in which sounds are presented in pairs. Two types of task can be performed: the pairwise similarity consists of a dissimilarity judgement between the two sounds, while the two-alternative forced choice consists of a dominance (or preference) judgement.

Pairwise similarity

The pairwise similarity is the application of pairwise comparison to dissimilarity judgements, and was widely used in precursor researches investigated the perception of timbre (McAdams et al., 1995; Grey and Gordon, 1978). In these studies, participants are asked to make dissimilarity judgements on pairs of sounds with a scale ranging from « not similar », to « very similar ». Then, a multidimensional analysis (MDS) allow to extract perceptual timbre spaces. Pairwise similarity uses relative information, and not an absolute rating like rating scales. It requires a number of annotations of the order of N^2, N being the number of sounds, and is therefore limited to small corpora (e.g. 12 sounds in the study of McAdams et al. (1995)). In a study on the perception of sound distortion, Michaud (2012) wanted to conduct a similarity task with a corpus of 37 sounds of about 10 seconds. Because of the size of the stimulus corpus, a pairwise similarity would have been too long. Therefore, the procedure was a preference permutation task where participants had to choose among three sounds, the sound most similar to a reference sound. Only 4 sounds were presented at a time in order to take heed of the auditory memory of the participant.

Two-alternative forced choice

In a two-alternative forced choice, participants choose between 2 items according to a specific attribute. As for pairwise similarity, this type of judgement requires a number of annotations of the order of N^2 and is limitated to small corpora.
Parizet et al. (2005) compared 6 different experimental methods adapted from rating scale

[1] freesound.org

and pairwise methods, in which participants evaluated the pleasantness of 9 in-car ventilation noises. Figure 2.1 presents the 6 methods employed.

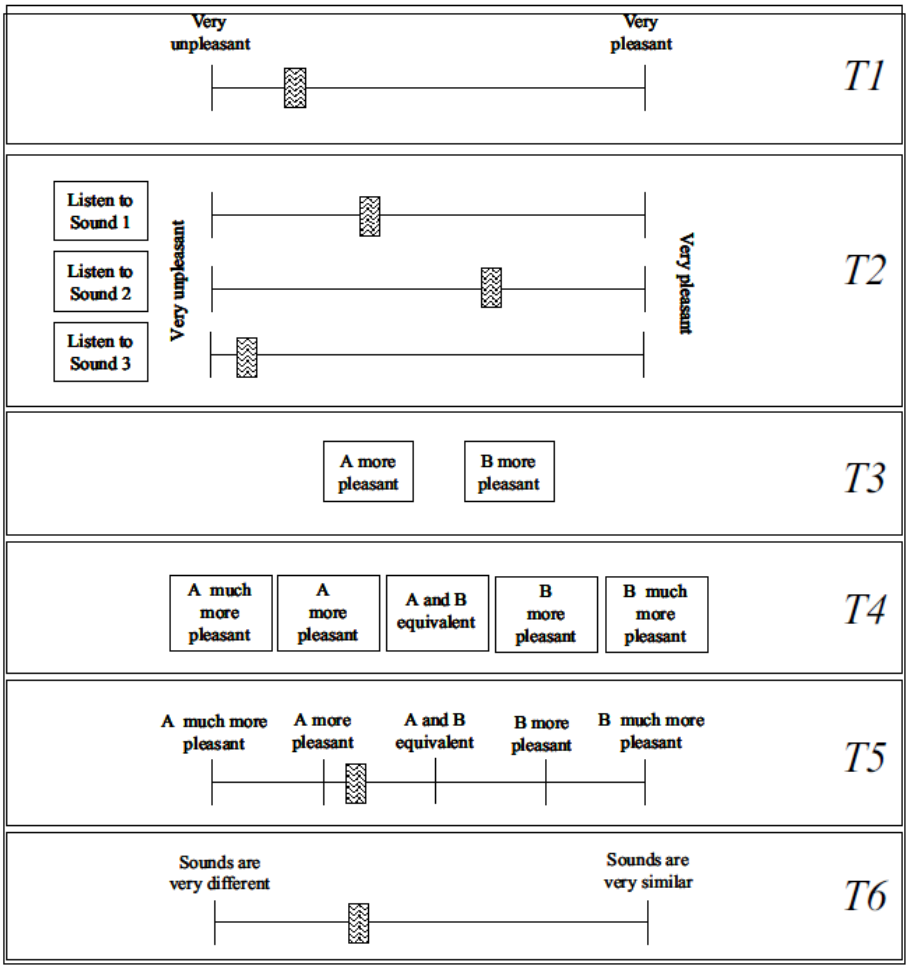


Figure 2.1: Answering scales of the six listening tests (Parizet et al., 2005)

- The first method (T1) was a rating scale ranging from 'very unpleasant' to 'very pleasant'. In the second method (T2), the 9 scales were displayed together so that participant could progressively adapt their ratings by comparing sounds all together.
- The third and fourth method were two-alternative forced (T3) and not forced (T4) choice between sound A and B. The fifth method (T5) was a slider between 'A more pleasant than B' and 'B more pleasant than A'.
- The last method (T6) was a pairwise similarity.

The rating scale methods took noticeably less time than the three pairwise comparisons, with 3min against 18min per participant. The subjects also found that the pairwise comparisons were longer.

As a result of clustering analysis, the pairwise comparison methods enabled a higher discrimination between two groups of listeners than rating scale methods.
In the end, the method consisting of displaying the 9 rating scales together offered the best compromise between the accuracy of the results and the time needed for subjects to realise the test, and is thus recommended if one only aims at evaluating the sounds pleasantness. Alternatively, the not-forced choice paired comparison maximised the amount of pairs of sounds having significantly different scores, and is recommended if one seeks the greatest possible discrimination between sounds. Additional dissimilarity judgements (T6) can be used to bring complementary information regarding the relationships between the items, but the not-forced choice paired comparison also provided this type of information. Unlike rating scales, it allowed to build a perceptual space of the items without the additional results of the pairwise similarity (T6), a task that doubles the duration of the entire experiment.

Pairwise comparison and pairwise similarity are interesting for the reason that they rely on relative judgements, unlike rating scale, but the literature showed that they require many annotations and take much longer than rating scale methods.
Since pairwise comparison is not a possible option for many-item contexts, we examined a new method of data collection that also relies on relative comparison of items with each other, but whose ergonomy is comparable to that of rating scales. This method is the Best-Worst Scaling (BWS), introduced by Louviere and Woodworth (1991). It has been increasingly used as an alternative to Rating Scale (RS) since it requires a similar amount of annotation while also overcoming some of the limitations of RS. As part of our goal to characterise the timbre of many sounds, we wish to determine which method is the most relevant between RS and BWS for the sound annotation task.

2.2 Rating Scale and Best-Worst Scaling

We present here the general principle of Rating Scale and Best-Worst Scaling, two experimental paradigms that interest us to account for the perception of timbre attributes.

2.2.1 Rating Scale (RS)

In the method of Rating Scale (RS), participants rate each item on a scale, one after the other (see the VAMEs in 2.1.1). The scale can be a slider with continuous values, or a discrete scale like a Likert scale which can take different discretisations (5,7,9-point scale) and which is now frequently used in social sciences (Likert, 1932). The score of one item is obtained by averaging all participants' rates for this item.

RS bias

While having the perk of a simple and ergonomic option for subjective annotation, the classic rating scale procedure is subject to a number of limitations (Schuman and Presser, 1996; Baumgartner and Steenkamp, 2001). The different response styles of participants when

answering to a rating scale can jeopardise the reliability of the final results. Here are the main limitation or biases for the rating scale:

- *Participant's bias:* Participants have their own personal perception of what each graduation on the scale is worth. Therefore, they may agree on the ranking of items while giving them different absolute rates.
- *Bias of extremes:* This bias occurs when participants avoid the extreme values of a scale, mainly focusing in the middle of the scale. Depending on the attribute, the opposite bias can also be observed: it is called the ceiling effect, met in satisfaction surveys for example (Masino and Lam (2014)).
- *Acquiescence bias:* This is the participant's tendency to mainly be in agreement with the studied concept, resulting in average scores above the scale's middle value. For instance when rating the importance of different strategies with a rating scale (Soutar et al., 2015), participants generally found that all strategies were important. Paulhus (1991) studied desirability bias and acquiescence bias, and highlighted that it can result in positive correlations between items that are not conceptually related.
- *Intra-participant consistency:* Participants are likely to change their opinion on the value of the scale's graduations over the course of the experiment. In other words, the longer the experiment, the harder it is for the participants to stay consistent.

Format and display

There are different options of format and display for the design of rating scales.

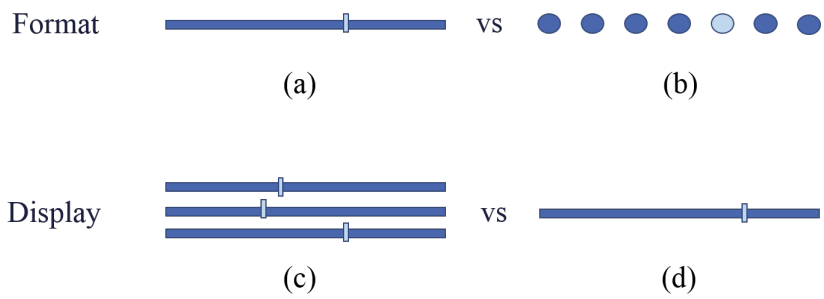


Figure 2.2: Different types of format and display for the Rating Scale: slider (a), Likert scale (b), multi-items display (c), one-by-one display (d).

Figure 2.2 presents the two main formats of the rating scale, the slider (a) and the Likert scale (b). A slider is a continuous bounded scale, providing interval data. A Likert scale is a discrete scale that can have 3, 5, 7... points, providing ordinal information. The interval interpretation of Likert scale data is very controversial, but it is common to see parametric models applied to Likert scales with 7 or more points when there is a large number of participants (Sullivan and Artino, 2013).

Some studies (Roster et al., 2015; Schaik and Ling, 2007) have compared these two formats without being able to identify the best one. In practice, above 7 points, the scores given by a Likert scale are very similar to the ones given by a slider, and the duration as well as the inter-participant consistency are also equivalent. The slider possibly attenuates the ceiling effect (Voutilainen et al., 2016), but the Likert scale seems to be the most ergonomic and most pleasant format for participants (Schaik and Ling, 2007; Voutilainen et al., 2016). This translates into higher amount of missing data for sliders in crowdsourcing contexts (Funke and Reips, 2012; Funke, 2016).

Figure 2.2 also presents two types of display of the rating scale, one displaying all the items simultaneously (c) and one displaying the items one by one (d). The study of Schaik and Ling (2007) compared both displays, but demonstrated no evidence of one being better than the other in terms of inter-participant consistency. The multi-item display (c) took much longer as participants go back and change their previous answers according to the new questions. Participants thus preferred the one-by-one paradigm (d). Similarly, Parizet et al. (2005) compared the two options with 9 sounds. The one-by-one display was faster, but the multi-item display provided significantly more accurate results. However, the experiment of Schaik and Ling (2007) covered more items (30), and it is likely that both display tend to provide similar results as the number of items increases.

Other design aspects are recommended in the literature, such as adding an "I don't know" option next to the scale in order to make the results more reliable (Roster et al., 2015).

2.2.2 Best-Worst-Scaling (BWS)

Best worst scaling (BWS) is a method of subjective evaluation introduced by Louviere and Woodworth (1991). In a BWS procedure, participants are asked to select the best and the worst item in tuples of k items (e.g. $k = 4$) according to the studied concept. Figure 2.3 illustrates an example of a BWS annotation where participants must select the most and the least attractive face among a 5-tuple. We note that a pairwise comparison is a Best-Worst Scaling with 2-tuples. By increasing the number of items per set from $k = 2$ to $k = 4$, the participant's task becomes more cognitively demanding, but more items are processed at once and more information can be derived from a trial.

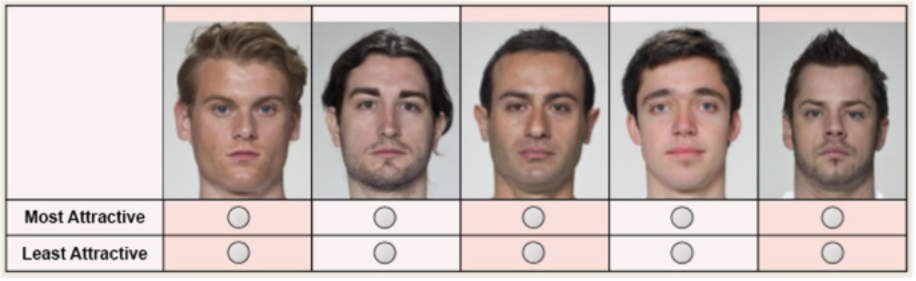


Figure 2.3: A trial in BWS (Burton et al. (2019)). The annotator must select the most attractive and the least attractive face.

Application

BWS has been increasingly used to distinguish consumer preferences. As a forced-choice method, BWS ensures a discrimination between the items and is thus an interesting method to distinguish preferences. Examples of the application of BWS in marketing contexts are the wine industry (Cohen and Goodman (2009)), the health sector (Flynn et al. (2007)), and more recently for energy policies (Aruga et al. (2021)).
Best-Worst Scaling can also be adapted to many-item problems, as done by Kiritchenko and Mohammad (2017a) and Hollis (2018a) in research on emotion and language. Both studies successfully applied BWS to thousands of items. Besides, Hollis (2018b) emphasises that the application of BWS to crowdsourcing contexts might be economically relevant, since the method has the potential to shorten the duration of a participant's task and therefore the global cost of the annotation process.

Scoring algorithms

To obtain a list of scores or a ranking of the items from the 'best'/'worst' responses, one uses a scoring algorithm. The most commonly used is a counting algorithm called the Best-Worst counting (Louviere et al., 2015; Kiritchenko and Mohammad, 2017a). The Best-Worst counting is quite simple as it corresponds to the substraction of the number of times an item has been selected as 'best' to the number of times it has been selected as 'worst'.

In an application of BWS to a many-item design, Hollis (2018b) introduces the use of another type of scoring, called tournament algorithms because they consider the whole of the trials as a tournament. The starting point of tournament scoring is that each judgement of a k-tuple allows to infer rankings also for items that have not been selected as 'best' or 'worst'. By choosing the best and worst on a 4-tuple, one can derive 5 duals, as illustrated on Figure 2.4. For example, if one chooses A as best and D as worst, one can infer A>D, A>B, ect.. BWS results thus generate a large amount of duals of items, which are the input data of tournament algorithm. The scoring process of the latter is similar to the Elo rating system[2] used in chess. Prior to a dual, the outcome of the match is predicted according to the scores of the players. After the dual, the scores of the two players are updated according to this prediction and according to the outcome of the match.

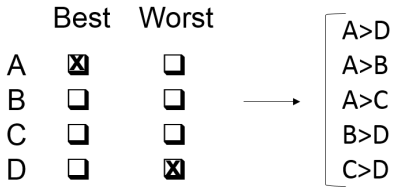


Figure 2.4: The 5 duals inferred from a BWS trial. Duals are the input data of tournament scoring algorithms.

[2]Method for calculating the relative skill levels of players in zero-sum games (Elo, 1978)

Hollis (2018b) compared counting algorithms, such as the Best-Worst counting, with more complex tournament algorithms such as the Elo scoring or the Rescorla-Wagner [3] scoring. Hollis (2018b) demonstrated with simulations that both tournament and counting methods produce similar scores, but that tournament scoring are more accurate and robust to inter-participant noise. Tournament algorithms are also longer to compute than counting algorithms, since scores need to be updated over many iterations before they converge.

Trials sequence design

The k-tuples that participants respond to can be optimised in order to collect the greatest amount of information. Different balanced incomplete block designs (BIBDs) exist (Louviere et al., 2015), but can not be used in many-item contexts. In order to score thousands of items, Hollis (2018b) established a design for the construction of 4-tuples, henceforth referred to as the *Hollis design*. The 4-tuples of all participants are generated all at once, following this three constraints:

1. Each 4-tuple is unique and seen by only one participant
2. Every item appears in an equal number of 4-tuple
3. No two 4-tuples share a pair of items. A pair of items is thus presented once for a participant.

Simulations conducted by Hollis (2018b) reported that the accuracy of scores was improved when applying these two constraints.

Number of annotations in BWS

When using scoring algorithms, BWS requires a certain number of annotations per item before satisfactory scores can be obtained. With 4-tuples, N/4 annotations are enough to see all items once but provide final scores distributed in three levels of responses, which correspond to the three choices 'best', 'unselected' and 'worst' (see simulation 1 in annex 2). Therefore, more annotations are needed to increase the sensitivity of BWS and to compute proper scores.
Kiritchenko and Mohammad (2017b) varied the number of annotations in a BWS experiment of N = 1367 words and sentences, by randomly selecting n annotations for each word among all the collected annotations. From n = 2 onward, that is to say with 2N annotations in total, the scores were 98% correlated with the scores obtained with 10N annotations. Kiritchenko and Mohammad (2017b) therefore set a minimum threshold of 2N annotations to compute BWS scores sufficiently close to their asymptotic value. However, this has not been verified in other experimental designs, on other types of items or attributes. Using BWS with 5-tuples, Burton et al. (2019) computed individual scores with N trials per participant.

[3] Model of classical conditioning used in discriminative-learning algorithms (Rescorla, 1972)

2.2.3 Summary

BWS appears to be a method subject to different rules and limitations than RS, as summarised on Table 2.1.
Firstly, it is based on relative judgement while RS is based on absolute rating, and the format of the results are thus very different. RS results are a matrix containing one rate per item per participant, while BWS raw results are k-tuples that can be different for all participants in Hollis design.
Secondly, BWS scores are derived with more or less complex scoring algorithms that require a certain number of annotations to provide accurate scores. In contrast, RS scores are simply derived by averaging participants' ratings, and can be derived for a single participant that saw each item once.
Thirdly, BWS avoids several biases that impact the RS. Participants do not need to remember the exact rates they used in previous trials to remain consistent with themselves and no longer face the intricacies of what each grade of the scale is worth. Unlike in RS, participants are forced to discriminate between items and cannot compress them into a neutral zone of a scale, although this has the drawback of forcing participants to choose between two items even when they consider them as close or equal regarding the studied attribute. Yet in the latter case, the two items have the same probability of being selected as 'best' and 'worst' later by other participants, and should eventually obtain similar scores as the amount of annotation increases.

	Rating Scale	**BWS**
Type of judgement	Rate an item on a scale	Choose the best and worst in a k-tuple
Scores computation	Average scores	Scoring algorithm
Limitation	Scale bias (e.g. extreme values)	Forced choice of participant

Table 2.1: Sumary of the main characteristics and differences between RS and BWS methods.

2.3 Comparisons of Rating Scale (RS) and Best-Worst Scaling (BWS)

Previously, we explained the functioning of RS and BWS. This section now presents different works comparing the two methods that have inspired this study.

2.3.1 Evaluation criteria

In the literature, the quality of annotation methods is assessed on the basis of their validity and their reliability at the group and at the individual level.

Validity is the accuracy with which the method measures what it is intended to measure, and can only be assessed if the 'true' values (or reference values) of the items are available. It is commonly measured by computing the correlation between the obtained scores and the true values. One can use the coefficient of Pearson that measures a linear correlation, or the coefficient of Spearman that measures a non parametric correlation based on the rank (see Annex 1).

Inter-Annotator Agreement (Inter-AA), or inter-participant consistency, is a type of reliability that indicates the extent to which participants agree with each other. Several metrics exist such as the Cronbach's alpha and the Krippendorff's alpha (see Annex 1), which are commonly used to assess the reliability of rating scales, and which are computed from the individual ratings of participants. This metric can be applied to BWS results by computing the individual scores of each participant from its trials.

Test-retest reliability, or intra-participant consistency, represents how consistent a participant is with himself throughout the experiment. It is commonly measured by conducting test-retests, in which participant answers to an item or to a trial twice at different points in time. For RS, a possible measure of the consistency of a participant is the correlation between the test and re-test ratings. This metric can be applied to BWS results by computing the individual scores of a participant from its trials, once for the test and once for the retest.

It can be noted that comparing RS and BWS raises several significant challenges. Firstly, measuring the validity of scores is not possible if reference values are not available, which is often the case when the studied dimension is a high-level or subjective concept such as a timbral attribute. Secondly, comparing the reliability of the two methods necessitates to compare two results of different formats. Common reliability metrics apply to a matrix containing one score per participant per item, that is to say vectors of individual scores, and RS raw results are already in this format. BWS raw results are not, and the participants' BWS trials can be used to calculate their individual scores, but only if the number of BWS annotations per participant is sufficient (see 2.2.2). Typically, if a participant sees each item only once in BWS, the individual scores can't be computed and usual RS metrics can't apply to BWS results.

2.3.2 Experimental comparisons of RS and BWS

Table 2.2 presents reference studies that compared the Rating Scale and BWS procedures with the help of the different criteria presented above. These studies follow various comparison strategies, shedding different lights on the comparison of the two annotation methods.

The experimental designs presented below are all different, but for each study, we aimed at reporting the total number of annotations collected in each method, knowing that:

- one RS annotation or RS trial is a scale

Author	**N items**	**Corpus**	**Attribute**	**Measures**
Soutar et al.	10	Concepts	Importance	Inter-AA
Kiritchenko and Mohammad	3207	Words, Phrases	Valence	Inter-AA, Test-retest
Hollis	1034	Words	Valence & others	Validity
Burton et al.	30	Faces	Attractiveness & others	Validity Inter-AA Test-retest
De Bruyne et al.	300	Tweets	Valence & others	Inter-AA

Table 2.2: List of reference studies comparing BWS and RS with their experimental parameters

- one BWS annotation or BWS trial is a k-tuple
- N is the number of items
- Seeing all items once means answering to N RS trials or to N/k BWS trials

Soutar et al. (2015)

In this experiment, BWS and RS were compared to judge the importance of N = 10 positioning strategies for the competitiveness of a company. For RS, 200 participants responded on a 7-point Likert scale, and each participant saw all items once. For BWS, 200 participants responded to 5-tuples. In total, there were 200N (200 x N) annotations in RS and approximately 200N in BWS.

The ranking of items is very similar in RS and BWS, suggesting that both methods evaluate the same latent dimension. After the individual scores were computed in BWS, the reliability was assessed by looking at the standard deviation of each item's score. The standard deviations of items were lower in BWS than in RS, suggesting a higher inter-annotator consistency in BWS. The study also highlighted the presence of biases in RS results, such as the endpilling effect which consists of overusing one extremity of the scale. A positioning map of the items built with a multidimensional analysis (MDS) showed that these biases resulted in little differentiation among the items rated in RS. BWS allowed for a better discrimination among the 10 strategies and provided additional information about interrelationships between them, such as the closeness of two given strategies. Soutar concluded that in this annotation task, BWS was not only more reliable but also provided more information than RS.

Kiritchenko and Mohammad (2017a)

In this study on semantic, participants judged the valence of N = 3207 words or phrases via the crowdsourcing platform CrowdFlower. For RS, 20 participants responded on a 9-point

Likert scale and each participant saw all items once. For BWS, 10 participants responded to 4-tuples and each participant answered to 2N tuples and saw all items 8 times. In total, there were 20N (20 x N) annotations in RS and 20N (10 x 2N) in BWS. In BWS, the 2N tuples were the same for all participants and were generated according to these rules:

1. There are no identical items within a tuple
2. Each item appears approximately in the same number of tuples
3. Each pair of items appears approximately in the same number of tuples

RS and BWS scores are highly correlated, indicating that both methods evaluate the same concept. The inter-annotator agreement is measured with *Split-Half reliability* (SHR), which consists of splitting the annotators in 2 groups and computing the correlation between the scores of the two groups. Further explanations on the SHR can be found in Appendix 1. In this experimental design, each half contained the answers to the same tuples than the other half. Overall, BWS obtains the highest SHR, and thus provides the most reliable scores. The difference of SHR between RS and BWS is higher on complex sentences than on simple words, suggesting that BWS particularly outperforms RS on the most complex judgements. The RS test-retest reliability is also lower for sentences than for words, confirming that the sentences where tougher to judge. Varying the number of annotations from 1N to 20N, Kiritchenko and Mohammad (2017a) found that the gap of SHR between RS and BWS reduces as the number of annotation increases, and that BWS particularly outperforms RS in terms of inter-annotator agreement when the amount of annotation is small.

Hollis (2018a)

In this experiment, participants evaluated N = 1034 words with four attributes which are Arousal, Concreteness, Age of acquisition and Valence. For BWS, 70 participants answered to 4-tuples, and each participant answered to a sequence of 260 trials, seeing all items once. The 70x260 distinct tuples were generated according to Hollis design (see 2.2.2). The RS scores were computed with an already existing database containing the results of 9-point Likert scales. To compare the efficiency of both methods, Hollis introduced the concept of data collection cost, which is the number of trials per items in a method. He compared RS and BWS at equal data collection cost according to this definition, that is to say with 16N RS annotations and 16N BWS annotations.

Hollis did not provide a measure of reliability for BWS scores, but outliers were discarded by measuring their compliance. The compliance of each participant to the group was computed as the proportion of the participant's 'best' and 'worst' choices that were in accordance with the scores of the group. The BWS scores were then recalculated keeping only the results of compliant participants.
Since there are no true values for abstract attributes such as for the Concreteness of a word, the validity can not be measured . An alternative is to use a 3rd task to evaluate the predictive validity of the scores. In this study, this 3rd task is the ability of the scores to predict the

Lexical Decision Reaction Time (LDRT) of the items, which is the reaction time it takes to tell whether a word is a real word or not. The LDRT of a word happens to be correlated with low-level features like the word's length, but also with high-level features like its Valence, its Concreteness and its Arousal. Hollis thus generated predictions of this LDRT with a regression model using variables including the studied attributes. The predictive validity of the scores was calculated as the correlation between the model's prediction and the reference values, which are experimental LDRT values taken from pre-existing databases[4]. Figure 2.5 reports the results for Arousal and Valence, where the previous norms correspond to the ones collected with rating scales.

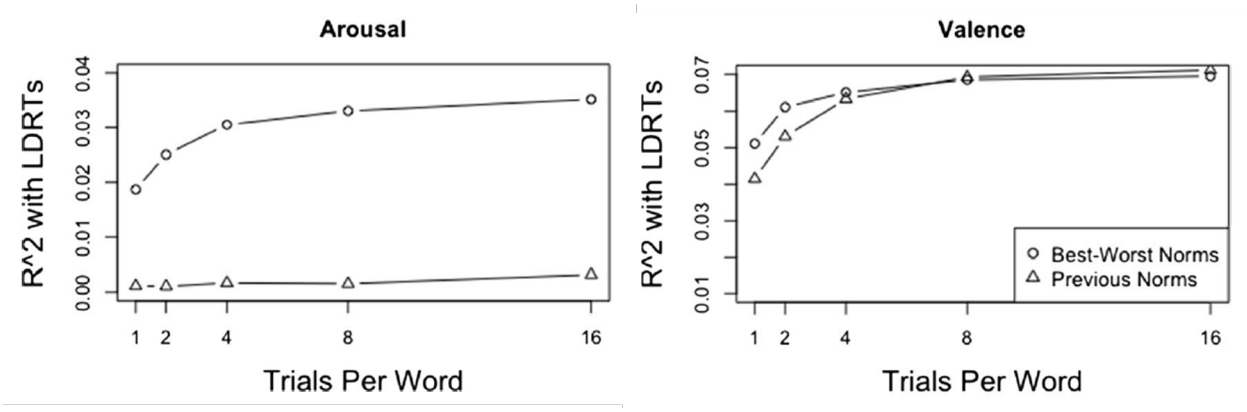


Figure 2.5: Validity of RS and BWS for two attributes in the study of Hollis (2018a)

Hollis observes that for Arousal, Concreteness and Age of acquisition, the predictions of the model based on BWS scores are more accurate than the predictions using RS scores. Furthermore, as the number of annotation increases, BWS tends towards a higher asymptotic validity than RS. The conclusion of this study is that RS and BWS provide different information, since "different response formats measure different aspects of a semantic construct", which is also in line with the previous findings of Soutar et al. (2015). Results are different for Valence, for which RS and BWS perform equally when the number of trials exceeds 8N (Fig. 2.5). This suggests that BWS particularly stands out for a small amount of annotations, but it also shows that the comparison of RS and BWS can lead to different conclusions depending on the studied attribute.

Burton et al. (2019)

Burton ran 3 experiments comparing RS and BWS on the judgement of N = 30 pictures of faces. The corresponding studied attributes were respectively the attractiveness, the distinctiveness and the trustworthiness of the faces. In each experiment, participants performed the RS task with a 9-point Likert scale and each participant saw all items once. In BWS, each participant answered to 30 5-tuples (see fig 2.3). RS and BWS were thus compared at

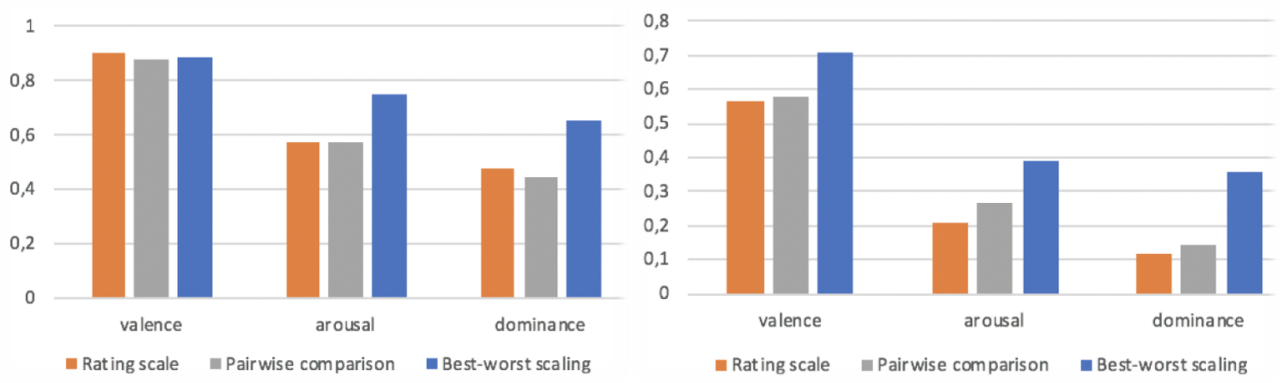
[4]English Lexicon Project (Balota et al., 2007) and British Lexicon Project (Keuleers et al., 2012)

an equal number of annotations, which was different for each attribute: 398N, 166N, 95N.

In the first experiment, Burton evaluated the validity of BWS and RS at the individual level. Participants performed a 3rd task consisting of the ranking of 3 faces, and the results of this 3rd task was compared to a prediction made with the results of BWS or RS results. Considering this validation task, BWS scores were more valid than the RS scores. Nevertheless, one might argue that the 3rd task used to assess the validity is closer to a BWS judgement than to an absolute rating, hence disadvantaging RS.
In the next two experiments, Burton asked participants to judge the faces twice, a few days apart. The individual scores were computed for both the test and the retest sessions, and the consistency of each participant was measured with the correlation of Pearson between the two sessions, although this does not detect an absolute difference of ratings between tests and retests but only a relative one. BWS proved to have a better test-retest reliability than RS for the two attributes.
A measure of inter-annotator agreement with Cronbach's alpha (see Appendix 1) on the results of RS and BWS gave a better inter-participant consistency score for the BWS.

Burton et al. (2019) points out that RS presents an additional cognitive challenge compared to BWS, and that RS might be particularly arduous for specific population having difficulties to maintain a good calibration of the rating scale, such as children or clinical population with memory issues. To test this hypothesis, Burton et al. (2021) re-iterated the study with children. BWS scores were again more reliable than RS scores, and the difference between RS and BWS in test-retest reliability was greater among children than among adults, confirming that children have more difficulty calibrating rating scales than adults.

De Bruyne et al. (2021)



(a) (b)

Figure 2.6: Two measures of inter-annotator agreement (De Bruyne et al., 2021). The Split-Half reliability (a) and the Krippendorff's alpha (b) lead to different conclusions.

In this study, a 7-point Likert scale, BWS and pairwise comparison were compared on the judgement of the Arousal, the Dominance and the Valence of N = 300 tweets. 6N annotations were collected in RS (6 x N) and 9N annotations were collected for pairwise comparison. For BWS, 1.5N distinct trials were generated according the same rules than Kiritchenko and Mohammad (2017a), detailed previously in this section. Each trial was then rated by 6 participants, providing 9N (1.5 x 6 x N) annotations for BWS in total.

Two different metrics of reliability were measured. The first is the Krippendorff's alpha (see Appendix 1), applied to RS and to BWS after individual scores were computed from participants' trials. The second is the Split-Half Reliability (SHR), a metric splitting participants in two halves and computing the correlation between the scores of the two groups (see Appendix 1). Figure 2.6 reports the values of these two metrics for each attribute. On 'Valence', the SHR value is slightly higher for RS whereas the Krippendorff's alpha is significantly higher for BWS. This shows that the SHR can lead to different conclusions than those of the Krippendorff's alpha, a much more common and recognised reliability metric. The authors propose the following explanation: "individual variability is by default already averaged out by taking the mean, so we believe this is not an optimal technique to assess inter-annotator agreement".
The reliability measures both convey that Valence is the most concrete and simple attribute, whereas 'Arousal' and 'Dominance' are more complex to judge. The authors conclude that "the benefit of best-worst scaling compared to rating scale annotations enlarges as the complexity of the annotation task increases", which was also the conclusion drawn by Kiritchenko and Mohammad (2017a) about the complexity of the sentences.
Regarding the ergonomy of the methods, participant found that BWS was the most difficult task. This is most likely related to the fact that BWS tasks were also 5 to 7 times longer than RS tasks, as each participant responded to 150 BWS trials and 100 RS trials.

Conclusion on previous experimental comparisons

Overall, BWS outperforms RS in terms of validity and reliability in all experiments.
A first finding is that the less annotations there are (above 1N), the more BWS outperforms RS (Kiritchenko and Mohammad, 2017a; Hollis, 2018a).
A second finding is that the more complex the task is, the more BWS outperforms RS. This idea is supported by the following ascertainment: Valence is the simplest and most consensual attribute to judge, as it has been observed in the previous experiments but also in other semantic annotation works (Wood et al., 2018). At the same time, with large amounts of annotations, the validity and reliability of RS and BWS are never as close as for the attribute Valence (Hollis, 2018a; De Bruyne et al., 2021). It is thus in line with the hypothesis that BWS and RS perform more similarly for simpler task.
Hollis is the only one comparing both methods at an equal data annotation cost, with four times fewer annotations in BWS than in RS. Even in this configuration, BWS still gives more valid results than RS. However, there is a lack of information regarding the reliability of the scores in this configuration.

2.4 Motivations for a BWS-RS experiment on sounds

This state of the art offers several valuable insights that motivate an experiment comparing Rating Scales (RS) and Best-Worst Scaling (BWS) on sound judgement.

Firstly, BWS appears to be a promising alternative to the commonly used RS. We saw that most of the annotation experiments involving a specific timbre attribute were using a type of RS called Verbal Attribute Magnitude Estimates, while Parizet et al. (2005) demonstrated with paired comparison that a relative judgement could be more suitable for this type of task. Unfortunately, pairwise comparison is not an option in many-item contexts, but the literature conveyed that Best-Worst Scaling can be an interesting alternative to the rating scale as it provides highly similar scores with often more valid and reliable results.

Secondly, this state of the art gives several guidelines to conduct the comparison of RS with BWS. The formerly presented studies compared BWS and RS's performances on the basis of the quality of the scores they provided, that is to say through the validity and reliability of the results. Other works investigated additional facets including the ergonomy of the methods. The studies finally showed that, depending on the complexity of the task and on the the number of participants, one could come to different conclusions regarding which method has the best performances.
The literature also offers us two different paradigms to juxtapose the RS and the BWS, the most common being the comparison at an equal number annotation, and the other one being the comparison at an equal data annotation cost, adopted by Hollis in the frame of considerations on crowdsourcing applications.

In the end, we have the opportunity to test a BWS procedure for the first time with a sound corpus. Using BWS could allow collecting new types of data labelling with hopefully more consistent, more accurate or larger datasets, and thereby benefiting to research on sound timbre. However, annotating sounds is a very different task than annotating faces, words or strategies as one can't have an overview of all 4 items simultaneously. There is a strong temporal dimension brought by listening to the items one after the other, and participants are for example likely to listen to all 4 sounds and then return to the first sound to listen to it again, having forgotten how it sounded. This is likely to affect the ergonomy of the annotation methods, and therefore their duration.

Chapter 3

Experiment

3.1 Objectives and progress

3.1.1 Objectives

Our objective is to compare Best-Worst Scaling (BWS) and Rating Scale (RS) on a task of judging the timbre of sounds.
A way to take further the characterisation of timbre attributes is use them to annotate a sound base and to see which audio features they are related with. One of the main challenges is to build up sufficiently rich and varied data sets to capture the subtleties of timbre perception, and thus to provide good material for future applications in machine learning and deep learning. Thereby, we wish to annotate the timbre of a large corpus of different instruments.
Having in mind the possible applications of BWS to crowdsourcing contexts and to large corpora, we aim to keep the participants' tasks reasonably short. Therefore, we set the BWS experiment in the same configuration than Hollis (2018b), where each participant sees each item exactly once in BWS. However, we choose to compare the efficiency of RS and BWS in a new paradigm where each sound occurs the same amount of time in RS and in BWS. This paradigm contrasts with the comparison of RS and BWS at equal number of annotation, mainly used in the literature and relying on the definition of data collection cost laid by Hollis (cf. 2.3.2).

Regarding the previous works presented in the state of the art, we wish to weigh RS and BWS on three main aspects: the performance (validity, reliability), the ergonomy, and the robustness to complexity.

Performance

Firstly, we want to compare the performances of the methods, that is to say the quality of the scores they provide. Data quality includes data validity and data reliability.
To assess the validity of the scores, we need a timbre attribute for which we can provide reference values. To do so, we chose the attribute 'Brillant', meaning bright, since the brightness of a sound was found to be highly correlated with an easily computable audio feature:

the logarithm of the Spectral Centroid (SC) (Schubert et al., 2004; Grey and Gordon, 1978; Schubert and Wolfe, 2006)).

To assess reliability, we will first check the consistency of participants at the group-level. This will tell us about the inter-annotator agreement, or in other words, the extent to which participants judge the sounds according to the same latent dimension. Reliability will also be evaluated at the individual level through test-retests, which will check whether each participant is consistent with himself over time. This will tell us about the repeatability of the annotation task, but also about how well-defined and comprehensive the studied concept is for each participant.

Ergonomy

Secondly, we aim to compare the ergonomy of both methods. Since the task's duration is a key economic factor in a crowdsourced data collection process, we wish to gauge the time and the number of annotations needed for each method. We also wish to collect the participant's perception of the difficulty, pleasantness and duration of each method.

Complexity handling

Thirdly, we want to evaluate how RS and BWS behave on tasks of different complexity. For this purpose, the experiment is replicated on another attribute, 'Riche', which we seek to make more complex than the task on 'Brillant'. Different levers can be used to differentiate the two tasks in terms of complexity, such as giving a definition to participants for one of the attribute and not for the other.

3.1.2 Progress and experiment stages

Previously, we have presented a state of the art for timbre annotation, that led us to define the objectives and the three main angles our experiment. This section presents the work stages that followed, from the design phases to the conduct of the experiment.

1. **Review of the metrics**
   Assessing the reliability of the scores is not straightforward, as BWS individual scores can't be computed in our design. Common metrics used for RS don't apply, and hence we adapted custom reliability metrics for Best Worst Scaling.

2. **Simulation**

   We conducted a phase of simulation in Python with the objective of exploring the behaviour of RS and BWS in different conditions and testing our metrics.

   As a first step, we built a simulation program with the following pipeline:

(a) Creation of random true values:
True values were generated randomly following a continuous uniform distribution.

(b) Generation of participants and their latent values:
The latent values of the participants were computed from the true values with different noise conditions. We implemented inter and intra-participant noises, but also 3 biases of the rating scale: the tendency to avoid the extremity of the scale, the participant's bias and the participant's bias changing over time (see 2.2.1).

(c) Building the BWS trials:
The BWS trials were 4-tuples built according to Hollis design (cf. 2.2.2).

(d) Performance of RS and BWS task:
RS and BWS responses of each participant were inferred from their latent values. For RS, we simulated the answers to a 9-point Likert scale.

(e) Scoring of BWS and RS results:
The scoring of RS consisted of a simple average of the results. For BWS, we used the counting and tournament algorithms made available by (Hollis, 2018b).

(f) Computation of the validity and reliability metrics:
Validity was measured as the correlation with the true values. We also computed inter-annotator reliability metrics and test-retest reliability metrics.

As a second step, we aimed at observing the influence of the number of participants, the number of items and the noise conditions over the validity. The main conclusions are that in the absence of RS biases, RS performed better than BWS no matter the number of participants. Adding RS biases, we found that RS validity was lowered whereas BWS scores stayed almost unchanged. BWS eventually became more valid than RS as RS biases were amplified. Also, we observed no impact of the number of items on the accuracy of scoring algorithms and on the validity of RS or BWS (above 20 items). The detailed simulations can be found in the Annex 2.

As a third step, we computed the Split-Half reliability, the compliance to the mean scores and test-retests metrics to better understand their behaviour in different noise conditions. We computed their values for perfect and dummy participants to characterise their bounds and to be able to correctly interpret their values in the real experiment. These metrics and their limits are presented further in the section 3.3.

3. **Building the test rig**

   After running the simulations, we prepared the test rig to conduct tests with real participants. We coded the interfaces of the RS and the BWS tests with Max/MSP, and we selected a corpus of sounds from an existing database made of thousands of instrumental sounds. We extracted their characteristics and applied a few selective criterion to form the experimental corpus. We also built the analysis pipeline.

4. **Pilot test** The pilot tests were conducted within the laboratory. We were able to check the feasibility of a first task on 'Brillant' with 100 sounds, and another task on

'Riche' with 200 sounds. This allowed us to test the robustness of the interface and to gauge the tasks' duration and the agreement of participants regarding this attributes. We also considered a task with more sounds, but regarding the state of the art and for practical purposes, we finally favoured testing several attributes on 100 sounds rather than several sizes of corpus.

5. **Experiment**
   In the light of the simulations and the pilot tests, we established the final protocol and recruited participants to perform the experiment.

3.2 Materials

3.2.1 Subjects

Subjects were 20 participants, 10 women and 10 men, aged between 20 and 30. They were all non-musicians and were paid 30€ for this experiment.

3.2.2 Sound corpus

The corpus was made of N=100 sounds of woodwind, brass and string instruments. The sounds were selected from the project Studio-Online Library (SOL) (Ballet et al., 1999), a sound library recorded at IRCAM. Each sound was the recording of an instrument playing a constant note for 5 seconds. All sounds were octaves of Cs ranging from C1 (32.70Hz) to C7(2093.00Hz). For comfort reasons, the loudness of each sound sample was normalised following the EBU norm on loudness (R-128), setting all the sounds to -23 LUFS[1].

Spectral Centroid extraction

To better fit to the judgement of the brightness ('Brillant'), the corpus was balanced in Spectral Centroid (SC) (see formula Eq. 3.1). We conditioned the choice of the 100 sounds by minimising the differences of SC values between two sounds, with respect to the just noticeable difference of SC (Allen and Oxenham, 2014). The idea behind such constraint is that we didn't want to design BWS trials in which it would be impossible for participants to discriminate the brightness of sounds. The sounds' SC values were thus logarithmically spaced between 300Hz and 7000Hz, but in spite of our efforts, there remained a few proportion of the trials that contained a pair of sounds with a difference of SC below the threshold (2% of all pairs in total).In annotation tasks involving subjective or high-level attributes, it is likely that some trials do not respect the corresponding discrimination levels.

$$SC = \frac{\sum_{k=0}^{N-1} f(k)x(k)}{\sum_{k=0}^{N-1} x(k)} \tag{3.1}$$

$f(n)$ = center frequency of the bin $n°k$
$x(n)$ = weighted magnitude of the bin $n°k$

[1]EBU R 128 is a recommendation for sound level normalisation - https://tech.ebu.ch/publications/r128/

The Spectral Centroid was computed in Python with the Librosa library [2]. Figure 3.1 represents the spectrogram of a violin playing a C4, with the tracking of the temporal Spectral Centroid plotted in blue. To compute the mean SC value of the sound, we first weighted the SC of each time window with the energy of the signal at this window. We discarded the low-energy windows, corresponding to silences or transitional periods, by retaining only the SC values superior to 95% of the maximum SC. Then, we computed the median of the remaining SC values to obtain the global Spectral Centroid value of the sound.

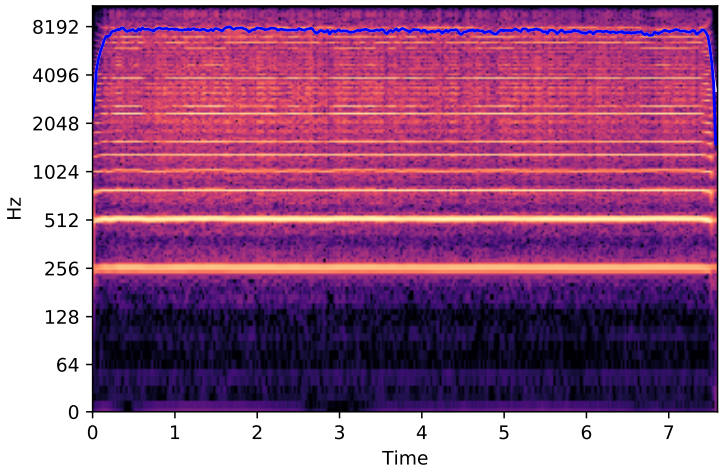


Figure 3.1: Spectrogram of a C4 violin with its Spectral Centroid (in blue)

Design of BWS and RS trials

The BWS trials are 4-tuples. We generated 20 distinct sequences of trials according to Hollis design (cf. 2.2.2) to have one unique sequence per participant. No two trials shared a common pair of sounds among all generated trials, and each sound appeared once per sequence. Participants were answering to the same trials for 'Brillant' and for 'Riche', but presented in a different order and with a shuffle of the 4 sounds at each trial.
20 sounds of retest were added at the end of each task. Consequently, in order to maintain the same amount of sounds between the two procedures, a BWS sequence was made of 30 trials including 5 trials of retest and a RS sequence was made of 120 trials including 20 retests. The retests sounds were different for all participants, and were taken among the second and third quartile of the participant's test, to avoid a too close proximity of tests and retests sounds. The order of appearance of the sounds was randomised for each participant in both methods. The scores were computed without taking the retests into account, that is to say with 5N (20 x N/4) annotations in BWS and 20N (20 x N) annotations in RS in total.

[2]Librosa is a python package for music and audio analysis, developed by McFee et al. (2015)

	Rating Scale	BWS
Test	100 trials	25 trials
Retest	20 trials	5 trials
Nb of sounds	120	120

Table 3.1: Summary of the BWS and RS tasks of one participant for one attribute. A BWS trial is a 4-tuple and a RS trial is a scale.

3.2.3 Apparatus

Sounds were presented to listeners through a DT 770 PRO headset at an average level of 65dB(A). The sound level was measured with the sound level meter type 2250-S of Brüel and Kjær. Participants were in a listening booth equipped with a Mac mini and isolated from exterior noise. The test interface was coded on Max/MSP, a widely used music software package developed by Ircam in the 1980s and now by Cylcing'74. BWS and RS interfaces are displayed on Figure 3.3 and Figure 3.2.
The RS interface is a simple RS 9-point Likert scale. Participants can click the button play/pause as much as he desires, and must give a rate before switching to the next sound. On the BWS interface, 4 sounds are displayed, and blue buttons flag the sounds that have not been listened yet. Participants must select the most 'Brillant' and the less 'Brillant' sounds, and switch to the next trial.
For both methods, a red spot would flash if participants listened to a sound more than 3 times. It aimed to encourage participants not to hesitate too long, and to answer spontaneously. For each interface, a loading bar informed participants of their progress in the task.

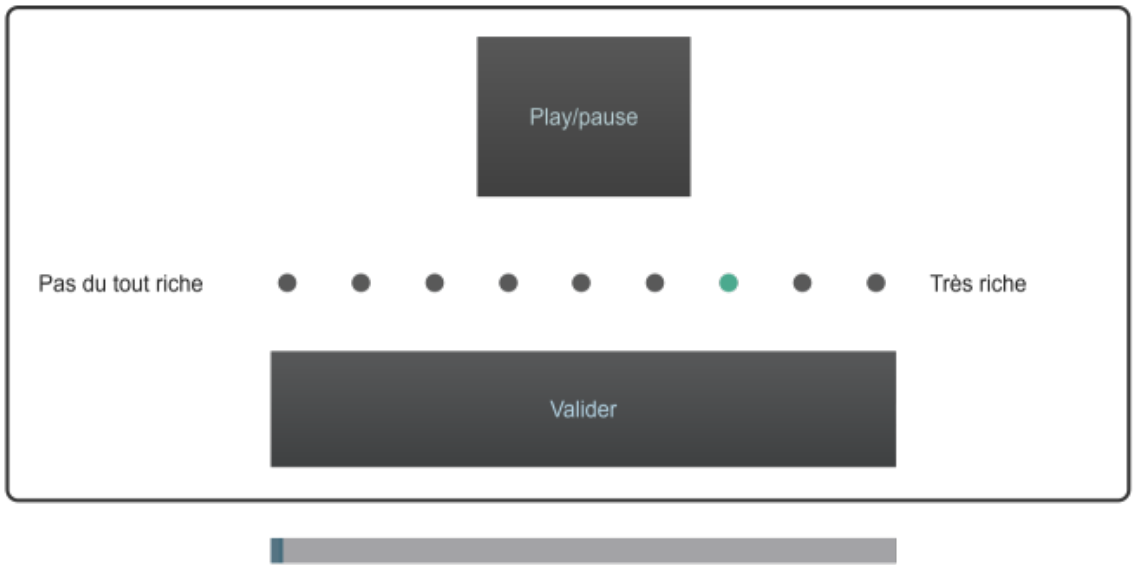


Figure 3.2: Max/MSP user interface for the Rating Scale

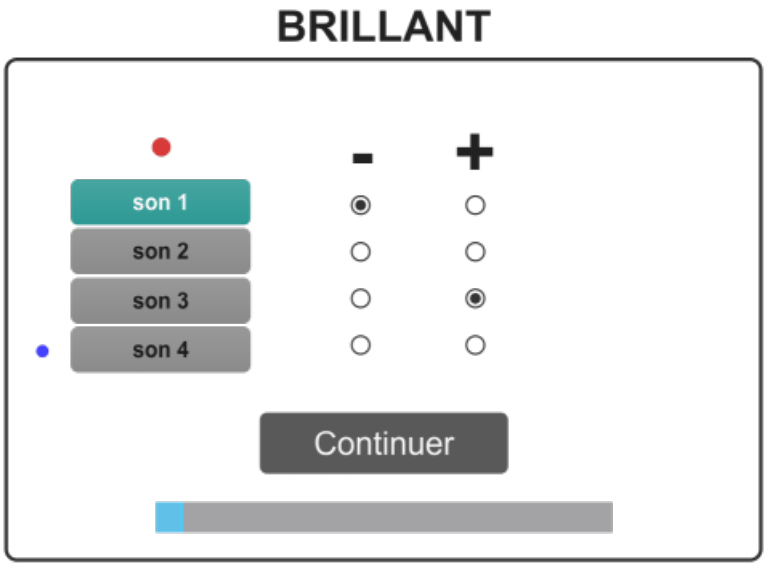


Figure 3.3: Max/MSP user interface for the Best-Worst Scaling

3.2.4 Procedure

The procedure was designed so that the task on 'Riche' was more complex than the task on 'Brillant'. We use three levers to differentiate the two tasks in terms of complexity:

A. The attribute 'Brillant' will be explained to the participants with a definition and audio examples, whereas no definition nor example will be given for 'Riche'.

B. Participants will achieve the task on 'Riche' first, discovering the sounds for the first time. For the task on 'Brillant', they will have seen each sound twice already (in RS and in BWS), and will thus have more insight into the corpus.

C. The corpus of sounds will be balanced in Spectral Centroid, thus suitable for a judgement of the brightness ('Brillant'), but not necessarily for a judgement of the richness ('Riche'). On 'Riche', RS and BWS are thus compared in a context closer to real annotation tasks, where the distribution of the studied dimension is neither known nor controlled.

The experiment lasted 1h30 per participant in average, and followed 8 steps:

1. **Instructions**
   The instructions were given through a 2 minutes video explaining the principle of the task and the functioning of the interfaces. The full script is available in the Annex 5.

2. **Training on RS and BWS interfaces** Participants trained at both interfaces with 8 sounds, for which they were asked to rate the pitch of the sound from high to low.

3. **RS-BWS tasks on 'Riche'**
   For the 1st task, participants judged the richness of 120 sounds on BWS and RS, with

the last 20 sounds being for retests. The order of the two tasks was permuted for each participant.

4. **Questionnaire for 'Riche**
   Then, participants answered a questionnaire (available in Annex 5) about their perception of the tasks. A five minutes break was taken after this questionnaire.

5. **Learning phase on 'Brillant'**
   For the timbral attribute 'Brillant', participants were given the following definition: "A bright sound is a sound that has a lot of high components. One can say that a sound which is not bright is dull or muffled". This definition was illustrated with four pairs of sounds. Each pair was made two similar sounds (e.g. two voice sounds, or two synth sounds), with one sound being bright and the other not very bright.

6. **RS-BWS tasks on 'Brillant'**
   For the 2nd task, participants judged the brightness of 120 sounds in BWS and in RS, with the last 20 sounds being for retests.

7. **Questionnaire for 'Brillant'**
   Participants answered to the same questionnaire than for 'Riche' (see Annex 5).

The order of sounds was randomised for each participant. Also, the order of the method was completely balanced within the group of subjects, that is to say that each possible ordering (e.g. RS-BWS-BWS-RS) occurred for 5 participants.

3.3 Analysis

Participants answers were collected in .JSON files and analysed in Python.
The first step of the analysis was the scoring of the results. The next steps follow the three comparison axis defined in the objectives: the assessment of the methods' performances through validity and reliability, their ergononomy and finally their ability to handle complexity.

3.3.1 Scoring the results

The RS score for one sound is computed as the mean of all participants' ratings for this sound.
The BWS scores are computed with two kinds of scoring algorithm: the counting and the tournament algorithms (see 2.2.2). To investigate the differences between these two, we computed BWS scores with the Best-Worst counting and with the Value scoring, a tournament algorithm adapted from the Rescorla-Wagner model (Hollis, 2018b).

The Best-Worst scoring is a counting algorithm that simply counts the number of times the items were selected as 'best' and 'worst'. Each item is scored according to equation 3.2.

$$V_{item} = \frac{N_{best} - N_{worst}}{N_{trial}} \tag{3.2}$$

N_{trial}: number of occurrence of the item

The Value scoring is a tournament algorithm based on the induced information between pairs of sounds in a trial. It has a good accuracy and robustness to noise (Hollis, 2018b), and needs about hundred of iteration to converge. First, the algorithm builds the list of all the duals directly inferred from the trials (see Fig.2.4). At each iteration, the algorithm goes through the entire list of duals in a random order. For each dual $V_{win} > V_{loos}$, the values of the two items are updated according to equations 3.3 and 3.4. Before moving on to the next iteration, the list of duals is reshuffled and the learning rate β decreases.

$$V_{win} = V_{win} + \alpha\beta\,(1 - V_{win}) \tag{3.3}$$

$$V_{loos} = V_{loos} - \alpha\beta\,(V_{loos}) \tag{3.4}$$

$$\alpha = 1 - \frac{V_{win}}{V_{loos} + V_{win}} \tag{3.5}$$

V_{win}, V_{loos}: values of the winning and loosing item
β: learning rate, decreasing at each iteration
α: salience of the dual

The salience parameter α is the extent to which the outcome of the match fits to the prediction, and is computed according to equation 3.5. For example, if $V_{loos} = 0$ and $V_{win} = 1$, the outcome was fully predictable and the salience parameter is 0, hence the scores don't change. In the opposite case where $V_{loos} = 1$ and $V_{win} = 0$, the outcome of the match is totally opposite to the prediction. The dual is thus considered as highly salient, the salience parameter is at its maximum 1, and the dual strongly modifies the items' values.

3.3.2 Performance metrics

Once the scores have been calculated for both attributes, we assess their quality using validity and reliability measures.

Validity

Our validity metrics are the correlations of Pearson and Spearman between the scores and the reference values. We do not have reference values for 'Riche', so the validity is only assessed for 'Brillant'. As previously introduced (see 3.1.1), the reference value for each sound is the logarithm of its Spectral Centroid, as it scales well with our perception of the brightness. The correlation of Pearson is commonly used to assess the validity of data regarding reference values. It doesn't require to normalise the scores, unlike the Mean Square Error (MSE) which is an other type of validity measure.
Since BWS scores are based on ordinal choices, we also want to know which method is the best in terms of ranking, which is why we compute the correlation of Spearman.
Significance - To tell whether the validity of RS and BWS are significantly different,

we conduct a statistical test of comparison between two dependent correlations, which is described in the Annex 1.

Reliability

We chose a design where each participant answers to N/4 trials in BWS, which is not enough to compute individual scores. Without individual scores, we cannot use the reliability metrics that are commonly used in the literature such as the Krippendorff's alpha or the Cronbach's alpha. Therefore, we used two alternative metrics of inter-annotator agreement: the Split-Half reliability (SHR) and the compliance to mean scores.

The **Split-Half reliability** (SHR) consists of splitting the group of participants in two halves, and computing the correlation of Pearson between their two scores. We repeated the process over 100 iterations, with different shuffled splits of participants at each iteration. The SHR is the average value of the 100 iterations.
Significance - To tell whether the SHR of the two methods are significantly different, we use the comparison of two independent correlations, described in the Annex 1. We compare the two final SHR values of RS and BWS, but since this final value is an average of all iterations, we also wish to test the significance of each iteration. Thus, we compute the proportion of iterations for which the SHR is significantly different between RS and BWS.
Shortcomings - The SHR was originally conceived to assess the internal consistency of a test by splitting the items in half, and was later altered to assess the inter-annotator agreement by splitting participants in half. Kiritchenko and Mohammad (2017a) used the SHR in that way and found that BWS had a higher inter-annotator agreement than RS. Yet, our experiment sets in different conditions. The two halves contained answers to the same trials and there were as much RS annotations than BWS annotations, whereas in our experiment, BWS trials are different in the two halves and there are four times less BWS annotations in each half that in RS. De Bruyne et al. (2021) also used the SHR to assess the inter-annotator agreement, but criticised its functioning (see 2.3.2) and observed that it occasionally led to different conclusions from those of the Krippendorff's alpha, a more certified reliability index.
In addition, simulations (see the simulation 5 in the Annex 2) suggested that the SHR can significantly benefit to RS, with a difference up to $\Delta r = 0.10$ between the SHR of BWS and RS for an equal simulated inter-participant noise. A possible explanation is that the SHR judges not only the inter-annotator agreement, but also the ability of the methods to compute accurate scores with only half of the participant. Regarding this latter criteria, RS might outperform BWS if they aren't enough BWS annotations in each half to compute proper BWS scores.

The **Compliance to mean scores** is the second inter-annotator agreement metric that we considered. It was introduced by Hollis (2018a) to spot non-compliant participants (see 2.3.2).
In BWS, the compliance to mean score of a participant is the proportion of duals inferred from a participant's answers (see Fig. 2.4) that are consistent with the mean scores computed for the group. To adapt this measure for RS, the RS results of participants are converted to

BWS results by simulating their answers to their BWS trials according to their RS ratings. Then, we can compute the proportion of duals consistent with the mean scores, like in BWS. The general compliance of each method is the compliance to the mean scores averaged over all participants.
Significance - Each participant has its individual value of compliance. One can therefore compare RS and BWS with a Student test (see Annex 2) on the two groups of individual compliances.
Shortcomings - The simulation 4 in Annex 2 presents the evolution of the compliance to the mean scores with the inter-participant noise. One can see that RS and BWS have a similar compliance. However, without noise, the compliance of BWS does not reach 1 whereas participants perfectly agree. It is due to the limitations of the scoring process that can't achieve perfect scores with only 5N BWS annotations.
A second limit is that the compliance to the mean scores is an average of individual measures, and converge when the results of the group converge. Thus, adding new participants can increase the reliability of results without increasing the compliance to the mean scores. Unlike usual reliability metrics, it doesn't assess the absolute and global reliability of the results, but only the mean agreement of each individual with the group.

3.3.3 Ergonomic criteria

With the Max/MSP interface, we recorded the listening duration of each sound in addition to the total duration of the task For each participant, we normalised the duration of the 4 tasks by the participant's average task time. An analysis of the variance allowed us to see the effect of the method and the effect of the attribute on the duration.
Furthermore, we collected participants' answers to the questionnaire (cf. Annex 6), and conducted an analysis of the variance to determine if a task was perceived as more pleasant, more difficult, longer or more adapted than the other.

3.3.4 Complexity handling

The choice of the two attributes 'Brillant' and 'Riche' was motivated by the will to investigate how each method handled complexity. For RS and BWS, we observed the evolution of the inter-annotator agreement and the ergonomic metrics between 'Brillant' and 'Riche', and additionally, we computed the three following test-retest reliability metrics:

- In BWS, we computed the **test-retest compliance**. It is the proportion of duals (cf. 2.4) answered similarly in the test and in the retest.
- In RS, we computed a **test-retest compliance adapted for RS**, that aims to reproduce the previous BWS measure. All possible duals are inferred from the 20 retest ratings. Then, one computes the proportion of those duals that are consistent with the ranking inferred from the test ratings of the participant.
- In RS, we also computed the **alpha of Krippendorff** between the test and retest.

Although these three metrics all meant to measure the intra-participant consistency, they are different and cannot be compared to contrast RS and BWS, but only from one attribute to the other.

3.4 Results

3.4.1 Scores

Once all participants completed the experiment, we computed the scores for the four tasks.

Value and BestWorst scoring

RS scores were computed by averaging the participants' rates, while BWS scores were computed with two different scoring algorithms.
BestWorst scoring and Value scoring provided very similar scores, correlated at r = 0.997 for 'Riche' and r = 0.998 for 'Brillant'. Figure 3.4 shows the ranked scores for the two scoring methods. One can see that the two curves overlap, but that the scores computed with BestWorst look more like a step line chart, and give more discrete values than Value scoring. This in line with the finding that tournament algorithms are more accurate than counting algorithms. Yet, it should be noted that the two sets of scores are highly similar while the Value scoring is significantly more complex and longer to compute. All the metrics of performance that we computed are almost equal for Value and BestWorst scores, thus for the sake of simplicity, we will limit to one type of scoring and present the results of BWS with Value Scoring.

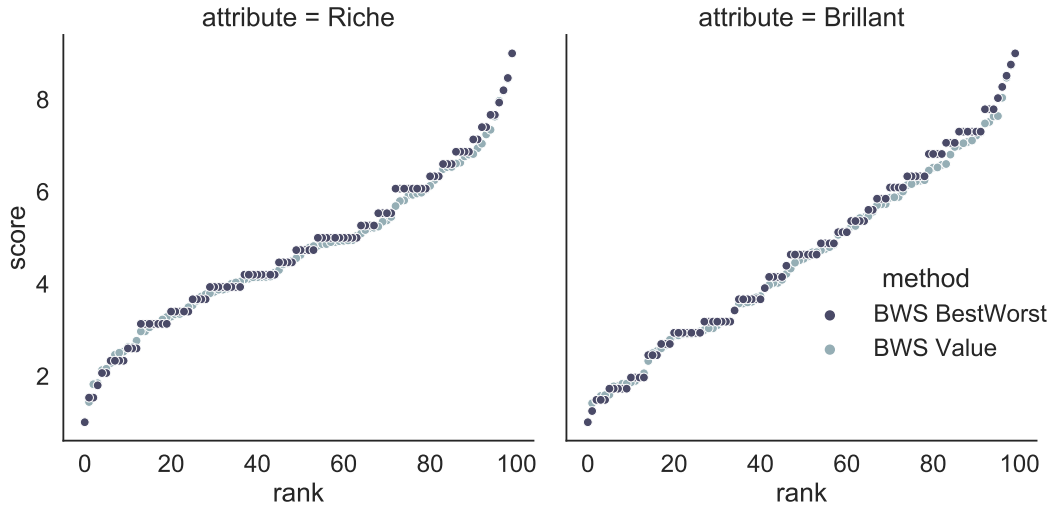


Figure 3.4: BWS results scored with Value and BestWorst algorithms

Similarity of RS and BWS scores

Table 3.2 reports the correlation between RS and BWS scores. The two methods are quite highly correlated, which indicates that they evaluate the same latent dimension. The results on 'Riche' are less similar than on 'Brillant', which shows that depending on the task, there can be more or less subtle differences between the scores of the two methods.

Attribute	Similarity	
Brillant	$r = 0.93$	$\rho = 0.94$
Riche	$r = 0.85$	$\rho = 0.84$

Table 3.2: Similarity between RS and BWS scores

Additionally, the Annex 3 presents the ranking of the instruments for 'Riche' and for 'Brillant'.

3.4.2 Performance

Validity

Figure 3.5 reports the validity of RS and BWS, measured with the correlation of the Spectral Centroid with the scores on 'Brillant'. The correlations report that both methods are equally valid. The scores of BWS are slightly more correlated to the Spectral Centroid than RS, but the difference between the correlations is not significant (Z = 0.24[3]).

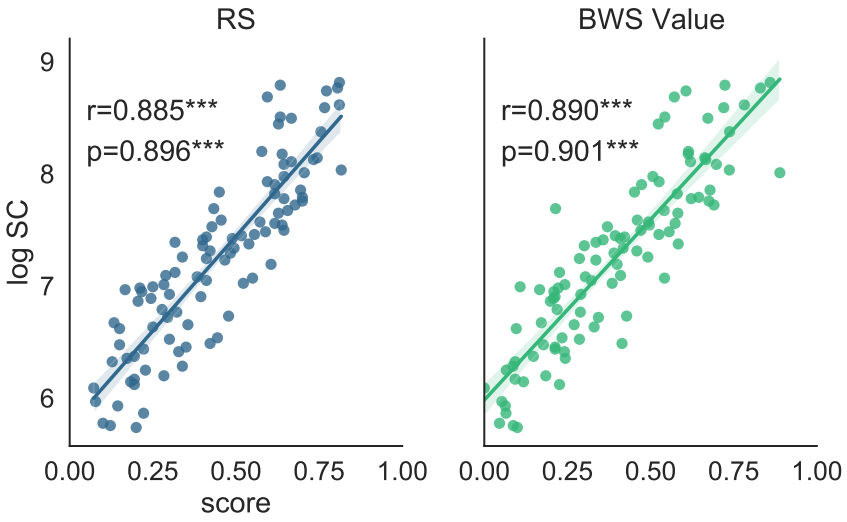


Figure 3.5: Validity of RS and BWS scores for 'Brillant'.

Confusion with pitch - The Figure 3.6 reports the correlation of the scores on 'Brillant' with the pitch (Fo). RS scores are significantly more correlated to pitch values than BWS

[3]The 95% significance level is reached when Z>1.96 (see Annex 1)

scores (Z = 3.45), which indicates that the pitch influenced participants' answers more in the RS task than in the BWS task. In addition, the correlation with pitch in RS (r = 0.90) is higher than the correlation with the Spectral Centroid (r = 0.885), which is meant to be the main dimension judged. This result is consonant with the study of Allen and Oxenham (2014) on the the discrimination and the confusion between pitch and brightness. They found a significant interference between the two dimensions, with an increase of the Spectral Centroid often being confused with an increase of the pitch and vice versa. A likely explanation is that when confronted with sounds of the same pitch, which was often the case since there were only 7 different pitches in the corpus, participants were forced to discriminate them on the basis of brightness in BWS. On the opposite, in RS, participant were free to judge sounds according to their pitch instead of their Spectral Centroid, without necessarily noticing it.
Therefore, the validity of RS and BWS is equal but the two methods provide qualitatively different scores.

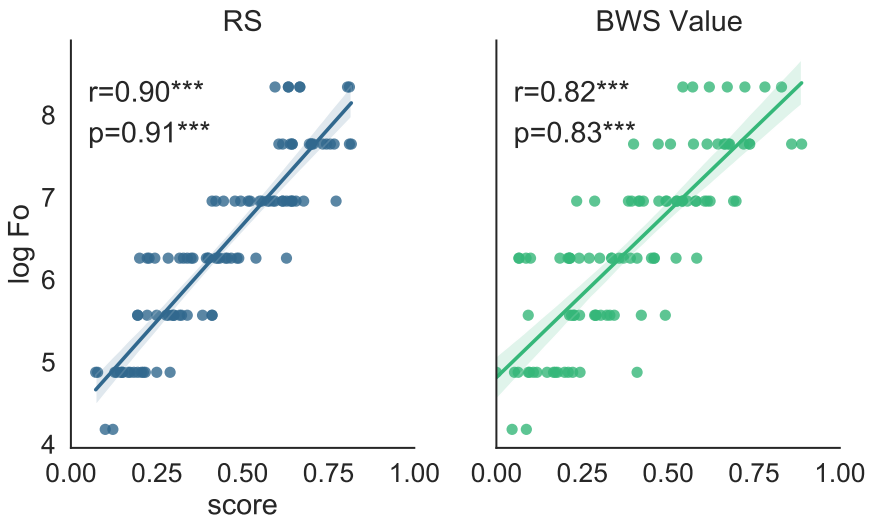


Figure 3.6: Correlation of scores with the logarithm of pitch for 'Brillant'.

Evolution of the validity - To investigate further the difference between RS and BWS validity, we plotted the evolution of the validity with the number of participants on Figure 3.7. The validity of BWS is increasing more steeply, and the curves suggest that RS might be closer to its asymptotic value than BWS. This firstly shows that each method can be the most relevant, depending on the number of participants. Regarding the previous results on the confusion of brightness with pitch, an hypothesis is that BWS scores can reach a higher asymptotic validity than RS scores since they are less biased by the pitch, but no conclusion can be drawn without pursuing the experiment with more participants.

To qualify the tendency observed on Fig.3.7, one can eventually notice that our results follow a pattern that is contrary to the results of Hollis on semantic norms. In the latter

(Fig. 2.5), BWS was more valid than RS for a small number of annotations (1N to 4N). In our case, it is RS that is more valid than BWS until 4N (equivalent to 16 participants). This suggests that the behaviour of RS and BWS as a function of the number of annotations also depends on the nature of the task.

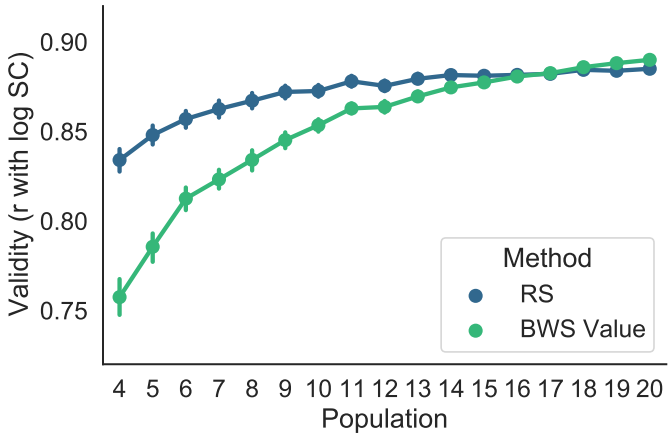


Figure 3.7: Evolution of the validity with the number of participant for 'Brillant'.

Inter-annotator agreement

Table 3.3 reports the Split-Half reliability (SHR) coefficients for the different methods and attributes. On 'Brillant', RS has the highest SHR value and the difference is significant ($Z > 1.96$) for every iteration. On 'Riche', RS still has the highest SHR, but is not significantly different from the SHR value of BWS. Looking at each iteration of the SHR computation, we also found that only 26% of the iterations presented a significant difference, therefore the SHR of RS and BWS are comparable on 'Brillant'.

Attribute	Coefficient	RS	BWS
Riche	r	**0.76**	0.68
	ρ	**0.75**	0.66
Brillant	r	**0.94**	0.83
	ρ	**0.94**	0.84

Table 3.3: Split-Half reliability of RS and BWS on two attributes

Figure 3.8 reports the compliance values to the mean scores for the two methods. The compliances of RS and BWS are almost equal on 'Brillant', and on 'Riche', BWS has a slightly higher compliance but the difference with RS fails to reach significance ($p = .09$).

On 'Brillant', the two metrics indicate that RS and BWS scores are equally reliable. On 'Riche', the compliance to the mean scores does not decide between the two methods, whereas the SHR is lower for the BWS. This difference in the conclusions can be enlightened

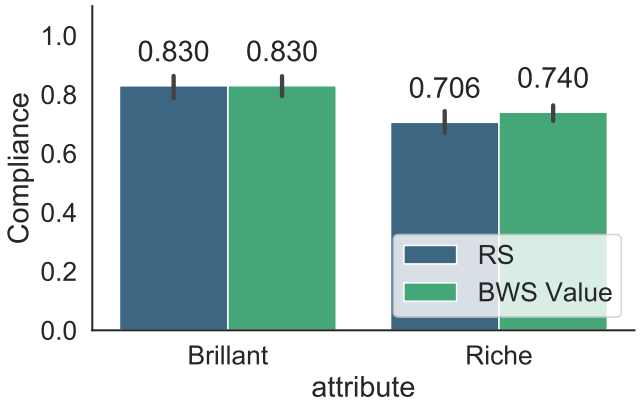


Figure 3.8: Compliance to the mean scores of RS and BWS on two attributes

by the limits of the SHR previously presented in 3.3, which suggest that the SHR might favour the RS over the BWS.
In sum, it can be concluded from the SHR and compliance that the reliability of RS and BWS scores are highly comparable for both attributes.

3.4.3 Ergonomy

Task duration

Figure 3.9 reports the average duration of each task, and shows that BWS took less time than RS for both attributes. BWS lasted in average 7min35 for 'Brillant' and 10min43 for 'Riche', whereas RS lasted in average 8min23 and 11min45. After normalising the duration of each task, we could conclude that RS was significantly longer than BWS on 'Brillant' ($p = .0.0028$) and on 'Riche' ($p = .017$).

The duration of both methods are yet of the same order of magnitude, and in average, a single BWS trial took three to four times longer than a single RS trial. This supports the idea that to compare RS and BWS at a similar temporal cost, the method should be compared at an equal number of appearance of each sound, and not at an equal number of annotation.

Listening mode and response styles

Participants listened to the sounds very differently in both annotation methods. In RS, participants listened to the presented sounds 1.2 times in average, while in BWS, the average number of listening per sound is 2.0 times. These two means are significantly different on 'Riche' and on 'Brillant' ($p<.001$) according to a Student test realised on all the listenings of the task. In BWS, participants were indeed more likely to switch from one sound to another to make the comparison, as some participants pointed out in the questionnaire. This shows that participants don't follow the same listening strategies to answers RS and BWS trials.

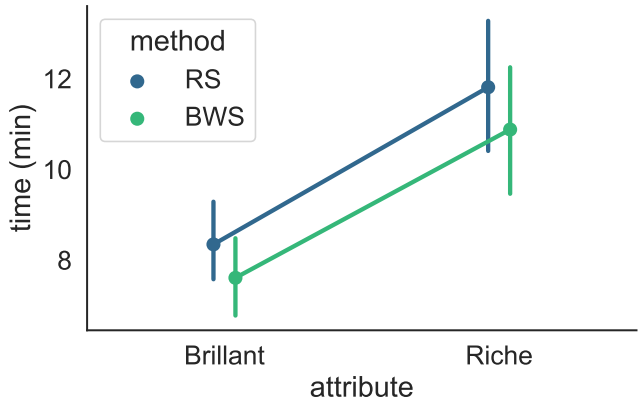


Figure 3.9: Duration of the RS and BWS tasks

In addition to the way participants listen to the sounds, we also observed the behaviour of participants regarding their use of the rating scales. The Annex 4 reports different response styles adopted by participants in the experiment, that are in line with some biases evoked earlier.

Participants' feelings

In the questionnaire, participants were asked to rate the pleasantness and the difficulty of each task. At the group level, the participants' opinions are balanced and no significant difference appears between RS and BWS. The average grades for the pleasantness and the difficulty of RS and BWS are equivalent, regardless of the attribute. At the individual level, each participant could clearly tell which method they found to be the easiest and the most pleasant, as they always answered by discriminating both methods. They put a difference of 1.6 point in average between RS and BWS's grades. Some participants argued that they struggled to calibrate their use of rating scale. Others found that the scale wasn't very accurate, and also that it was hard to use extreme values. Some also felt like they were contradicting themselves when answering to RS trials. In BWS, participants found it hard to choose between similar sounds and said that they had to listen to them several times. Also, some felt more worried about making a wrong choice than they did with RS. The participants differentiated the pleasantness and the difficulty of RS and BWS more on the first task on 'Riche' (mean difference of 1.9) than on 'Brillant' (mean difference of 1.3).

Participants elected BWS as the method reflecting the best their opinion in the two judgement tasks. More precisely, 75% of participants (see 3.10) chose BWS which is a significant majority according to the Chi-square test ($p = .025$). This result can be qualified by the fact that participants discovered the BWS during this experiment. They might have been attracted by the novelty of the method and been subject to desirability bias.

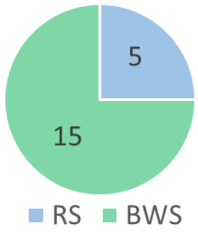
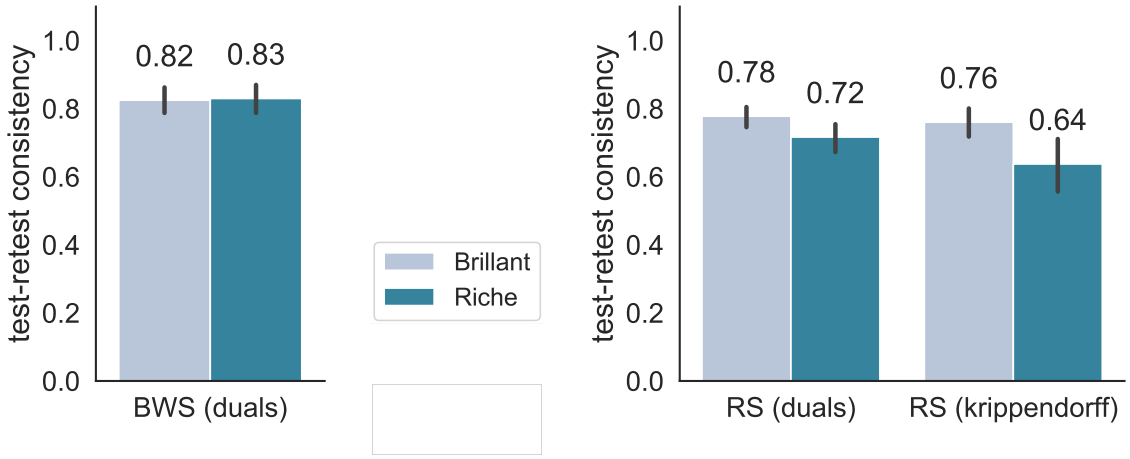


Figure 3.10: Which method was the most adapted to reflect your opinion ?



(a) BWS test-retests

(b) RS test-retests

Figure 3.11: Test-retest consistency measured with three metrics

3.4.4 Complexity handling

As explained previously in the protocol (see 3.2.4), we used 3 levers to differentiate the tasks on 'Riche' and 'Brillant' in terms of complexity: the definition of the attribute 'Brillant', the task ordering and a corpus conditioned by the Spectral Centroid.
Several metrics reveal that this levers managed to make the task on 'Riche' effectively more complex than the second task on 'Brillant'. First, an analysis of the variance (ANOVA) of the questionnaire revealed that participants significantly perceived the task on 'Riche' as longer and more difficult than the task on 'Brillant'. Secondly, an ANOVA of the tasks' durations show a main effect of the attribute ($p<.001$): participants spent more time on 'Riche' than on 'Brillant (see Fig. 3.9). Thirdly, the inter-annotator agreement metrics indicate that scores are less reliable on 'Riche' than on 'Brillant'. For both methods, the compliance to mean scores is significantly lower for 'Riche' than for 'Brillant'. The SHR is also significantly lower for 'Riche' in 100% of the iterations in RS, and in 83% of the iterations in BWS.

Although both methods reflect a gap of complexity, the differences of reliability across attributes suggest that BWS is more robust to the task's complexity than RS. The Figures 3.11a and 3.11b report the test-retest reliability measures of BWS and RS. In BWS, the test-retest compliance is not different between 'Riche' and 'Brillant'. On the contrary, for

RS, both test-retest reliability metrics show a significant drop of the consistency at the individual level.
This finding can be qualified by the fact that RS and BWS measure two different intra-participant consistencies. The BWS test-retest compliance evaluates the ability of a participant to stay consistent with himself in statements of the type 'sound A > sound B', whereas this is not a choice that participants explicitly make in RS. Looking at the evolution of the compliance to the mean scores (fig 3.8) from 'Brillant' to 'Riche', one can also see that the compliance of BWS decreases less than the compliance of RS, suggesting a better robustness of the BWS to complexity.

3.5 Critic of the protocol

3.5.1 Critic of the metrics

The reliability metrics used constitute the main limit of our protocol.
On the one hand, the process of Split-Half reliability requires to compute accurate scores with only half of the participants, which is likely to disadvantage BWS over RS according to the following observations:

- In simulations, the SHR is lower in BWS than in RS for a same simulated inter-participant noise (see the simulation 5 in Annex 2). This difference reduces as the number of participants increases.
- In this experiment, BWS half-scores are less accurate than the RS half-scores, since the RS scores at 10 participants are correlated on average at r = 0.99 with the scores at 20 participants, whereas this correlation is r = 0.96 in BWS. The difference between the two is significant (Z = 3.2[4]). The evolution of the validity (Fig. 3.7) also showed that RS scores converged sooner than BWS.
- In our experiment as in the study of De Bruyne et al. (2021), the measure of SHR disadvantaged BWS over RS when compared to a second metric of reliability (compliance to the mean scores in this study, Krippendorff's alpha in the work of De Bruyne et al. (2021)).

On the other hand, the compliance to the mean-scores allowed us to compare both methods on equal terms, but it does not assess the absolute reliability of the results since it converges as the scores of the group converge. Thus, adding new participants can increase the reliability of results without increasing the compliance to the mean scores. It has already been used (Hollis, 2018a) to spot outliers and discard non-compliant participants, but was never used to rate the consistency of a method before, and there is a global lack of knowledge concerning this metric.

We are also limited in our interpretation of the test-retest reliability metrics. To assess the impact of the complexity, we compared the evolution of RS metrics with the evolution

[4]Significance tested with the comparison of two independent samples (Annex 1).

of a BWS metric, assuming that they all increased monotonously with the intra-participant consistency. To address this issue, we verified the evolution of the RS and BWS metrics in the simulation,(see Annex 2), however the simulations are limited by the modelling of the intra-participant noise and don't perfectly reflect the reality.

3.5.2 Critic of the experimental design

The lack of proper reliability metrics for BWS results are a direct consequence our experimental design, in which participants had no pair of items in common. This aspect of the design, meant to maximise the collected information, jeopardises the measure of reliability and it thus questionable.

Low number of participants

When comparing RS and BWS, some trends couldn't be considered as significant, possibly due to too few participants. Some doubts remain as to the compliance to the mean scores on 'Riche', and to the questionnaire's answers to pleasantness, difficulty and estimated duration. It is uncertain whether they would be dispelled with more participants or not. In addition, we noted that the BWS validity on 'Brillant' tends to outperform the RS validity as the number of participants reaches 20 participants (see Fig.3.7). We could therefore expect this tendency to significantly grow with more participants and to have different asymptotic validity for RS and BWS.

Improved versions of the methods

Participants might have been biased by the novelty of the BWS, particularly when answering about the pleasantness and the difficulty of the methods.
Furthermore, we did not investigate the possible axes of improvement for the RS and the BWS. For instance, Greenleaf (1992) aims to improve the reliability of rating scales by modelling RS biases and correcting the distribution of each participant's rates. The collected results can also be more consistent when the choice between two items is not forced in paired comparison'(Parizet et al., 2005) or when participants can answer 'I don't know' instead of giving a rate (Roster et al., 2015) on rating scales. For the sake of simplicity and to collect complete results in both methods, we choose to force participants to answer to all items or trials, but it is likely that the designs of the two methods can be improved.

Chapter 4

Conclusion

We tested the Rating Scale and the Best-Worst Scaling with two timbre attributes at different levels of complexity, "Brillant" and "Riche". The two methods provided very similar scores and are both relevant for timbre annotation. They also presented a similar validity, reliability and ergonomy, with a small efficiency advantage for the BWS as it was a tad faster. Regarding the scores obtained with both methods, BWS proved to be a valid alternative to RS for sound annotation.
We also observed qualitative differences between the two methods, notably regarding the convergence of the scores and the confusion with pitch. They stressed that RS and BWS involve different judgement mechanisms and thus don't carry the exact same information. Eventually, we put in evidence the influence of two key factors on the performances, which are the task's complexity and the number of participants. A complex task lowers the reliability of both methods, but BWS was found to be more robust to complexity than RS. BWS hints at a higher asymptotic validity than RS as the number of participants increases, whereas RS scores proved to be more valid than BWS scores for a small number of participants. We left aside other possible facets of the comparison such as the ability of RS and BWS to cluster participant, or to provide interrelationships among the sounds. Yet, this would provide new insights on the nature and the extent of the differences between RS and BWS data.

Our experimental design allowed us to compare RS and BWS on a large dataset and at a similar cost of data collection. The challenge posed by this approach was the lack of reliability metrics applicable to RS and BWS results. Since the validity is not always measurable as we have seen for the attribute 'Riche', the measure of reliability really is a crucial issue. A solution would be to apply BWS in a configuration where the trials are not constrained by the Hollis design, thus making a compromise with the optimisation of the information. A design where participants share some pairs of items would allow to compare their answers on same duals, and would pave the way for a proper reliability metric in BWS which doesn't require the computation of individual scores and which can be compared to RS.

Our experimental comparison can be extended to different attributes and types of sounds. A corpus of stimulus can be conditioned in terms of size or variability, and all attributes do not have the same discrimination level or dichotomous/continuous aspect. The annotation task involves interactions between these qualities, and some combinations might suit better

to relative judgements like in BWS while other might suit better to absolute ratings. The mechanisms of interaction between the type of judgement and the nature of the attribute and the corpus are yet to be investigated.
Which method is the most valid depends very much on the number of participants considered, and it is a field that we wish to explore in further works by increasing the number of participants. The extension of the comparison to more items would also give a perspective on which method is the most robust to the fatigue bias.
Finally, a possible development line of this experiment is the investigation of improved version of RS and BWS, with no forced choice for example. The literature has investigated the biases of RS and possible improvements, but we have little knowledge about BWS biases.

BWS has been successfully applied to a sound corpus and could provide new insights compared to the Verbal Attribute Magnitude Estimation methods, with possibly different conclusions regarding the acoustical analysis of the scores as it was seen with the pitch in this experiment. We correlated the scores with the pitch and the Spectral Centroid, but many other audio features can be extracted to refine the definition of attributes. This experiment was designed for a methodological comparison and thus was performed by non-musicians, but one can also chose to collect BWS annotations from sound professional to have more consistent and relevant judgements that consider the multiple dimensions of the sounds' timbre.

An interesting application of the BWS is the partial annotation of datasets. If a corpus is too large to be seen entirely by a single participant, a solution is to give only a slice of the corpus to each participant, possibly with a crowd-sourcing approach (Yuen et al., 2011). In that case, relative judgement methods like BWS could be more suitable than absolute ratings like RS, where participants wouldn't calibrate and use their scale on the same parts of the corpus.
The rise of crowd-sourcing platforms eventually calls for online applications of the BWS on sounds, under less controlled conditions than in the laboratory as it has already be done in semantic works.

Annex 1 - Lexicon

‘

Reliability metrics

- **Pearson correlation:** measure of linear correlation between 2 vectors x and y of a same length n. It is used to assess the similarity between the values of x and y.

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \bar{y})^2}}$$

- **Spearman correlation**: measure of ordinal correlation between 2 vectors. It is used to assess the similarity of the rankings of x and y.

$$\rho = 1 - \frac{6 \sum d_i^2}{n(n^2 - 1)}$$

d = the pairwise distances of the ranks of the variables xi and yi .
n = the number of samples.

- **Cronbach's alpha**: reliability metric assessing the internal consistency of a test, that is to say the extend to which the questions are consistent with themselves and study a same construct.

$$\alpha_{cronbach} = \frac{k}{k-1}\left(1 - \frac{\sum_{i=1}^{k}\sigma_{x_i}^2}{\sigma_X^2}\right)$$

x_i = answer to the item i
$X = x_1 + x_2 + .. + x_k$ = sum of all items in a test of k items

Cronbach's alpha can be altered to assess the inter-annotator agreement instead of the internal consistency. For this purpose, the formula above is applied with items regarded as annotators and annotators playing the role of items (Burton et al., 2019).

- **Krippendorff's alpha**: reliability metric assessing the inter-annotator agreement. One interpretation of Krippendorff's alpha is:

$$\alpha_{krip} = 1 - \frac{D_{observed}}{D_{expected}}$$

$D_{observed}$ = the observed disagreement
$D_{expected}$ = the disagreement expected by chance

Krippendorff's alpha can be used whenever some annotators each assign one value to one item. Krippendorff's alpha can be applied to any number of items, to incomplete (missing) data, to any number of values available for rating an item, to binary, nominal,

ordinal, interval and ratio metrics. Therefore, it allows to compare the reliability of results obtained with different numbers of annotators and items, different metrics, and unequal sample sizes. The computed reliability coefficient takes into account the probability of an agreement by chance. Hayes and Krippendorff (2007) detail the computation of Krippendorff's alpha and argue why it should be used as the standard reliability measure.

- **Split-Half reliability (SHR)**: reliability metric originally used to measure the internal consistency of a test by splitting the items from the measurement procedure in half, and then calculating the scores for each half separately. The SHR coefficient is the Spearman's or Pearson's correlation coefficient between the two halves.

  SHR can also assess the inter-annotator agreement by splitting the participants in half instead of splitting the items. The SHR value is the correlation between the two scores of the two subgroups of participants.

  In order to have a more reliable SHR coefficient, the process is generally repeated over a few iterations (typically 100). A new combination of subgroups is generated at each iteration, and the final SHR value is the average of the correlation coefficients over the iterations.

$$SHR = \frac{\sum_{i=1}^{i=N} r_{i(S1,S2)}}{N}$$

  N = number of iteration
  S1, S2 = scores computed from group 1 and group 2

Significance tools

- **Student's t-test**: statistical test of a null hypothesis, where the test statistic follows a Student t-distribution. It is commonly used to test if the means of two sets of data are significantly different from each other. For two sets of data, the t-value is:

$$t = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\frac{\sigma_1}{n1}^2 + \frac{\sigma_2}{n2}^2}}$$

  Once the t-value has been determined, a p-value can be found using a table of values from the Student t-distribution. If the calculated p-value is less than the chosen threshold for statistical significance (usually 0.05), the null hypothesis is rejected in favour of the alternative hypothesis.
  To perform a test with three or more means, an analysis of variance must be used.

- **Fisher Z-Transformation**: transformation of the sampling distribution of Pearson's r (i.e. the correlation coefficient) so that it becomes normally distributed.

$$z = \frac{1}{2}\ln(\frac{1+r}{1-r})$$

- **Comparison of correlations from independent samples**
  This test evaluates the significance of the difference between two independent correlations. With A,B,C and D being independent vectors, the null hypothesis is:

$$Ho : r_1^2(A, C) = r_2^2(B, D)$$

  The first step is to apply Fisher Z-transformation to both correlation coefficients.

  Then, one introduces the following statistic:

$$D = z_1 - z_2$$

  D asymptotically follows a law of parameters

$$\mu = 0$$

$$\sigma = \sqrt{\frac{1}{n_1 - 3} + \frac{1}{n_2 - 3}}$$

  If the observed value of D falls inside the critical region, then Ho is rejected at the chosen significance level.

  Typically, if $Z = |\frac{D_{obs}}{\sigma}| \geq 1.96$, then r_1 and r_2 are significantly different at the confidence level $\alpha = 0.05$.

  The mathematical demonstration is explained by Rakotomalala (2015), and several online calculators exist, such as Lenhard (2014). This significance test is also commonly applied to Spearman's coefficients when the data are not normal, as it is recommended by Myers and Sirois (2004).

- **Comparison of correlations from dependent samples**
  This test evaluates the significance of the difference between two dependant correlations sharing one variable. In other words, it evaluates whether two independent vectors A and B are equally correlated to a reference vector R. The null hypothesis is:

$$Ho : r_1^2(A, R) = r_2^2(B, R)$$

  The value Z is obtained with the equations (3), (10) and (4) of Steiger (1980), in "Case A". Similarly to the other tests, the null hypothesis is rejected if the observed value of D falls inside the critical region, that is to say if $Z \geq 1.96$.

  For this study, we built a python script taking up Steiger's equations. The obtained Z values were equal to those obtained with other online calculators, built by Lenhard (2014) or Lee (2013).

Annex 2 - Simulations

We ran several simulations preliminary to the experiment in order to better understand the functioning of RS ans BWS, and to observe the behaviour of our validity and reliability metrics.

Simulation framework

For each simulation, we set up the number of items, the number of participants, and a random vector of true values. The true values are ranged between 0 and 1, and follow a continuous uniform distribution. Participant's latent values are generated from the true values with specific noise conditions. The BWS trials are generated according to Hollis design, as in the real experiment. Then, the responses of each participant to the 9-point Likert scale and to the BWS trials are simulated, according to the participants' latent values.

Inter-participant noise implementation

The inter-participant noise is implemented by applying a noise to the vector of fictive true values V_{true}. For each virtual participant, a random vector of values V_{rand} is generated, and the participant's latent values V_{part} is computed as:

$$V_{part} = (1 - x)V_{true} + xV_{rand}$$

- x is the amplitude of inter-participant noise, between 0 and 1

This technique allows to control the noise x like a slider, with x=1 corresponding to a fully random behaviour, and x = 0 corresponding to a perfect participant, fully in agreement with true values. Figure A2.1 shows a participant's values with different levels of inter-participant noise x.

Intra-participant noise implementation

The intra-participant noise was generated by applying a Gaussian noise to the participant's latent values, all along his test.

$$V_{part,k} = expit(logit(V_{true}) + xg(k))$$

- x is the amplitude of the intra-participant noise
- $g(k)$ is the gaussian noise's value at the question k

We apply the expit and the logit function so that the addition of the noise occurs in the interval]-inf, +inf[. In that way, the latent values stay bounded between 0 and 1. With a gaussian noise of zero amplitude, the participant is perfectly consistent with himself accross time, and always answer the same way to a trial.

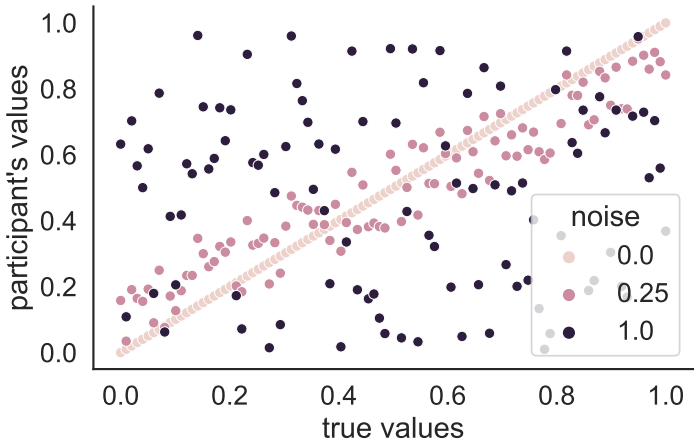


Figure A2.1: Implementation of inter-participant noise in the latent values of a participant.

RS biases implementation

In addition to the two generic noises presented above, we implemented 3 different biases related to RS. The first bias is the behaviour that consists of avoiding the extreme values of a scale. The second bias is the positive or the negative bias that a participant can have on a scale. Those biases are implemented by applying a filter to the participant's latent values, as represented on the figure A2.2. The third bias is the positive or negative bias changing over time, which is an other type of intra-participant noise. For example, a participant is likely to start giving high rates at the beginning of the experiment, and then to get more severe along the test because new items made him re-calibrate his rating scale.

We implemented those response styles because previous studies (Soutar et al. (2015), Baumgartner and Steenkamp (2001)) show that they are likely to be adopted by participants in judgement tasks. However, we don't aim to model the true decision process of participants or to model reality, but rather to check if those biases can indeed be responsible for a decrease of RS performances.

Simulation 1: Number of annotation needed

Figure A2.3 plots the distribution of RS and BWS scores in Hollis design with N = 100 items.

With one participant, each item is seen once in both methods which corresponds to N/4 BWS annotations and N RS annotations. In RS, the individual scores are distributed over the nine levels of the scale and fit approximately to the true values, but in BWS, the individual scores of one participant are distributed in only three levels, corresponding to 'best', 'worst' and unchosen. They are not usable, and this is why individual scores are not available for BWS in Hollis design.

With eight participants, we added a small inter-participant noise to the simulation. Without this noise, each virtual participant would rate each item exactly the same way in RS, and

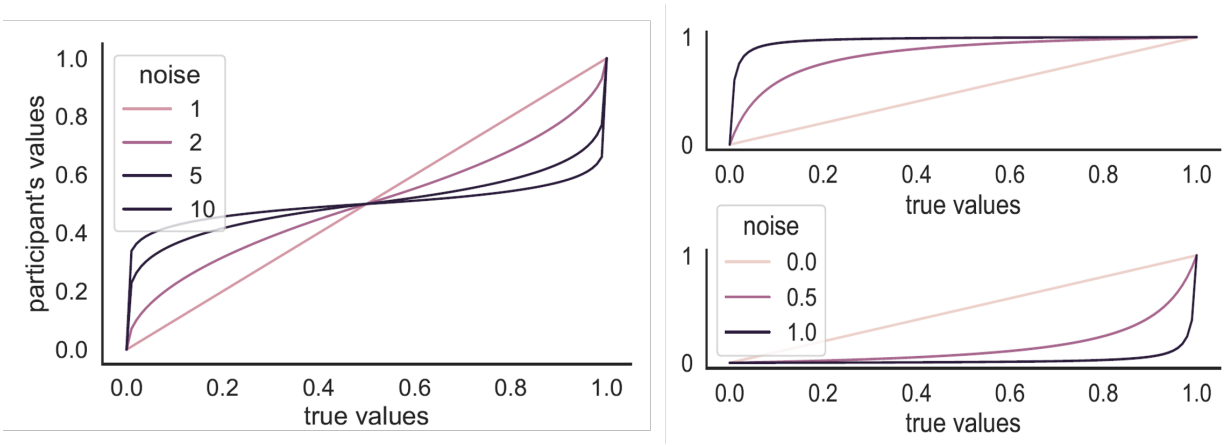


Figure A2.2: Implementation of RS biases in the latent values of a participant. Avoid-extremes bias (left) and positive and negative bias (right).

the average scores would be identical to the ones obtained with 1 participant, that is to say discretised into nine levels. A small inter-annotator disagreement is actually what makes the RS scores accurate. The BWS scores are now computed with 2N annotations, and are more accurate than with one participant (giving N/4 annotations) because more distinct 4-tuples have been answered, providing new duals to the scoring algorithm.

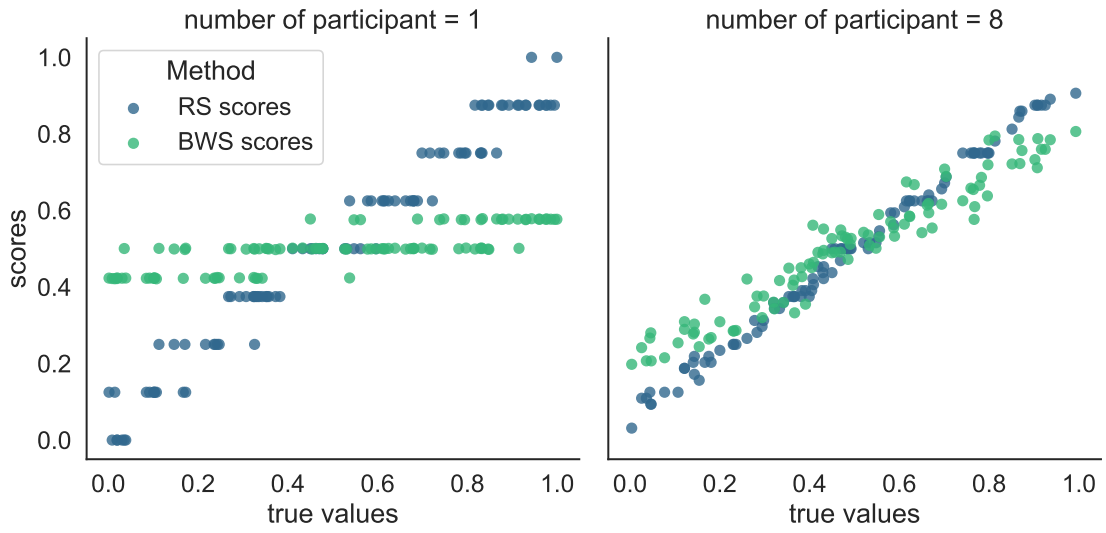


Figure A2.3: Simulation of RS and BWS scores in Hollis design. Individual scores (left) are usable in RS but not in BWS.

Simulation 2: Validity

Simulations were conducted by varying the number of participants, the number of items and the noise conditions. Validity was measured using Pearson's correlation with the true values.

Therefore, the validity is about 1 with perfect participant, and about 0 with dummy participants giving random answers. The simulations of the validity led to the following conclusions:

Validity without RS biases

- Above 20 items, the number of items doesn't influence the validity of RS or BWS.
- The validity of RS and BWS increases with the number of participants.
- Applying only inter and intra-participant noise, RS results are more valid than BWS results, no matter the amount of participants.
- Both methods give similar scores and similar validity as the number of participants increases.

Validity with RS biases

For these simulations, RS biases were added to the participant's behaviours.

- RS validity decreases as the amplitude of RS biases increases, whereas BWS validity isn't impacted.
- RS becomes less valid than BWS as RS biases are amplified.

These results indicate that these biases can possibly be responsible for a decrease of RS validity in real life. They support the idea that the BWS is a relevant alternative to avoid the response styles of RS, since BWS validity does not seem to be affected by them.

The following simulations aim to explore the behaviour of three reliability metrics: the test-retest compliance, the compliance to mean-scores and the SHR.

Simulation 3: Test-retest compliance

To measure the intra-participant consistency with the same type of measurement for RS and BWS, we made up the test-retest compliance measure. It is the proportion of duals answered similarly by a participant in tests and in retests trials. In BWS, we counted the 5 duals inferred by each trial of test and retest. In RS, the duals that we considered are all the possible pairs among the 20 sounds of retests.

To check the behaviour of this measure in RS and in BWS, we plotted the test-retest compliance with different intra-participant noise values on figure A2.4. Each point corresponds to the average of 5 simulations. Each simulations takes the average of the test-retest compliance of 20 participants, as in the experiment.

One can see that the two methods behave similarly, and that their consistency both decreases approximately linearly with the amount of intra-participant noise. However, the measure clearly advantages the BWS. Therefore, we can't use the test-retest compliance to compare directly RS and BWS intra-participant consistency. We only use it to compare how each method's consistency evolves from one attribute to the other.

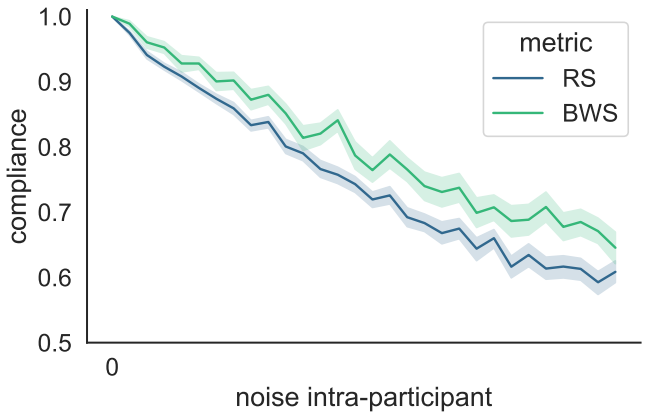


Figure A2.4: Simulation of test-retest compliance. The test-retest compliance is the proportion of respected duals between tests and retests' trials in BWS, and between the test and retests rankings in RS.

Simulation 4: Compliance to mean scores

To measure the inter-participant consistency with the same type of measurement for RS and BWS, we used the compliance to mean scores, already used by Hollis (2018a) to spot outliers. This is the proportion of duels to which a participant responded in a manner consistent with the average scores of the group. In BWS, the 5 duals inferred by each trial are compared to the global ranking. In RS, we re-simulated the BWS duals of the participant with his ratings.

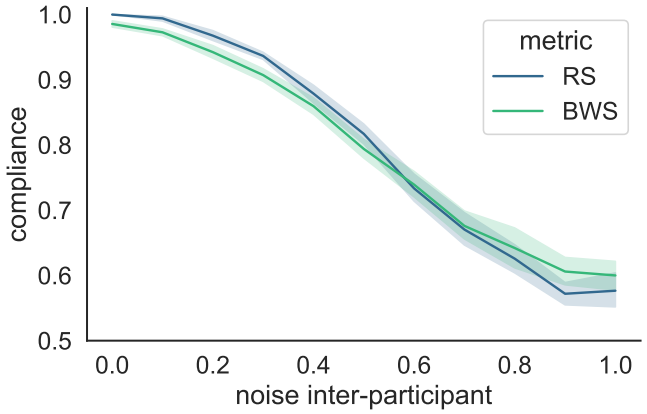


Figure A2.5: Simulation of the compliance to mean scores

The simulation is conducted with 20 participants and 100 items. The intra-participant noise varies from 0 to 1 (that is to say from perfect agreement to a random behaviour). The result is plotted figure A2.5, where each point corresponds to one simulation with 20 participants and 100 items.

It can be seen than both methods have generally very similar compliance scores. However, RS is significantly advantaged when compliance scores are above 0.90, with low inter-participant noise.

Lower bound explanation: One could expect that with fully random players, the corresponding compliance to mean scores would be 0.5 since the probability to agree on a dual one out of two. Yet, the lower bound of the compliance simulated with 20 participants is approximately 0.6, and not 0.5. This is caused by the fact that mean scores take into account all participants, including the participant of interest. Therefore, the mean scores and the participant's judgment share a bit of the same information. Another version of this metric where the mean scores are calculated on all participants but the participant of interest is possible, and it converges to 0.50 when the noise is maximum. However, differences between RS and BWS are a bit heightened with the latter version, and this is why we kept the first version where the mean scores include all participants.

Simulation 5: Split-Half Reliability (SHR)

The second way to measure the inter-participant consistency with the same type of measurement for RS and BWS was to use the Split-Half reliability, as it was done by Kiritchenko and Mohammad (2017a). It consists of splitting the participants in two halves, and computing the correlation between the scores of the two halves.

We simulated the SHR of 20 participants judging 100 items, with an inter-participant noise varying from 0 to 1 (perfect agreement to to fully random behaviours).
We first plotted the SHR of RS and BWS with different trials, as in our experiments. This means that participants in the first halves answered to different 4-tuples than participants of the other half. On figure A2.6, one can see that the two SHR are significantly different, with differences up to $\Delta = 0.12$. We aim to find why the SHR of BWS was lower than the SHR of RS for a same inter-participant noise.

Same or different trials - A first explanation was that the SHR of BWS was lower because the 4-tuples were different in each half. Without any noise, that is to say when all participant perfectly agree, the SHR of BWS is around 0.95 and not 1. It is because the two halves don't answer to the same BWS trials, and therefore the two scores don't contain exactly the same type of information. We tested this hypothesis by simulating a third method: *BWS - same trials*. It corresponds to the SHR between two groups of participants that answered to the same trials. Within each group, participants have different series of trials, but each participant of group A has a twin participant in the group B who answered the same trials than him. On fig.A2.6, one can see that without noise, the BWS with same trials has an SHR equal to 1, as RS. However, as the noise increases, the difference with BWS with different trials is reduced, and both BWS designs have an equivalent SHR. The gap with the SHR of RS has therefore an other cause than the fact that BWS trials are different in the two halves.

Number of annotations - Most likely, by splitting a group in two, BWS scores lose

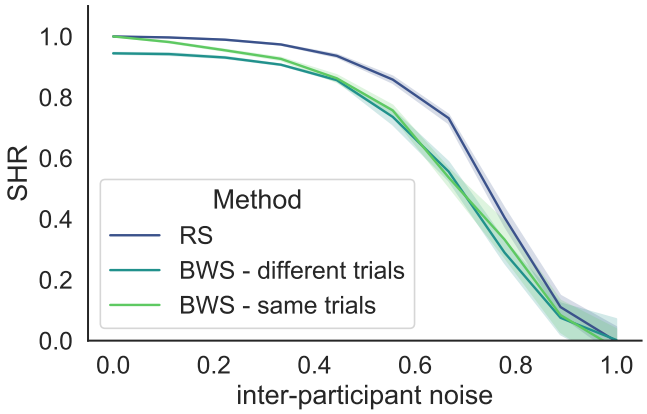


Figure A2.6: Simulation of SHR with 20 participants

much more precision than RS scores. Both the simulation and the experiment show that RS scores converge faster than BWS scores, as the scores at 10 and 20 participants are more correlated in RS than in BWS. With more participants, both BWS and RS scores are likely to be very close to their asymptotic values with half the annotations.

To test this hypothesis, we plotted the SHR of RS and BWS with 100 participants on Figure A2.7. The difference of SHR between RS and BWS is a lot smaller than with 20 participants, however it is still significant for inter-participant values around 0.6.

This differences between RS and BWS in the simulation can reach 0.12 with 20 participants, and should be taken into account when comparing the SHR of real data. Nevertheless, these simulations are limited by the fact that they don't model reality, and the inter-participant noise for instance is likely to correspond to a much more complex model than the uniform noise adopted here.

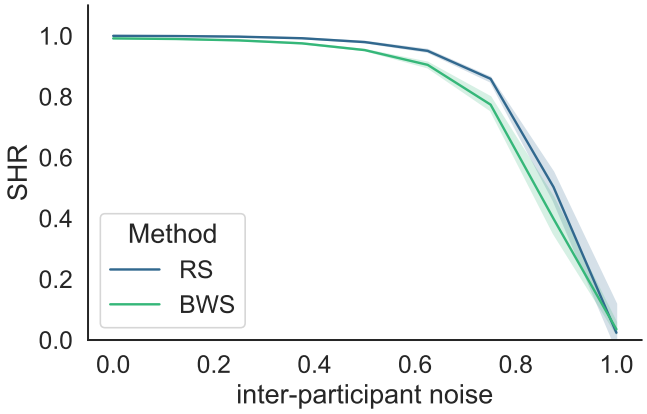


Figure A2.7: Simulation of SHR with 100 participants

Annex 3 - Instruments ranking

Below are the results of the instruments' timbres judgements regarding the richness ('Riche') and the brightness ('Brillant'). Rankings were obtained by averaging the scores of RS and BWS. These rankings should be considered with caution since all instruments weren't represented with the same octaves.

Richness	**Brightness**	**Rank**
doublebass	trumpet harmon	1
cello	violin	2
alto	trumpet	3
violin	trumpet cup	4
french horn	oboe	5
bassoon	flute	6
trombone	alto	7
bass tuba	cello	8
trombone harmon	accordion	9
alto saxophone	alto saxophone	10
clarinet	clarinet	11
oboe	trombone	12
trumpet	doublebass	13
accordion	bassoon	14
trumpet harmon	french horn	15
flute	bass tuba	16
trumpet cup	trombone harmon	17

Table A3.1: Ranking of the instruments according to participants

Richness

According to the ranking, string instruments offer the richest sounds. In the questionnaire, many participants associated 'richness' with 'vibration', 'variation' or 'modulation', which presumably refer to the vibrato of the string sounds presented. Other participants also defined a rich sound as a sound having 'many harmonics','not a single frequence', 'containing several notes', thus showing a spectral approach in the perception of timbre. Richness was eventually associated with pleasantness and beauty, and with complexity and depth.

Brightness

There are 3 types of trumpets in the top 4 of the brightest sounds. The less bright sounds are produced by larger brass instruments. In the questionnaire, participants mainly associated 'bright' with 'high-pitched'. They also associated brightness with light and smoothness.

Corpus and pitch

The sound corpus extended over 7 different octaves of Cs, and the brightness was thus strongly correlated with the pitch ($r = 0.77$). Participants reported that they often used their perception of the pitch to judge the brightness of the sounds. Therefore, the final ranking of the instruments above is strongly influenced by the octaves in which those instruments were present.

Annex 4 - RS responses styles

We have seen that the main drawback of the rating scale is that participants may answer to the task with different response styles, affecting the quality of the results. In the experiment, we could observe several such response styles among the participants. We extracted the responses of 4 participants for Rich and Brilliant, which each illustrate a different response style.

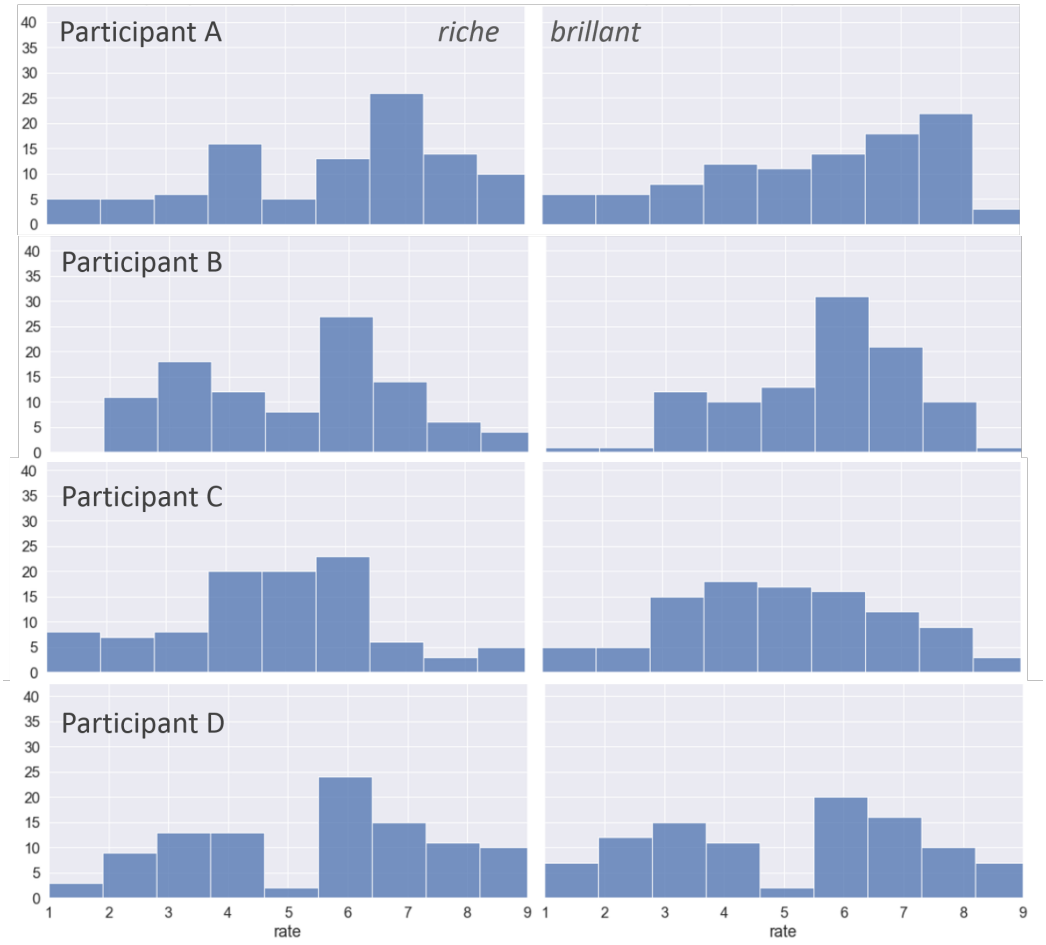


Figure A4.1: Rating scale responses styles

Participant A has a **positive response bias**: one can see that his answers are concentrated on the right part of the scale. However, this bias isn't present at the group-level. It is more commonly found when the attribute has a strong positive or negative valence, such as "important", or "pleasant", whereas 'Brillant" (birght) or 'Riche' (rich) are quite neutral attributes.

Participant B avoids to rate with **extreme values**. He almost never gave a 1 or a 9 to an item. Asking for the participants feedback in the questionnaire, some participants said that they had felt like they weren't using the extreme values of the scale..

Participant C answers mainly with the **middle of the scale**. This is likely to provide results with low discrimination between items.

Participant D, on the contrary, answers avoiding the middle value, in a more **categorical** way than the others. This shows that some participants can perceive a latent dimensions as a dichotomous variable while others can perceive it as a continuous variable.

Annex 5 - Instructions

Before starting the experiment, the participants were watching a 2 minutes video explaining the interface, and giving the following instructions:

> " Hello, and welcome to this experiment on timbre. This experiment consists of judging instrumental sounds using 2 methods, the Rating Scale and the Best-worst scaling.
>
> Let's introduce the two methods using the attribute 'high-pitched' as an example. In the Rating Scale method, the sounds are presented to the participant one by one. Click on the Play button to play the sound. Then assign a score on this 9-point scale, depending on whether the sound is high or low.
>
> Try to answer as spontaneously and intuitively as possible. It is recommended that you do not listen to more than 3 times per sound. If you have listened to a sound more than 3 times, a red light will come on to indicate that it is time to move on to the next sound. Answer as best you can, and move on to the next sound.
>
> The bar at the bottom of the screen shows you how far you have progressed in the test.
>
> Let's now introduce the second method, best-worst scaling. This time, the sounds are presented 4 by 4. Click on a sound to play it. The blue dots show you the sounds you have left to listen to. Next, indicate which sound seems to be the highest and which seems to be the lowest. Click on continue to move on to the next test.
>
> Again, try to answer spontaneously and intuitively. Some sounds may be very similar, but it is recommended that you do not exceed 3 listenings per sound. When this quota is exceeded, a red signal indicates that it is time to move on. Then answer as best you can and move on to the next test.
>
> Finally, please leave your phone outside the booth.
>
> Thank you again for your participation Have a good experience!"

Annex 6 - Questionnaire

Participants were asked to complete the following questionnaire once they had judged 'Rich' with both methods, and again when they had judged 'Brillant'.

- In your opinion, how long (in minutes) did the RS method last ?
- In your opinion, how long (in minutes) did the BWS method last ?
- On what criteria did you judge the richness of the sounds? What was your strategy ?
- Rate the level of difficulty of the RS method. (7-point likert scale)
- Rate the level of difficulty of the BWS method. (7-point likert scale)
- To what extend was the RS method pleasant ? (7-point likert scale)
- To what extend was the BWS method pleasant ? (7-point likert scale)
- Which method do you think was most relevant to reflect your actual perception of the richness of the sounds?
- Free comments: challenges encountered, global perception of the experiment..

List of Figures

List of Tables

Bibliography

Allen, E. J. and Oxenham, A. J. (2014). Symmetric interactions and interference between pitch and timbre. *The Journal of the Acoustical Society of America*, 135(3):1371–1379.

Alluri, V. and Toiviainen, P. (2010). Exploring Perceptual and Acoustical Correlates of Polyphonic Timbre. *Music Perception*, 27(3):223–242.

Aruga, K., Bolt, T., and Pest, P. (2021). Energy policy trade-offs in poland: A best-worst scaling discrete choice experiment. *Energy Policy*, 156:112465.

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*.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., and Treiman, R. (2007). The english lexicon project. *Behavior research methods*, 39(3):445–459.

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

Burton, M., Fisher, C., Pea, P. G., Rhodes, G., and Ewing, L. (2021). Beyond likert ratings: Improving the robustness of developmental research measurement using best–worst scaling. *Behavior Research Methods*, pages 1–7.

Burton, Gillian, M., Rigby, D., Sutherland, C. A., and Rhodes (2019). Best-worst scaling improves measurement of first impressions. *Cognitive research: principles and implications*, 4(1):1–10.

Carron, M. (2017). Speaking about sounds: a tool for communication on sound features. *Journal of Design Research*, 15(2):85–109.

Cohen, E. and Goodman, S. (2009). Applying best-worst scaling to wine marketing. *International journal of wine business research*.

Darke, G. (2005). Assessment of timbre using verbal attributes. In *Conference on Interdisciplinary Musicology. Montreal, Quebec*. sn.

De Bruyne, L., De Clercq, O., and Hoste, V. (2021). Annotating affective dimensions in user-generated content. *Language Resources and Evaluation*, pages 1–29.

Disley, A. C. and Howard, D. M. (2004). Spectral correlates of timbral semantics relating to the pipe organ. *Speech, Music and Hearing*, 46:25–39.

Elo, A. E. (1978). *The rating of chessplayers, past and present*. Arco Pub.

Faure, A. (2000). *Des sons aux mots, comment parle-t-on du timbre musical?* PhD thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS).

Flynn, T. N., Louviere, J. J., Peters, T. J., and Coast, J. (2007). Best–worst scaling: What it can do for health care research and how to do it. *Journal of Health Economics*, 26(1):171–189.

Funke, F. (2016). A web experiment showing negative effects of slider scales compared to visual analogue scales and radio button scales. *Social Science Computer Review*, 34(2):244–254.

Funke, F. and Reips, U.-D. (2012). Why semantic differentials in web-based research should be made from visual analogue scales and not from 5-point scales. *Field Methods*, 24(3):310–327.

Greenleaf, E. A. (1992). Improving rating scale measures by detecting and correcting bias components in some response styles. *Journal of Marketing Research*, 29(2):176–188.

Grey, J. M. (1977). Multidimensional perceptual scaling of musical timbres. *the Journal of the Acoustical Society of America*, 61(5):1270–1277.

Grey, J. M. and Gordon, J. W. (1978). Perceptual effects of spectral modifications on musical timbres. *The Journal of the Acoustical Society of America*, 63(5):1493–1500.

Hayes, A. F. and Krippendorff, K. (2007). Answering the call for a standard reliability measure for coding data. *Communication Methods and Measures*, 1(1):77–89.

Hollis, Geoff et Westbury, C. (2018a). When is best-worst best? a comparison of best-worst scaling, numeric estimation, and rating scales for collection of semantic norms. *Behavior research methods*, 50(1):115–133.

Hollis, G. (2018b). Scoring best-worst data in unbalanced many-item designs, with applications to crowdsourcing semantic judgments. *Behavior research methods*, 50(2):711–729.

Kendall, R. A. and Carterette, E. C. (1993). Verbal Attributes of Simultaneous Wind Instrument Timbres: II. Adjectives Induced from Piston's "Orchestration". *Music Perception*, 10(4):469–501.

Keuleers, E., Lacey, P., Rastle, K., and Brysbaert, M. (2012). The british lexicon project: Lexical decision data for 28,730 monosyllabic and disyllabic english words. *Behavior research methods*, 44(1):287–304.

Kiritchenko, S. and Mohammad, S. M. (2017a). Best-worst scaling more reliable than rating scales: A case study on sentiment intensity annotation. *arXiv:1712.01765*.

Kiritchenko, S. and Mohammad, S. M. (2017b). Capturing reliable fine-grained sentiment associations by crowdsourcing and best-worst scaling. *arXiv:1712.01741v1 [cs.CL] 5 Dec 2017*.

Lee, K. J. (2013). Calculation for the test of the difference between two dependent correlations with one variable in common [computer software].

Lenhard, Wolfgang, A. (2014). Testing the significance of correlations.

Likert, R. (1932). A technique for the measurement of attitudes. *Archives of psychology*.

Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: Theory, methods and applications*. Cambridge University Press.

Louviere, J. J. and Woodworth, G. G. (1991). Best-worst scaling: A model for the largest difference judgments. Technical report, Working paper.

Masino, C. and Lam, T. C. (2014). Choice of rating scale labels: implication for minimizing patient satisfaction response ceiling effect in telemedicine surveys. *Telemedicine and e-Health*, 20(12):1150–1155.

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. *Psychological research*, 58(3):177–192.

McFee, B., Raffel, C., Liang, D., Ellis, D. P., McVicar, M., Battenberg, E., and Nieto, O. (2015). librosa: Audio and music signal analysis in python. In *Proceedings of the 14th python in science conference*, volume 8.

Michaud, P.-Y. (2012). *Distorsions des systèmes de reproduction musicale : Protocole de caractérisation perceptive*. Theses, Aix-Marseille Université.

Myers, L. and Sirois, M. J. (2004). Spearman correlation coefficients, differences between. *Encyclopedia of statistical sciences*, 12.

Osgood, C. E. (1964). Semantic differential technique in the comparative study of cultures 1. *American Anthropologist*, 66(3):171–200.

Parizet, E., Hamzaoui, N., and Sabatie, G. (2005). Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica*, 91(2):356–364.

Paulhus, D. L. (1991). Measurement and control of response bias. *Measures of personality and social psychological attitudes (pp. 17–59)*.

Pearce, A., Brookes, T., and Mason, R. (2019). Modelling timbral hardness. *Applied Sciences*, 9:466.

Rakotomalala, R. (2015). Analyse de corrélation. *Cours statistique à l'université de lumière Lyon*, 2:89.

Rescorla, R. A. (1972). A theory of pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. *Current research and theory*, pages 64–99.

Roster, C. A., Lucianetti, L., and Albaum, G. (2015). Exploring slider vs. categorical response formats in web-based surveys. *Journal of Research Practice*, 11(1):D1–D1.

Schaik, P. and Ling, J. (2007). Design parameters of rating scales for web sites. *ACM Trans. Comput.-Hum. Interact.*, 14.

Schubert, E. and Wolfe, J. (2006). Does timbral brightness scale with frequency and spectral centroid? *Acta acustica united with acustica*, 92(5):820–825.

Schubert, E., Wolfe, J., Tarnopolsky, A., et al. (2004). Spectral centroid and timbre in complex, multiple instrumental textures. In *Proceedings of the international conference on music perception and cognition, North Western University, Illinois*, pages 112–116. sn.

Schuman, H. and Presser, S. (1996). *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage.

Soutar, G. N., Sweeney, J. C., McColl-Kennedy, J. R., Louviere, J. J., Flynn, T. N., and Marley, A. A. J. (2015). *Best-worst scaling: an alternative to ratings data*, page 177–188. Cambridge University Press.

Steiger, J. H. (1980). Tests for comparing elements of a correlation matrix. *Psychological bulletin*, 87(2):245.

Sullivan, G. M. and Artino, Anthony R., J. (2013). Analyzing and Interpreting Data From Likert-Type Scales. *Journal of Graduate Medical Education*, 5(4):541–542.

von Bismarck, G. (1974). Timbre of steady sounds: A factorial investigation of its verbal attributes. *Acta Acustica united with Acustica*, 30(3):146–159.

Voutilainen, A., Pitkäaho, T., Kvist, T., and Vehviläinen-Julkunen, K. (2016). How to ask about patient satisfaction? the visual analogue scale is less vulnerable to confounding factors and ceiling effect than a symmetric likert scale. *Journal of advanced nursing*, 72(4):946–957.

Wallmark, Z. (2019). Semantic crosstalk in timbre perception. *Music & Science*, 2:2059204319846617.

Wood, I., McCrae, J. P., Andryushechkin, V., and Buitelaar, P. (2018). A comparison of emotion annotation schemes and a new annotated data set. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Yuen, M.-C., King, I., and Leung, K. (2011). A survey of crowdsourcing systems. pages 766–773.

Zacharakis, A., Pastiadis, K., and Reiss, J. D. (2012). An interlanguage study of musical timbre semantic dimensions and their acoustic correlates. *Music Perception: An Interdisciplinary Journal*, 31(4):339–358.