Comparison of Perceived and Imagined Instrumental Blend

Linglan Zhu

Master of Arts

Music Technology Area
Department of Music Research
Schulich School of Music
McGill University
Montreal, Canada

April 2022

A thesis submitted to McGill University in partial fulfillment of the requirements for the degree of Master of Arts

© 2022 Linglan Zhu

Abstract

Timbral blend is a fundamental aspect of various musical activities for shaping sounds and musical intentions, most prominently in composition and performance. It also underpins how people listen to and understand music. Although blend is most intuitively rendered with sounds through "external" hearing, the notion of "internal" hearing provides an alternative but still musically meaningful angle to approach blend. Previous studies have mostly focused on timbral instrumental blend of heard sounds, with little evidence on how imagined blend mediated by mental images functions in comparison with "externally" heard blend. Given the implicit role of imagining blends in different musical practices, another question that can be asked is whether musical background influences the properties of imagined blend. To investigate these questions, two groups of participants (musicians and non-musicians; 32 per group) were presented with pairs of short instrumental sounds in unison from 14 different instruments in two different conditions. In the first condition, individual instruments were played sequentially, and participants were instructed to imagine them being played simultaneously and rate their degree of blend. In the next condition, pairs of instruments were played simultaneously, and participants were asked to rate the perceived degree of blend. Results showed significant interaction effects between the type of instrument pairs and the two presentation conditions, and among instrument pairs, presentation conditions, and musical backgrounds. Similar effects were also observed after different instrument pairs were agglomerated into different instrumental family pairs, suggesting both specific and general instrumental contingency for the quality of imagined blend. Acoustic analyses were conducted on the sound stimuli and were used in modeling blends in the two conditions. Results suggested that while certain acoustic factors function consistently both in heard and imagined blend, some acoustic features contribute differently to the two types of blends, a result confirmed by two "blend spaces" generated with multidimensional scaling. Overall, it appears that the perception of heard and imagined blends draws on potentially different abstractions of certain acoustic features. In practice, how these two types of blends might differ is a result of complex interactions involving instrumental choices and listeners' musical backgrounds.

Résumé

Le mélange timbral est un aspect fondamental de diverses activités musicales pour façonner les sons et les intentions musicales, notamment dans la composition et l'interprétation. Il sous-tend également la manière dont les gens écoutent et comprennent la musique. Bien que le mélange soit le plus intuitivement rendu avec des sons par l'audition « externe », la notion d'audition « interne » fournit un angle alternatif mais toujours musicalement significatif pour aborder le mélange. Les études antérieures se sont principalement concentrées sur le mélange instrumental timbral des sons entendus, avec peu de preuves sur la façon dont le mélange imaginé médié par des images mentales fonctionne par rapport au mélange entendu « extérieurement ». Étant donné le rôle implicite de l'imagination des mélanges dans différentes pratiques musicales, une autre question qui peut être posée est de savoir si le contexte musical influence les propriétés du mélange imaginé. Pour étudier ces questions, deux groupes de participants (musiciens et non-musiciens ; 32 par groupe) ont reçu des paires de sons instrumentaux courts à l'unisson provenant de 14 instruments différents dans deux conditions différentes. Dans la première, les instruments individuels étaient joués séquentiellement, et les participants devaient imaginer qu'ils étaient joués simultanément et évaluer leur degré de mélange. Dans la deuxième, des paires d'instruments ont été jouées simultanément, et les participants devaient évaluer le degré de mélange perçu. Les résultats ont montré des effets d'interaction significatifs entre le type de paires d'instruments et les deux conditions de présentation, et entre les paires, les conditions et la formation musicale. Des effets similaires ont été observés après que les paires instrumentales aient été agglomérées en paires de familles instrumentales, suggérant une contingence instrumentale à la fois spécifique et générale pour la qualité du mélange imaginé. Des analyses acoustiques ont été menées sur les stimuli sonores et ont été utilisées pour modéliser les mélanges dans les deux conditions. Les résultats suggèrent que si certains facteurs acoustiques fonctionnent de manière cohérente dans les mélanges entendus et imaginés, certaines caractéristiques acoustiques contribuent différemment aux deux types de mélanges, un résultat confirmé par deux « espaces de mélange » générés par une mise à l'échelle multidimensionnelle. Dans l'ensemble, la perception des mélanges entendus et imaginés semble se reposer sur des abstractions potentiellement différentes de certaines caractéristiques acoustiques. En pratique, la façon dont ces deux types de mélanges peuvent différer est le résultat d'interactions complexes impliquant les choix instrumentaux et la formation musicale des auditeurs.

Acknowledgments

Time has flown by since I began the new journey in music technology two years ago. The COVID pandemic has shadowed this journey since the very beginning and brought unexpected challenges. Luckily, with the wonderful people in MPCL I am able to finally put a final period on my master's thesis, getting ready for the new challenges in the forthcoming PhD studies. Firstly, I would like to thank my supervisor, Professor Stephen McAdams, for recognizing my passion for music research and welcoming me to the wonderful MPCL at the very start (while I was still a chemistry graduate with no academic background in music whatsoever!). I am grateful for his guidance and inspiration throughout my studying at McGill, and I am looking forward to continuing our research for my PhD. I sincerely thank our wonderful technical manager Bennett Smith for programming my experiments with all the fussy details I asked for every now and then. I was always amazed by the prompt and precise updates he made to my experiments following our email correspondence. Our statistical consultant Marcel Montrey also offered valuable input into clarifying some of the problems I encountered in data analyses. The wonderful MPCLers: Yuval Adler, Marcelo Caetano, Lena Heng, Erica Huynh, Iza Korsmit, Behrad Madahi, Lindsey Reymore, Joshua Rosner, Jade Roth, Kit Soden, and Matthew Zeller, have always been inspiring me with their wisdom and passion for music. I also thank them for their feedback on my thesis research and piloting my experiments. I would like to especially thank Lindsey Reymore for giving me (a first-time experimenter) hands-on guidance on how to run experiments with human participants. My lovely composer friends: Benjamin Lavastre, Omer Barash, Francis Battah, and Theo Haber, for our endlessly exciting discussion on great composers and music. Lastly, to my wonderful parents: without their unconditional support I wouldn't be able to embark on this great adventure of music.

Author Contributions

Under the supervision of Professor Stephen McAdams, I was responsible for coming up with the research questions, designing and running the experiments, and analyzing and interpreting the data. Bennett Smith programmed the experiments and helped me prepare them.

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

The perceptual complexity of timbre has been well described in both the existing music-theoretical and psychophysical literature. A recent review by Siedenburg & McAdams (2017a) listed four conceptual distinctions of timbre which clarified some of the foundations behind timbre. The last one among them, which states that “Timbre is property of fused auditory events”, speaks to the interactive potential of different sonic events and the operational aspect of creating timbre from different constituents. Auditory scene analysis (Bregman, 1990) elucidates how complex auditory information is organized on a perceptual level and contributes to different perceived sound sources. On a higher level, an analogy can be made concerning how distinct sound sources can fuse into a single perceptual unity or remain separated. In the case of fusion, the resulting composite sound creates the illusion of a “virtual source image” (McAdams, 1984b). Fused concurrent sounds can have a range of distinct timbral characteristics (Sandell, 1995), which offer great potentials to orchestration and composition in the form of different instrumental combinations. Unsurprisingly, orchestration treaties have given great attention to the choices of concurrent combinations of different instruments for achieving different sonic intentions (e.g., Adler, 2002; Berlioz & Strauss, 1948; Rimsky-Korsakov, 1964). Various psychoacoustic studies have also focused on the perception of concurrent instrumental sounds (for a recent review on this, see McAdams, 2019b), many of which have specifically dealt with the perception of blend of the resulting concurrent timbres and its potential acoustic correlates.

Despite the predominant discussion in these studies addressing blend on actually sounding combinations of instruments, blend is also applicable to the mental image of imagined instrumental combinations. The imagination of different instruments interacting with each other and hearing instruments being played internally is commonly referred to as "inner hearing", and is not unfamiliar to musicians, especially composers or orchestrators. Imagined concurrent timbres can be subjected to composers' evaluation for subsequent modification, bridging written scores (initial ideas) and concrete sounds (realization). Previous studies have shown the authenticity of imagined timbral imagery based on long-term memory and prior learning of single instruments emulating that from hearing actual physical stimuli (Crowder, 1989; Halpern, Zatorre, Bouffard & Johnson, 2004). However, little empirical research has been done on the mental image of concurrent timbres, let alone the quality of blend of imagined timbres. One of the basic questions concerning imagined blends would naturally be how comparable they are to perceived actual blends, and furthermore, how acoustic features might contribute to the found relationship. The current study tried to take a first step in this direction, examining imagined blends of different acoustic instruments in comparison with the perception of actual blends of the same instruments, as well as their potential acoustic correlates. This chapter gives an overview of existing research on two essential elements involved in the present study: a) concurrent timbres and perception of instrumental blends, and b) timbre in working memory, followed by a general description of motivations and questions proposed for the current study.

1.1 Concurrent timbres and blends

1.1.1 Musical and perceptual background

Modern investigations on the timbral quality of concurrent instrumental sounds appeared relatively later than those conducted on single instrumental sounds, the first of the former might be credited to Kendall and Carterette's research on the perceptual, verbal, and acoustical attributes of wind instrument dyads (Kendall & Carterette, 1991). On the other hand, the application of concurrent combination of instruments has always been a crucial component in composition and orchestration practices (Goodchild & McAdams, 2018). As Sandell (1989) identified, "Combining timbres, such as for melodic doubling, has been an important part of ensemble writing for centuries … and is likely to remain an important compositional concern in the future".

Generally speaking, the concurrent grouping of sounds is fundamental in how we perceive and understand the sonic environment we experience every day. Auditory scene analysis (Bregman, 1990) provides a framework for how the auditory system makes sense of multiple sound objects and thus derives a meaningful representation of them. Whereas segregation of sound objects facilitates listeners' attention to different events separately (to thus "understand" the sonic environment), under certain conditions sounds from different sources can fuse together into a "virtual" sound source, where new timbres emerge from this perceptual fusion (McAdams, 2019b, p. 218). This latter case is especially pertinent in the creation of musical sounds. In the context of musical activities such as orchestration, the perceptual process involved follows a hierarchy of different organizations that contribute to different levels of orchestration effects (McAdams, Goodchild & Soden, in press). The most basic level of them is the perceptual grouping of concurrent sounds into events where elemental perceptual attributes like pitch and timbre emerge, allowing higher level orchestration effects to function.

Because of its fundamental role in orchestration and rich possibilities amenable to perceptual investigations, concurrent timbres have been one of the earliest focuses of empirical studies on orchestration-related issues. Sandell (1991) discussed several potential topics on orchestration that could be meaningful for perceptual investigations, including semantic descriptors of instrumental timbre, characterizing strength of instruments, and concurrent timbres. Sandell argued that concurrent timbre is a more suitable topic given its great relevance to musicians and central role in orchestration teaching. Several specific sonic objectives concerned with concurrent timbre were listed by Sandell, including augmenting existing timbres, softening timbres, inventing timbres, timbral imitation, etc. These objectives (or techniques) were further distilled into three sonic goals relevant to concurrently sounding timbres (Sandell, 1995): timbral heterogeneity (emphasizing the independence or segregation of timbre), timbral augmentation (having one timbre embellishing another), and emergent timbre (synthesizing a new timbre from existing timbres). These were later borrowed by McAdams et al. (in press) in formalizing different concurrent grouping aims within the taxonomy of orchestral grouping effects.

What underlines these timbral techniques is the perceptual attribute of blend (Sandell, 1991 & 1995), which often functions as a criterion of how well the chosen instruments combine together. According to the definition found in Merriam-Webster, "blend" means "to combine or associate so that the separate constituents or the line of demarcation cannot be distinguished". In the context of music, a similar idea of instrumental blend can be found in orchestration treatises such as by Piston (1955), which generally suggest instrumental combinations "in which the distinctiveness or individuality of the constituent instruments is subordinated to obtaining an overall, uniform timbral quality" (Sandell, 1991). This musical definition is in line with the perceptual fusion process of creating a “virtual source image” described by McAdams (1984b). The primary role of employing blend in orchestration was further stated by Blatter (1997), describing the usage of blending, mixing, matching, and contrasting different instrumental and timbral colors as “one of the chief goals of the orchestrators”. Because of its clear definition across different orchestration treatises and underlying role in evaluating composite sounds, blend is an intrinsic topic when it comes to discussing the effect and quality of different instrumental combinations, rendering itself of great interest to both musicians in actual musical practice and researchers who want to investigate the perceptual groundings of such practice. Additionally, as the previously mentioned different concurrent timbral techniques suggest, blend is not an all-or-none phenomenon (McAdams et al., in press). The amenability of blend to continuous manipulation (and hence a continuous perception of its strength), where the combined instruments can vary from completely blended to completely segregated, facilitates its perceptual modeling. All these facts confirm the importance and pertinence of investigating instrumental blends through the lens of perception.

1.1.2 Previous studies on non-instrumental blend

Prior to the earliest perceptual studies on instrumental blends conducted by pioneers such as Kendall and Carterette (1991; 1993a) and Sandell (1989; 1991; 1995), there had been several auditory perceptual studies conducted concerning blend-related issues for non-instrumental sounds, and results from these studies sometimes address similar perceptual groundings as in the case of instrumental blends. As part of the general background of existing studies on blend, a summary of some exemplary research and findings on non-instrumental blend will be discussed below.

A few studies investigated the recognition of concurrently sounding different vowels with their fundamental frequencies separated by different amounts (e.g., Scheffers, 1983; Halikia, 1985). Results from these studies generally suggest that as the two vowels are separated further in pitch, the accuracy of recognition increases, implying plausibly a decrease in the degree of blend. When vowels are presented in unison the accuracy of recognition is the lowest. These results suggest an effect of pitch separation alone on the perception of blend where unisons have the biggest advantage of promoting blend. Certain non-unison harmonic intervals like fifths and octaves are also more likely to promote blend compared with inharmonic intervals due to a larger number of partials shared by the two sounds (Sandell, 1991).

Studies on the perceptual fusion of different acoustic components within a complex tone also have great implications for the higher-level fusion of different sound sources. The harmonic series is an important cue for yielding an unambiguous pitch sensation for a sustained complex sound. Sustained inharmonic sounds are more likely to yield the perception of multiple pitches which in many cases can be interpreted as the existence of multiple sources (McAdams, 1984b). In terms of the temporal aspect, a series of experiments by McAdams (1984a) showed that the coherence of frequency and amplitude modulation of acoustic components contributes to perceptual fusion. In the amplitude case, this also implies the cue of synchronous onsets of acoustic components, as demonstrated by Rasch (1978) where asynchrony between components in a two-voiced tone leads to easier identification of individual voices and lower masking threshold for the softer voice. These cues of temporal coherence can be summarized under the Gestalt “common fate” principle, which states that sounds changing in similar ways are likely to have originated from the same source (Bregman, 1990). Additionally, if the coupled interaction of amplitude and frequency modulation defines a familiar spectral envelope (e.g., vowel formants), this would suggest a stable resonance structure and hence contributes to perceptual fusion (McAdams, 1984b).

A more empirical investigation of the usage of blend in actual musical practice was the choral blend study by Goodwin (1980). In the experiment, singers were asked to sing a sustained vowel in a soloistic manner and then in a choir singing manner where the singers listened to a pre-recorded well-blended choir sound and attempted to blend with the choir. Spectral analysis of the singing showed that compared with vowels sung in the soloistic manner, those sung in the choir manner had stronger fundamentals with fewer and weaker partials. This transformation of spectral features indicates lowered spectral centroid and darkened timbre, which are in line with singers’ practice of "darkening" their tone to blend with other singers (Sandell, 1995).

These results have either direct or indirect implications for blends of instrumental sounds. The cues of harmonicity, onset synchrony and coherent frequency and amplitude modulation can be applied to the combination of different sound sources and function as concurrent grouping principles of event formation, which leads to a blended timbre. The more these principles converge, the stronger the degree of fusion (McAdams et al., in press). Results from Goodwin’s study suggested an advantage of sound with darker timbre to blend with other sounds, which might also be applicable to instrumental blend.

1.1.3 Previous studies on instrumental blends

As discussed in the background section, the use of concurrent timbres of different instruments constitutes an indispensable aspect of orchestration, and blend is an important perceptual underpinning of the quality of concurrent timbres. Several studies have been conducted on the perception of concurrent timbres and blends of different acoustic instruments and have suggested their potential acoustic correlates. Sandell (1989) investigated blends of fifteen different instrumental sounds synthesized with the line-segment approximations form by Grey (1975). Participants rated all possible pairs of instruments in unison on the degree of blend between the "separated" and "fused". The blend ratings of pairs with a given instrument were averaged to generate a single averaged blend of each instrument, its blendability within the set of sounds. The resulting averaged blends showed significant negative correlation with both spectral centroid and perceptual attack time of single instruments (Gordon, 1987), suggesting that the presence of any dark instrument with quick attack in an instrument pair generally leads to better blend than a pair of two bright instruments with slow attack.

The blend ratings were also used as distances in multidimensional scaling, where the resulting space shows the degree of blend between instruments by their spatial proximity. The technique of multidimensional scaling (MDS) has been often used in perceptual research on timbre where listeners' dissimilarity ratings of sound pairs are mapped onto a spatial configuration in a given number of dimensions. The resulting geometrical structure, often called a "timbre space", is thought to reflect the perceptual qualities listeners use to compare the sounds (McAdams, 2019a). In Sandell's study, the blend ratings were conceived as a measure of psychological proximity, analogous to perceived similarity, between a pair of instruments. The two-dimensional "blend" space generated has two dimensions that correlate well with spectral centroid and perceptual attack time, thus suggesting that the more similar the two sounds are in terms of these two parameters, the better they blend. The way these two acoustic factors interact (specifically in terms of sum and difference of their respective values for individual instruments) and affect blend was further corroborated by Sandell (1991) using the same stimuli and was summarized as a "gravitational effect", which means the lower and closer the values of spectral centroid or perceptual attack time are for both sounds, the better the sounds blend.

Additionally, Sandell compared blends in unison and minor third intervals and the relative explanatory power of different acoustic factors. The results showed that at the unison, the overall lower sum of spectral centroid was more important than the similarity in centroid for promoting blend; of lesser importance were the overall sum of perceptual attack times and their similarity. When the two sounds were separated by a minor third, the relative significance of sum and difference criterion for the two acoustic factors was swapped: the closeness between spectral centroids of the two sounds became more important than their overall lower sum, and the closeness between perceptual attack times emerged as being more important than their overall smaller sum. Both interactions (sum and difference) of the perceptual attack time were of less importance compared to the unison context.

Another important finding from Sandell (1991) is the perceptual evidence of intrinsic blending power of individual instruments, i.e., the fact that "certain instruments tended to impose a certain degree of blend regardless of what they were paired with". "Good blenders" revealed by the experiments, such as bassoon and French horn, were also confirmed in different orchestration treatises. The underlying acoustic factors were identified as lower spectral centroid and shorter attack.

Finally, Sandell (1995) re-investigated the acoustic correlates of instrumental blends using the same stimuli from his experiments in 1991, considering more acoustic parameters derived from the composite concurrent sounds (e.g., in the case of spectral centroid, it is computed on the composite sound itself, which can be seen as a counterpart of the sum of spectral centroids in Sandell's research in 1991). He also examined potential interaction parameters between pairs of instruments (such as the correlation of temporal centroids between two instruments). Stepwise correlation identified composite spectral centroid as the most important factor affecting blend (lower composite centroid leads to better blend) when the two sounds were in unison, followed by the attack contrast (a measure of how different the attack envelopes are between two sounds). When the two sounds were separated by a minor third, the difference between spectral centroids for the two sounds replaced composite spectral centroid as the most important factor (smaller difference leads to better blend), corroborating the findings in Sandell (1991). Several temporal features that were found to be significant in the unison context became non-significant in the minor third context, which, together with the emerging importance of centroid difference, might be explained by the increased spectral distinguishability of the two sounds following the pitch separation. As a result, listeners can more reliably use the spectral difference to make blend ratings, whereas temporal features lose their relative perceptual importance as they don't provide further information beyond that of spectral features.

The stimuli used by Sandell's experiments are constrained by their synthetic nature and short duration (around 300 ms). Some early explorations on instrumental blend using authentic recordings in longer musical contexts were conducted by Kendall and Carterette (1991; 1993b; 1993c). They investigated the perceptual similarities of concurrent timbres of different wind instruments with multidimensional scaling. Stimuli were all played by instrumentalists, covering several musical contexts for concurrent timbres with varied durations (unison, unison melody, major third and harmonized melody). Subsequent verbal attribute rating experiments identified two stable dimensions in the resulting similarity space across different musical contexts as "nasal" vs. "not nasal", and "rich" vs. "brilliant". Preliminary findings in their experiments also suggested a direct relationship between timbral profiles of single instruments and their concurrent pairs, as the concurrent timbral similarity space can be largely reconstructed from unweighted vector sums of single instruments placed in the same perceptual similarity space. Using the same stimuli, Kendall and Carterette conducted further experiments (1993a) about the blend and identifiability of concurrent wind timbres. There was a moderately strong inverse relationship between blend and identifiability, meaning that the less two instruments blend, the easier it is to identify them individually in their concurrently sounding pair. As in Sandell's experiments, the unison context produced the highest blend ratings and lowest identification. Specifically, oboe dyads which were identified as "nasal" in their previous experiments received the lowest blend ratings. They also found that the degree of blend can be rather well predicted by the distance between the constituent instruments in their 2-D perceptual similarity space, with greater distance corresponding to less blend. This was supported again when they submitted blend ratings to multidimensional scaling resulting in a "blend" space, which was nearly identical to the perceptual similarity space of single instruments.

More recent studies on instrumental blend expanded both the scope of instruments and musical contexts for blend to happen. Tardieu & McAdams (2012) studied specifically the blend between a sustained instrument (woodwind, brass, bowed string) and an impulsive instrument (pitched percussion, plucked string), which had never been investigated before. They found that longer attack times and lower spectral centroids increased blend where the properties of the impulsive instrument had more contribution than those of the sustained instrument. On the other hand, the overall emergent timbres of such dyads were controlled primarily by the spectral envelope of the sustained instrument and the attack of the impulsive instrument. They concluded that in orchestration practice for dyads with such mixing of instruments, their perceived blends and overall timbres can be controlled almost individually by choosing the impulsive and sustained instruments. Lembke, Parker, Narmour and McAdams (2019) included pizzicato strings in studying blends of instrumental dyads and triads. Partial least-squares regression was used to investigate the relationship between perceived blends and different acoustical predictors related to timbres and musical contexts (e.g., pitch and articulation). The type of articulation (e.g., impulsive vs. gradual attack) was proven to be important for predicting blend where the presence of plucked strings resulted in clearly lower blend ratings than for combinations of sustained instruments.

Blends within larger real-world orchestral contexts were also investigated by McAdams, Gianferrara, Soden and Goodchild (2016). The stimuli they used were orchestral excerpts involving different instrument family combinations with varying numbers of instrumental parts. Potential factors affecting blends included in the analysis were type of timbral blend (timbral emergence or timbral augmentation), timbral category (combinations of instrument families), number of instrument parts and degree of parallelism of musical lines. Two blend-related perceptual criteria were used in experiments: blend vs. no blend and unity vs. multiplicity. Stepwise regression showed a significant effect of degree of parallelism in accounting for the variance of ratings. Results also suggested complex interactions between the above factors in explaining orchestral blends rather than independent effects of isolated factors. Additionally, the two tested perceptual criteria seemed to reveal a more complex nature of musical blend, as combinations can be perceived as “multiple” but still blended. This suggested that “the notion of musical blend is not synonymous with complete perceptual fusion”.

An approach investigating instrumental blend that focused on the aspect of performers’ efforts can be seen in Lembke, Levine and McAdams (2017). They studied how performers, specifically a bassoon player and a horn player, achieve blended timbres together in various musical contexts when assigned different performance roles as “leader” or “follower”. In line with the timbral darkening strategy found in Goodwin’s vocal blend study (1980), they found that when assigned the role “follower”, musicians adjusted the sounds they produced towards a darker timbre with reduced frequencies in the main formant or lower spectral centroids, together with slight reductions in sound level. The findings suggested that instrumentalists’ roles in the performance would determine how they coordinate with each other and adjust the sounds to achieve a common sonic goal as in the case of timbral blend.

An acoustic explanation of blends often seeks to correlate blend ratings with generalized global acoustic descriptors like spectral centroid. The perceptual importance of local descriptors for instruments' formant structures has also been studied by Lembke & McAdams (2015). They used pitch-generalized spectral envelope descriptions to characterize the formant structures of wind and brass instruments. The perception of their blends was found to be affected by the relative position and prominence of instruments' main formants. Specifically, for a dominant instrument and a subordinate one to blend well, the higher upper bound of the main formant for the subordinate instrument should not exceed than that of the dominant instrument.

1.2 Timbre in working memory

As discussed in the previous sections, existing research in blend mostly addresses it as a perceptual phenomenon characterizing concurrent sounds. However, the perception of blend can also happen in an imagined combination of sounds, where the sounds are "perceived" and evaluated in the form of mental image. To investigate such "virtual blend", it is necessary to first know how timbre is stored in working memory, comparing it to the sensory representation of timbre resulting from real stimuli. A few existing studies on this topic will be discussed below.

1.2.1 Maintenance of timbre in working memory

According to the classic multistore model, as elaborated by Atkinson and Shiffrin (1968), human memory can be thought of as structured in terms of different independent entities ("stores") functioning on different time scales. Between the lowest storage of sensory register with fast-decaying pre-attentive information and longer-term memory (LTM), short-term memory (STM) is responsible for retaining categorical information for a measurable time, where the duration of retention can be lengthened by active rehearsal. In the case of exerting conscious rehearsal and maintenance, the concept of STM is more closely related to working memory (WM), which "is usually defined as an active form of memory that, as a whole, underpins a range of important cognitive faculties such as problem solving and action control" (Siedenburg & Müllensiefen, 2019, p. 99). An example of active maintenance in WM is explicit or implicit vocalization of memorized items in the case of verbal WM, where the original memory trace is consciously refreshed and kept in a "phenological loop" by (sub)vocal rehearsal (Baddeley, 2012).

The concepts of STM and LTM can also be applied to the case of non-verbal auditory memory, with STM operating on sensory representations rather than on verbal items (Cowan, 1984). Of particular interest and relevance to the context of the current study is how timbre is kept in WM, as the imagination of instrumental blends requires active control of mental images of different timbres. The way timbre is maintained in short-term working memory has been demonstrated to be different compared to other non-verbal auditory memory as in the case of melody. Using a dual task paradigm, Nees et al. (2017) demonstrated that melodic short-term memory is maintained by subvocal articulatory rehearsal (i.e., by singing or humming internally to oneself), which suggests a strong similarity with verbal short-term memory. Unlike simple melodies, timbres of most instruments cannot be faithfully reproduced vocally by humans, which renders them less likely to be recoded into other formats such as an internal motor code for subvocal rehearsal of melodies in WM (Crowder, 1993). The concept of visual image thus may apply well to the case of timbre in WM. Similar to a visual image, a "timbral image" preserves the sensory coding of the original experience of hearing the timbre without being contaminated by other forms of recoding (Crowder, 1993).

Some experiments have provided evidence that timbre in WM is likely to be maintained by "attentional refreshing" (Camos, Lagner & Barrouillet, 2009) of the initial sensory trace of hearing it, i.e., by "replaying" the timbral image internally that was just activated. Schulze & Tillmann (2013) found that concurrent articulatory suppression (by asking participants to count out loud from 1 to 5) didn't impair the performance of backward recognition of timbral series of different acoustic instruments. This result suggests that WM of timbre is unlikely to involve verbal labeling of timbres and maintenance by articulatory rehearsal of the labels, because otherwise the recognition performance would be worse as it shares similar mental resources required by the suppression task. Moreover, Soemer & Saito (2015) showed that attention-driven reenactment of the auditory memory trace can be the underlying resource for timbre in STM. Participants in their experiments were exposed to series of artificial sounds differing in their timbres and they had to judge if a delayed probe sound was in the heard series. Different types of secondary suppression tasks were included during the delay between the series and probe sound. Their experiments showed that maintenance of timbral information was robust to articulatory suppression (where participants were asked to articulate the syllable “da” at a specified interval). However, a secondary auditory imagery task (where participants were asked to evaluate the imaged pitch height of auditory imageries evoked by certain onomatopoeic words) was disruptive to main task, suggesting possible shared mental resources between auditory imagery (which is obviously attention-driven) and maintenance of timbre in WM.

A similar experimental scheme was used in Siedenburg & McAdams (2017b) where they tested item recognition on timbres with both familiar acoustic instrumental sounds and unfamiliar transformed sounds. Both concurrent articulatory suppression (by asking participants to count aloud) and concurrent visual distractor task (by asking participants if there was a direct repetition of grid in a black-and-white grid pattern sequence) impaired the main recognition task for both familiar and unfamiliar sounds. The fact that unfamiliar sounds were unlikely to be labelled argues against verbal labelling and rehearsal of labels as a maintenance strategy for their timbres. The negative effect of articulatory suppression on the main task was therefore mainly attributed to its interference with the auditory trace (on which attentional refreshing relies). The negative effect of the visual distractor task on the main task was attributed to reduced attentional resources that attentional refreshing also relied on.

Overall, these studies show support for attentional refreshing (i.e., mentally replaying timbres) as the underlying mechanism of how timbre is preserved in WM, which is unlikely to be mediated by “the persistence of the sensory memory trace” (Siedenburg & McAdams, 2017b) or verbal labelling of timbres and subsequent rehearsal of the labels.

1.2.2 Timbral imagery

Timbral imagery is a closely related topic which provides additional helpful clues to the mental representation of timbre. The concept of “imagery” needs to be differentiated from that of “image” (as mentioned in the previous section) in that the concept of “image” applies to the original sensory representation from actually seeing (in the case of a visual image) or hearing (in the case of an auditory image) the target object, whereas in the case of “imagery” the same representation is derived “top-down” (from long-term memory contents) without prior stimulation (Crowder, 1993); examples include “picture an apple” or “imagine a piano sound”.

Some empirical evidence has demonstrated that imagery for timbre closely resembles the concrete sensory representation (mental “image”) of actually hearing the timbre, a characteristic that aligns with how timbre is maintained in working memory as discussed in the previous section. Crowder (1989) compared two experimental conditions where the tasks were both to judge whether two consecutively presented tones, which could vary in timbre, were of the same pitch or not. In the first condition, participants listened to the two consecutive tones, whereas in the second condition the first tone was imagined by participants with a presented pitch height and a specified instrumental timbre. In both conditions, the same qualitative effect was found: matched timbres facilitated correct "same-pitch" responses and the response times were faster than in tones of different timbres. This result was interpreted as evidence for the sensory-based nature of timbral imagery, that "the neural consequences of hearing an instrumental timbre and imagining it are, to some extent, equivalent" (Crowder, 1993). With the help of functional magnetic resonance imaging, Halpern et al. (2004) was able to directly compare the brain activity from hearing and imagining sounds with different timbres. Participants were asked to compare pairs of timbres and rate the perceived dissimilarity while their brain activities were recorded. In another experimental condition the same procedure was repeated expect that stimuli were to be imagined by participants. Results showed a significant correlation between the dissimilarity data in the perceived and imagined conditions, suggesting a strong parallel between perceived timbre and timbral imagery. Moreover, the brain activity in the two conditions featured the same pattern in the auditory cortex. These results, taken together, point to the "authenticity" and accuracy of timbral imagery, that "sensory representations activated by imagery can resemble those activated by sensory stimulation" (Siedenburg & Müllensiefen, 2019, p. 102).

Overall, existing studies on the maintenance of timbre in WM and imagery for timbre speak to the active and "experiential" nature of timbre cognition: timbre can be maintained in working memory by refreshing the initial sensory trace of hearing the timbre, and the mental image of timbre can also be re-constructed from long-term memory, which resembles actual sensory stimulation. Both processes involve the re-creation of aspects of original perceptual experience, i.e., of hearing the timbre itself. Relating back to the phenomenon of "inner hearing", these facts suggest aspects of its cognitive nature: imagining and actively maintaining timbre in WM leads listeners to actually "hear" the original sounds internally.

1.3 Current study

1.3.1 Motivations

The capability of humans to recreate authentic mental images of sounds without necessarily hearing them underlies implicitly many different musical activities. In terms of performance, it could be the case that orchestral players would imagine the pitch of an upcoming tutti entrance in unison to facilitate tuning (Zatorre & Halpern, 2005). It is not difficult to extend this to a scenario of achieving timbral blend between an existing orchestral pedal and an upcoming new chord where players of the latter would adjust their articulation and intonation based on how they imagine they would blend with the ongoing orchestral sounds, for example. When composers and conductors study or write a score, instead of conceiving the music at hand as a set of abstractions of musical rules, they readily imagine all the musical aspects (pitch, rhythm, timbre, etc.) to mentally stage the musical scene.

As discussed in the previous sections, many studies utilizing various neurological methods have suggested in general that neural activities responsible for internal auditory representation can occur without sound stimuli, which possibly mediates the experience of imagining music (Zatorre & Halpern, 2005). It is still not clear, on a higher perceptual level, how different concrete musical treatments (or techniques) involving <u>interactions</u> of sounds (e.g., segregation, blend, layering, etc.) render themselves comparatively in actual perception and imagination. In contrast to a single tone or a single melody, real-world music is much more complex when considering all the possibilities of sound interaction. It is therefore of great interest to further investigate these higher-level musical organization methods from the two alternative perceptual angles, i.e., “corporeal” vs. imaginary.

Out of all the sound interaction strategies available, instrumental blend is probably the most basic one, the simplest form being two different instruments sounding together. As shown in the previous introduction, blend is also the building block of many higher-level orchestration techniques. Given its formal simplicity and musical importance, instrumental blend would an excellent candidate for a preliminary investigation into the working mechanisms of “imaginary musical soundscape”.

1.3.2 Objectives

For composers and conductors, the scenario of imagining different combinations of instruments working together is mostly associated with timbral imagery activated from long-term knowledge of the instruments. This dependency on prior knowledge of instruments makes long-term timbral imagery highly individual and most exclusive to trained musicianship. Extra limitations will be needed to facilitate a controlled experimental setup where imageries generated by different people can be meaningfully compared. Thus, to reduce the complication of individual knowledge and allow comparison between musicians and non-musicians, the imagined instrumental blend studied here will be generated from short-term timbral images, i.e., after hearing the sounds. Participants will hear single sounds of instruments, and the imagined blends will be constructed based on these concrete stimuli. This design also allows a straightforward comparison between imagined blend and heard blend: in a separate condition, single instrumental sounds used in the imagining condition will be paired and played together, allowing participants to rate the heard blend.

The main objective of this study is thus to compare how listeners perceive the physically heard and imagined instrumental blend. This includes identifying how different the two types of blends are for individual instrument pairs and the overall agreement between heard and imagined blend. Additionally, whether the factor of musical background (musicians vs. non-musicians) influences how the two types of blends are perceived will be studied. To understand how acoustic parameters might contribute to the perception of the two types of blends differently, acoustic analyses will be conducted on the stimuli. Extracted acoustic features will be used for regression modeling of blending ratings in the two conditions separately. Finally, to visually compare how instruments blend in the two conditions, multidimensional scaling (MDS) will be used to construct two "blend spaces" by treating the degree of blend as a similarity measure between instruments. Similar treatment can be seen in Sandell (1989; 1991) and Kendall & Carterette (1993a). Acoustic features extracted from the previous analyses will be correlated with the scaling configurations to help interpret the MDS results.

1.3.3 Research questions and hypotheses

In line with the objectives, a few questions can be proposed for the current study:

i. How comparable are the perceived degrees of blend for imagined and heard blends? Additionally, does the factor of musical background play a significant role?

ii. Do listeners rely on similar or different acoustic parameters in their perception of imagined and heard blends?

Although there is very limited knowledge about imagined blend, some preliminary hypotheses can be drawn from available evidence about blend and timbral image/imagery outlined in earlier sections. Stiller (1985) suggested, from the point of view of orchestration practice, that the timbre of two different instruments sounding together is intermediate between the two constituent timbres, and one is soon able to imagine the sound of such combinations without having heard them before. This observation doesn't address blend directly, but it does offer a possible clue as to how listeners, especially musicians, might evaluate the blending sounds in imagination. Based on Stiller's observation, for people who have been working with combinations of instruments (e.g., composers, orchestrators, conductors, chamber performers, etc.) it might be easier to imagine combinations of sounds that match how they would sound together. They could have developed a set of strategies based on previous experience that allows them to efficiently conjure up mental images of combinations of instruments. On the contrary, it might be hypothesized that people who don't usually and intentionally engage with musical activities involving the evaluation and creation of combinations of instruments could be more puzzled by the imagination task, thus giving ratings that are more likely to deviate from those of physically heard blends.

Regarding the second question, it has been shown in the introduction of timbral imagery that the neural activities associated with imagery can match well those from hearing real sounds. In the context of the current study, it seems logical that listeners would still give overall coherent ratings between heard and imagined blends (i.e., the correlation between ratings for heard and imagined blends should be high). However, there is no evidence on whether the quality of timbral images can still stay the same under mental manipulations as the imagination of blends involves retrieving separate timbral images and superimposing them together (a "virtual blend"), which potentially draws on additional attentional resources. Studies such as Soemer and Saito's (2015) and Siedenburg and McAdams' (2017b) have shown that the maintenance of timbral information relies on attentional refreshing. This might imply that timbral imageries could be “downgraded” or further abstracted as additional efforts are being made to superimpose them internally. As a result, it is possible that features of stimuli will be extracted differently by listeners when imaging blends compared with when evaluating heard blends. How exactly different acoustic features might be drawn upon by listeners in the two conditions will need to be uncovered in the analysis.

Chapter 2

Method

2.1 Main experiment: perception of heard and imagined blends

2.1.1 Participants

Sixty-four participants in total were recruited (male = 26, female = 35, non-binary = 3; mean age = 23.1) which were categorized as musicians (who are currently pursuing a degree in music with at least five years of formal music training) and non-musicians (who had never pursued any music degrees). Musicians and non-musicians came from Schulich School of Music at McGill and the general Montreal community, respectively. Participants recruited for the non-musician group were asked to provide a summary of their musical backgrounds, if any; two of them were later grouped into musicians as they received more than five years of formal music training and identified themselves as serious amateur musicians. Overall, there are 32 people in the musician group (male = 18, female = 11, non-binary = 3; mean age = 23.8), with a mean average of 15.5 years of musical training (SD = 5.45), and 32 people in the non-musician group (male = 8, female = 24; mean age = 22.4), with a mean average of 2.1 years of musical training (SD = 3.34)[1]. Before the experiment, participants passed a pure-tone audiometric test at octave-spaced frequencies from 125 Hz to 8 kHz (ISO 389–8, 2004; Martin & Champlin, 2000) and were required to have thresholds at or below 20 dB HL to proceed to the experiment. Participants were compensated for

[1] It is worth mentioning that many non-musician participants reported years of music training including various forms of non-continuous informal learning. After cross-checking with verbal feedback gathered by the experimenter, only those whose background was sufficient were considered eligible for the musician group.

their participation. This study was certified for ethical compliance by McGill University's Research Ethics Board II and all participants signed written consent forms before the experiment.

2.1.2 Stimuli

To cover a wide range of instruments allowing for different degrees of blend, 14 different instruments were chosen as stimuli in this study, covering all major orchestral families: the woodwind family includes flute (abbreviated as "FL" in subsequent tables and figures; same for the other instruments), oboe (OB), English horn (EH), bassoon (FA), B-flat clarinet (KLB); the brass family includes C trumpet (TrC), French horn (HO), tenor trombone (TP), tuba (TU); the (pitched) percussion family includes celesta (CE), vibraphone (Vib), tubular bell (RGL); the string family includes violin (VI) and cello (VC). Stimuli for the fourteen instruments were selected from Vienna Symphonic Library (https://vsl.co.at). All played a sustained D#4 note at a forte or mezzo-forte dynamic with ordinary articulations. Instrumental samples were downmixed to mono by averaging across the two channels and then trimmed to 2.2 seconds in duration, where a linear fade-out envelope was applied to the ending 0.2 seconds. The resulting stimuli all have a sampling rate of 44.1 kHz with 16-bit amplitude resolution. Single instrument stimuli were paired up with each other, forming 91 pairs in total. These paired stimuli were presented differently in two different consecutive conditions ("sequential" and "concurrent") in the experiment aiming for testing the perception of imagined and heard blend, respectively. In the sequential condition, the two paired instruments were played one after the other with a gap of 0.5 seconds and participants were asked to imagine them being played simultaneously and rate the degree of blend of this imagined pair. In the concurrent condition, the two instruments were played simultaneously and participants were asked to rate the perceived degree of blend directly.

As the present study is mainly concerned with blends of concurrent timbres, other musical parameters that may affect blend should be controlled as much as possible in both sequential and concurrent conditions. One such parameter is loudness (Sandell, 1991). To equalize the perceived loudness of the 14 stimuli, six volunteers participated in a loudness-matching experiment in which they had to adjust the loudness of all other stimuli to match that of the oboe sample (which functions as the reference). One volunteer ran the experiment only once and the rest five volunteers ran the experiment twice, generating 11 loudness adjustment values in total. The medians of adjustment values were applied to the corresponding stimuli except for the celesta sample where an additional 2dB boost was applied because of inadequate loudness boost with the original median value[2].

Another important factor affecting blend is the onset synchrony between constituent sounds (McAdams, 1984b). Because non-synchronized sound events can prevent blending, it is necessary to ensure that the onset asynchrony for a given concurrent pair is minimized. Given that different instruments have different perceptual attack times (Gordon, 1987), aligning physical onsets does not necessarily ensure the perceptual attack synchrony. To align instrumental stimuli in the concurrent condition, seven volunteers participated in an attack synchronization experiment following a similar design in Gordon (1987) where all 91 pairwise combinations of instruments were synchronized. The medians of time shift values were applied to respective instruments in all pairs. The experimenter did a final listening check on the synchronized stimuli and made minor adjustments to pairs where the synchrony was not very satisfactory. The adjusted values were used to prepare stimulus pairs in the concurrent condition.[3]

2.1.3 Experimental design

The experiment was a three-way mixed design with one between-subjects factor—the musical background with two levels: musicians and non-musicians—and two within-subject factors—91 instrument pairs and two blending conditions (imagined and heard).

2.1.4 Procedure

The experimental session was run with the PsiExp computer environment (Smith, 1995). Sounds stored on a Mac Pro 5 computer running OS 10.6.8 (Apple Computer, Inc., Cupertino, CA) were amplified through a Grace Design m904 monitor (Grace Digital Audio, San Diego, CA) and presented over Dynaudio BM6a loudspeakers (Dynaudio International GmbH, Rosengarten, Germany) arranged at about ±60°, facing the listener at a distance of 1.5 m. Participants were seated in an IAC model 120act-3 double-walled audiometric booth (IAC Acoustics, Bronx, NY). The amplification level of the monitor was chosen in advance by the experimenter after pilot

[2] Values of loudness adjustment applied to instruments are documented in Appendix 1

[3] Values of time shifts applied to instrument pairs are documented in Appendix 2.

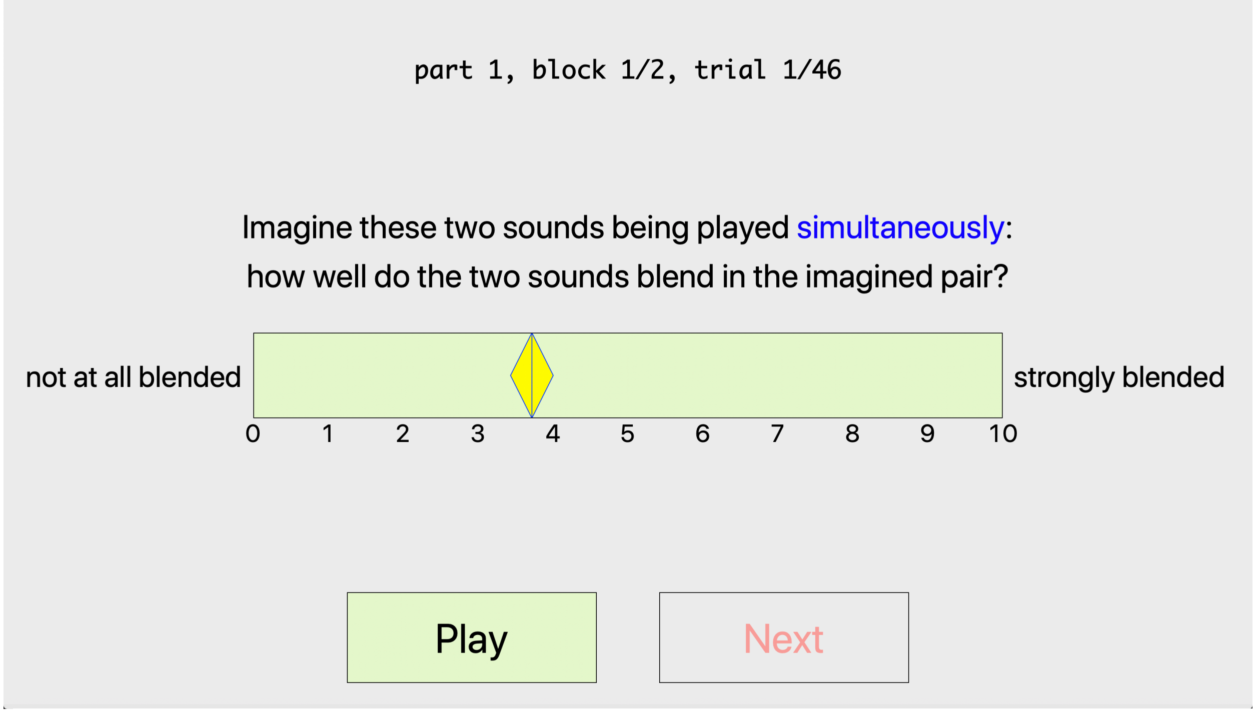
sessions to ensure a comfortable level for listening to all stimuli in the experiment and remained fixed for all participants.

Participants were briefly introduced to the experimental procedure first and were corrected for any misunderstandings about the notion of blend that was being tested in the experiment (i.e., instead of how pleasant the combined sounds sound like, blend in this study is about whether different sounds fuse into one virtual sound source).

At the very beginning of the experiment, participants were provided two examples of instrument pairs that are generally perceived to blend well (violin + cello) and poorly (flute + tubular bell). “Well-blended” was described as “the sounds fuse together as a single unity in perception when they sound together”; “poorly-blended” was described as “the sounds are more easily perceived as separate sources when they sound together.” For these examples, constituent instruments were played one after the other followed by the concurrently sounding pair.

Participants then entered a familiarization phase where 21 instrument pairs (randomized possible combinations of flute, oboe, trumpet, tuba, violin, celesta, tubular bell) were played sequentially. This was designed to allow participants to establish an idea of the possible range of blends they would be rating on and to decide how they would use the range of the rating scale.

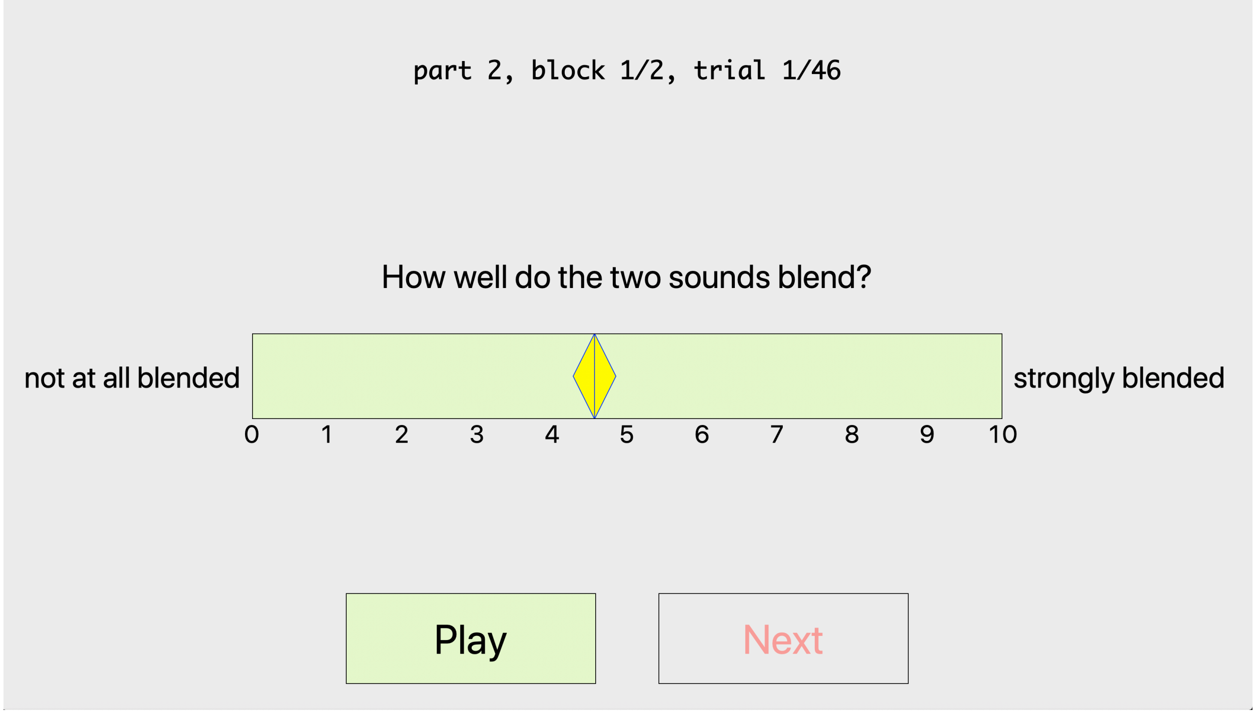
The entire experiment was divided into two parts, corresponding to the two conditions described in the previous part. As hearing concurrent pairs first might prime listeners on the quality of blends and interfere with how they might imagine blends, the sequential condition (where participants had to imagine) was always presented first. In the first part (interface shown in Fig. 2-1), paired instruments were played one after the other, which could be replayed altogether by pressing the “Play” button. The order of presentation was randomized. Participants were asked to imagine the two instruments being played simultaneously and rate how well the two sounds would blend in the imagined pair by placing a freely movable cursor on a bar, where a rating towards the left represents a low degree of blend and a rating towards the right represents a high degree of blend. Ratings were scaled to 0 ~ 1 for analysis. In the second part (interface shown in Fig. 2-2), paired instruments were played simultaneously and could be replayed several times. Participants were asked to rate the degree of perceived blend of the sounding pairs. The order of presentation of instrument pairs were randomized for both parts. Each part was further divided into two blocks. Participants could choose to take a short break between blocks in a part and between the two parts.



part 1, block 1/2, trial 1/46

Imagine these two sounds being played simultaneously:
how well do the two sounds blend in the imagined pair?

Figure 2-1. Experimental interface of the sequential condition.



part 2, block 1/2, trial 1/46

How well do the two sounds blend?

Figure 2-2 Experimental interface of the concurrent condition.

2.2 Acoustic analysis

2.2.1 The Timbre Toolbox

For the extraction of potential acoustic features related to blend, the Timbre Toolbox (Peeters, Giordano, Susini, Misdariis & McAdams, 2011, revised Kazazis, Depalle & McAdams, 2021) was used, which offers a wide range of audio descriptors useful for musical perceptual studies. Previous studies have pointed out the importance of several temporal and spectral factors that contributed significantly to blend, common ones such as attack time and spectral centroid (Sandell, 1989 & 1991; Tardieu & McAdams, 2012; Lembke et al., 2019), along with various global and local spectral-envelope features (Lembke & McAdams, 2015; Lembke et al., 2017). In light of these findings, several potential descriptors were chosen in the analysis that sought to cover various spectral, temporal and spectro-temporal aspects of the stimuli while not overusing a large number of features with unnecessary redundancy. Given the results of hierarchical cluster analysis in Peeters et al. (2011) showing correlations among audio descriptors in the Timbre Toolbox, eight descriptors were chosen (see Table 2-1) aiming to span different clusters (i.e., represent various acoustic aspects of the stimuli) without introducing redundant colinear descriptors. Attack slope (AttSlope) is a global descriptor computed on the temporal energy envelope of the audio signal that measures the average temporal slope of the energy envelope during the attack segment. RMS energy (RMSErg) is a time-varying descriptor measuring the root-mean-square energy of overlapping time frames of the audio signal (window length = 1024 samples, hop length = 512 samples). Spectral centroid (SpecCent), spectral crest (SpecCrest) and spectral variation (SpecVar) are time-varying descriptors addressing the content, shape, and temporal varying aspect of the frequency spectrum of the sound, calculated from the magnitude-squared STFT representation of the signal with Hann windows (window length = 2048 samples, hop length = 512 samples). Spectral centroid measures the center of gravity of the spectrum and is usually associated with the brightness of the sound (McAdams, 2013, p. 41). Spectral crest measures the peakiness of the spectrum which can be used to distinguish between noise-like and tone-like sounds. Spectral variation represents the amount of variation of the spectrum over time. Inharmonicity (InHarm), noisiness and tristimulus (three ratio values abbreviated as “Tri1”, “Tri2” and “Tri3”) are time-varying descriptors addressing the harmonic properties of the sound, calculated from the sinusoidal harmonic partial representation with Blackman windows (window length = 2048 samples, hop length = 512 samples). Inharmonicity measures the deviation of frequencies of partials from pure harmonic frequencies of the fundamental. Noisiness is calculated as the ratio of noise energy (the remainder after harmonic energy has been removed from the total energy) to total energy in the signal. Tristimulus characterizes the distribution of energies among different partial regions of the spectrum, with the first value (Tri1) measuring the proportion of total energy at the fundamental frequency, the second value (Tri2) measuring the proportion of energy of the second to fourth partials, and the third value (Tri3) measuring the proportion of energy of higher partials.

For time-varying descriptors, different summary statistics were used to reflect the general trend (median) and variability (inter-quantile range [IQR]) of the descriptors in a single stimulus. It has been shown that for certain descriptors the median and IQR values are closely correlated, but for others these two statistics reflect different aspects of the stimuli. Following the clustering results in Peeters et al. (2011), for the current analysis descriptors showing significant correlations between their median and IQR values (linked at the bottom of the dendrogram) were summarized by median values only. For the rest of the descriptors, both median and IQR were calculated (denoted by suffixes "_Med" and "_IQR", respectively, following the abbreviated names of acoustic descriptors). In total, this gives 13 acoustic descriptor values for single instruments (tristimulus has three values). The feature extraction process was carried out in Matlab version R2020a (The MathWorks, Inc., Natick, MA).

Table 2-1. Acoustic descriptors calculated on single instrument sounds, the respective input audio representations used for calculation, and statistics used for summarizing time-varying descriptors. Pairwise associations were used to generate pairwise features.

Descriptors	Input representations	Summary statistics	Pairwise associations
AttSlope	temporal energy envelope	-	sum, diff
RMSErg	audio signal	Med, IQR	sum, diff
SpecCent	power STFT	Med, IQR	sum, diff
SpecCrest	power STFT	Med, IQR	sum, diff
SpecVar	power STFT	Med	sum
InHarm	sinusoidal harmonic	Med	sum
Noisiness	sinusoidal harmonic	Med	sum, diff
Tristimulus	sinusoidal harmonic	Med	sum, diff, correlation

2.2.2 Pairwise features

To model blends with acoustic descriptors, it is necessary to come up with pairwise features that summarize relations between descriptors of constituent instruments. One way to do so, as can be seen in Sandell (1991), Lembke & McAdams (2015) and Lembke et al. (2019), is by simply taking the absolute difference and composite (sum) of associated descriptors for the paired instruments, which will be adopted in this study (denoted by suffixes "_sum" and "_diff", respectively). Modeling like this provides two alternative ways to investigate how the association between instruments affects blend, i.e., when two instruments show similar/different/overall high/overall low values for a specific acoustic descriptor, how does the perceived blend change. Additionally, the Pearson correlation of tristimulus values between paired instruments (TriCor) was also considered as a pairwise feature. Overall, this gives 13×2+1 = 27 pairwise features. A correlation matrix was calculated on all pairwise features over all 91 instrument pairs. Two features (InHarm_Med and SpecVar_Med) were found to show significant correlation (r > .9) between their own composite and difference pairwise features. Thus, only their composite pairwise features were included in the blend regression analyses. At the end, 25 pairwise features were used as potential regressors for the acoustic modeling of blend ratings.

Chapter 3

Results

Blend ratings in the two conditions were first analyzed using correlation to show the degree of their overall similarities. ANOVA analyses were subsequently carried out to test the effects of participants' musical backgrounds, instrument pairs, blending conditions, and their interactions. Results of acoustic modeling of blends are presented next, followed by multidimensional scaling of the blend ratings. All analyses were carried out with R version 4.2.0 (http://www.r-project.org).

3.1 Correlation between the two blending conditions

Plotting the distributions of blend ratings for musicians and non-musicians separately for all combinations of the two within-subjects factors (instrument pair and blending condition) revealed that many of them were not normally distributed and had outliers, which was confirmed with Shapiro-Wilk tests.[4] Thus, to correlate ratings between concurrent and sequential conditions, the median was used to represent participants' blend ratings for a given instrument pair as it is less biased by outliers. Fig. 3-1 and 3-2 show the variation of median blend ratings in the two blending conditions across all 91 instrument pairs for musicians and non-musicians, respectively. Both musicians and non-musicians gave "comparable" ratings across heard and imagined blends which are highly correlated: for musicians, $r = .96$, $p < .0001$; for non-musicians, $r = .95$, $p < .0001$. This result shows that despite local differences between the two conditions, the perceived degree of

[4] For musician + concurrent condition, 55 out of 91 instrument pairs have non-normal rating distribution; For musician + sequential condition: 22 out of 91 instrument pairs have non-normal rating distribution; for non-musicians within concurrent and sequential conditions, the numbers of pairs with non-normal rating distribution are 79 and 63, respectively (alpha level = 0.05).

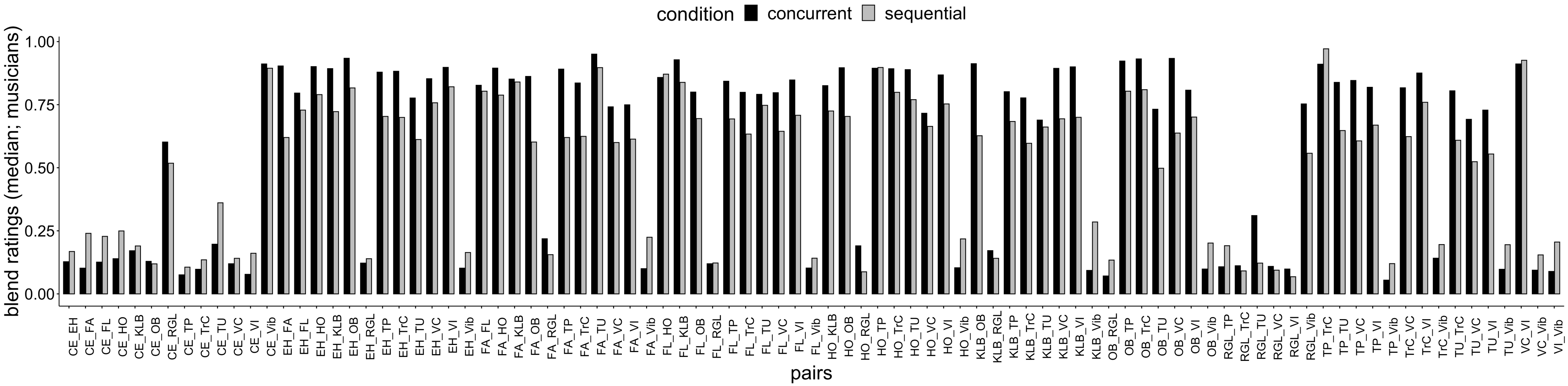


Figure 3-1. Median blend ratings for each instrument pair within musicians.

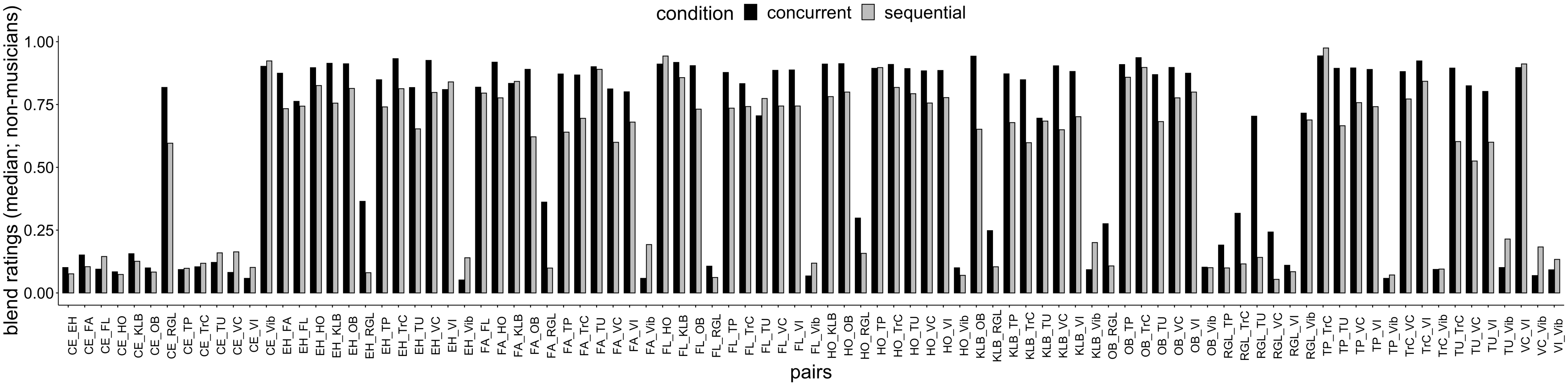


Figure 3-2. Median blend ratings for each instrument pair within non-musicians.

blend is fairly consistent between them for both musicians and non-musicians, i.e., when two instruments are perceived to blend well when sounding together, they would also blend relatively well in imagination.

3.2 ANOVA results

The original within-subjects factor of instrument pair contains 91 levels. While this retains the specificities of instruments and allows close examination of how perception of blend might differ between the heard and imagined conditions, it might be also be helpful to determine whether any patterns can be found in which specific instruments are generalized into their instrumental families, i.e., when this factor of instrumental combinations is further abstracted as family combinations. Thus, two different analyses of the ratings are examined here, the first being the original design where blends are studied on an "instrument-wise" basis, the second being an abstracted organization with "family-wise" blends which will be presented in later sections.

3.2.1 Instrument-wise blends

For the original organization of rating data, there are one between-subjects factor—musicianship (musicians vs. non-musicians)—and two within-subjects factors—one being instrument pair (91 different pairs), the other being blending condition (concurrent vs. sequential). As the sphericity assumption test were not able to be calculated properly in R for the within-subjects factors due to greater number of factor levels than that of observations (number of participants in this case), an alternative approach for calculating the significance of effects and interactions of factors with linear mixed models was used, which does not require the sphericity assumption (Field, Miles & Field., 2012, p. 621). Participants were treated as grouping factors across which intercepts are allowed to vary, modeling individual differences.[5] The ANOVA table with F-tests and associated p-values was extracted from the linear mixed model using

[5] It has been suggested recently (Heisig & Schaeffer, 2019) that for multilevel models with cross-level interactions, random slopes also be added to lower-level components (in the current case, they are the two within-subjects factors). Random-intercept-only models could give "severely anti-conservative statistical inference" with inflated Type I error rates. Due to time constraints encountered in the analyses with computational difficulty, however, only random intercepts were included in the model.

Satterthwaite's method (Satterthwaite, 1946) via the R package "lmerTest" (Kuznetsova et al., 2017). Results were summarized below.

3.2.1.1 Main effects

The musical background of participants had no significant effect, $F(1, 62) = 0.54$, $p = .46$. Blending condition had a significant main effect on the blend ratings, $F(1, 11222) = 458.67$, $p < .0001$. Fig. 3-3 shows that pairs presented in the concurrent condition were generally perceived to blend better than in the sequential condition where participants had to imagine the blends. As expected, the instrument pair had a significant main effect on blend ratings, $F(90, 11222) = 337.17$, $p < .0001$. Fig. 3-4 shows the blend ratings of different pairs averaged across the two blending conditions. This result reflects the intrinsic timbral qualities of different instruments which result in different degrees of perceived blend. Fig. 3-4 also suggests that pairs involving one percussive and one sustained instrument generally received much lower ratings, corroborating the findings in Lembke et al. (2019).

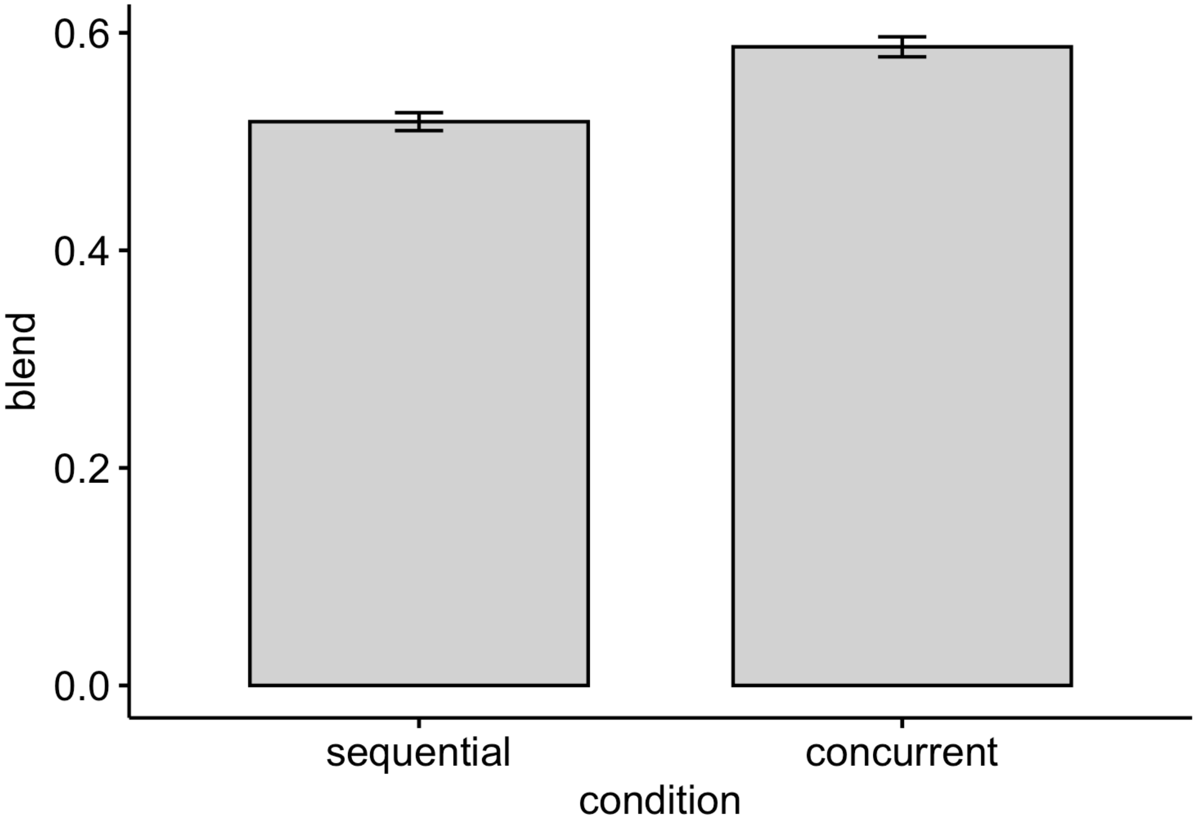


Figure 3-3. Main effect of blending condition on blend ratings (error bars correspond to the 95% confidence intervals [CI] of the means).

3.2.1.2 Two-way interaction effects

The interaction between blending condition and musical background was significant, $F(1, 11222) = 15.51$, $p < .0001$. The two groups rate blends similarly in the sequential condition. As can be seen in the interaction graph of these two factors (Fig. 3-5), for both musicians and non-musicians overall better blends were perceived in the concurrent condition than in the sequential condition. However, non-musicians tend to differentiate blends in the two conditions more than musicians do, with concurrent blends being overall rated prominently higher by non-musicians. There was a significant interaction effect between instrument pair and musical background, $F(90, 11222) = 2.06$, $p < .0001$. The interaction graph between these two factors (Fig. 3-6) suggests that musicians and non-musicians gave quite parallel ratings across all instrument pairs, albeit with varying degrees of divergence between the two groups for different pairs. Neither group gave consistently higher or lower ratings than the other group. The interaction between instrument pair and blending condition was also significant, $F(90, 11222) = 10.89$, $p < .0001$. Fig. 3-7 shows how the rating differences between the two blending conditions vary across instrument pairs.

To show which pairs were significantly different in the ratings between the two blending conditions, multiple pairwise comparisons were conducted focusing on the simple effects of condition × instrument pair interaction, i.e., the effect of blending conditions on blend ratings for each instrument pair. Shapiro-Wilk tests conducted on the difference scores of blend ratings between the two conditions (across the two musicianship groups) showed that several instrumental pairs had non-normally distributed difference scores.[6] Several extreme outliers[7] were also found in the difference scores for different instrument pairs by different participants. Therefore, the non-parametric alternative of a paired t-test—the Wilcoxon signed rank test—was used instead. Multiple tests were conducted between the two blending conditions for all 91 pairs of instruments with Holm correction for controlling family-wise error with multiple comparisons.

[6] $p < .05$ for 53 out of 91 pairs.

[7] Data points that are three times the interquartile range above or below the median.

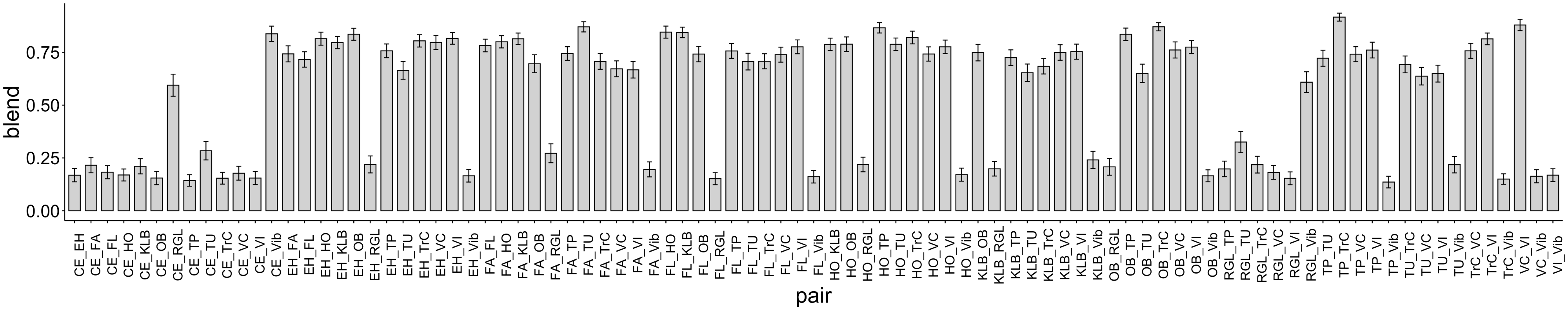


Figure 3-4. Main effect of instrument pair on blend ratings (error bars correspond to the 95% CI).

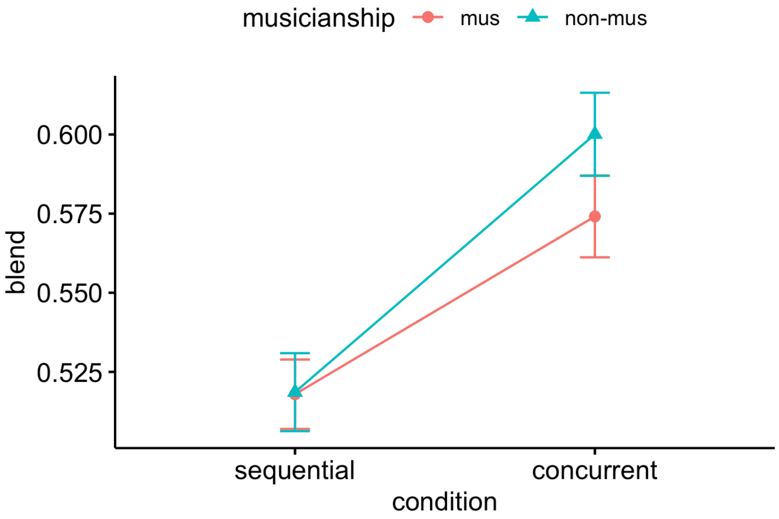


Figure 3-5. Interaction effect between blending condition and musical background on blend ratings (with 95% CI).

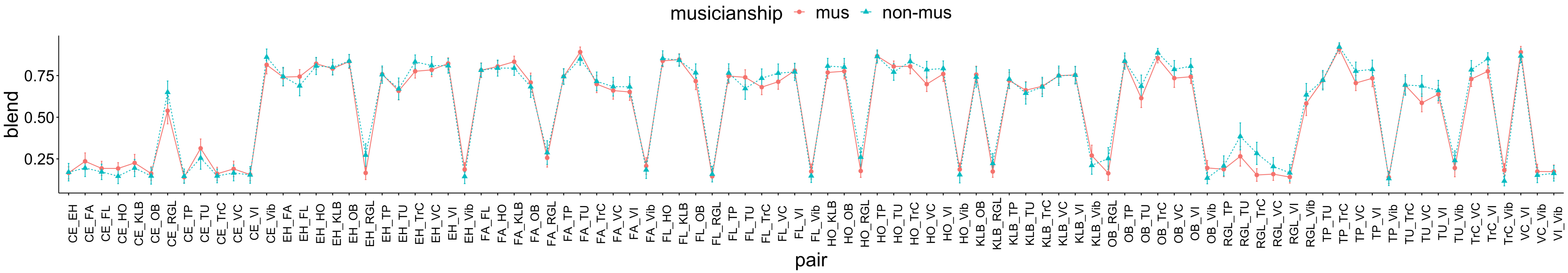


Figure 3-6. Interaction effect between instrument pair and musical background on blend ratings (with 95% CI).

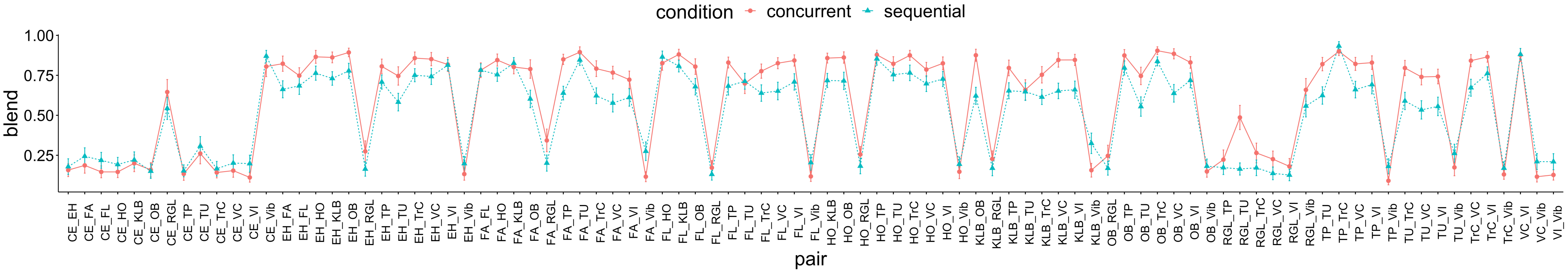


Figure 3-7. Interaction effect between blending condition and instrument pair on blend ratings (with 95% CI).

Fifty-one pairs were found to have significant conditional blend differences. Fig. 3-8 shows the median blend ratings in the two blending conditions for 36 instrument pairs out of the 51 pairs that had large effect sizes ($r \geq .5$), with corresponding p-values and effect sizes annotated in the figure. Both Fig. 3-7 and 3-8 showed that pairs containing two sustained instruments in general blended better when they were heard than when they were imagined (whether the difference was significant or not). Pairs containing one sustained instrument and one percussive instrument showed the opposite trend, except for tubular bell: all pairs involving tubular bell were rated somewhat higher (better blend) when the pairs were heard than when they were imagined.

To zoom in further on a few pairs that show bigger heard vs. imagined disparities across musicians and non-musicians, which may facilitate later discussion, multiple one-sample Wilcoxon signed-rank tests were conducted on the "absolute" difference scores between the two blending conditions for all 91 pairs where the null hypothesis was that the median of the "absolute" difference scores is smaller than a predefined positive threshold, here set to the average of medians of "absolute" conditional blend differences among all 91 pairs.[8]

[8] Difference scores were calculated by subtracting sequential ratings from concurrent ratings for each pair of instruments and for each participant. Thus, for each instrument pair, there was a difference score distribution constituted by 64 observations. As the focus is to find large conditional differences regardless of the sign, absolute difference scores were used. The tested threshold was taken as the average of all medians of such absolute differences.

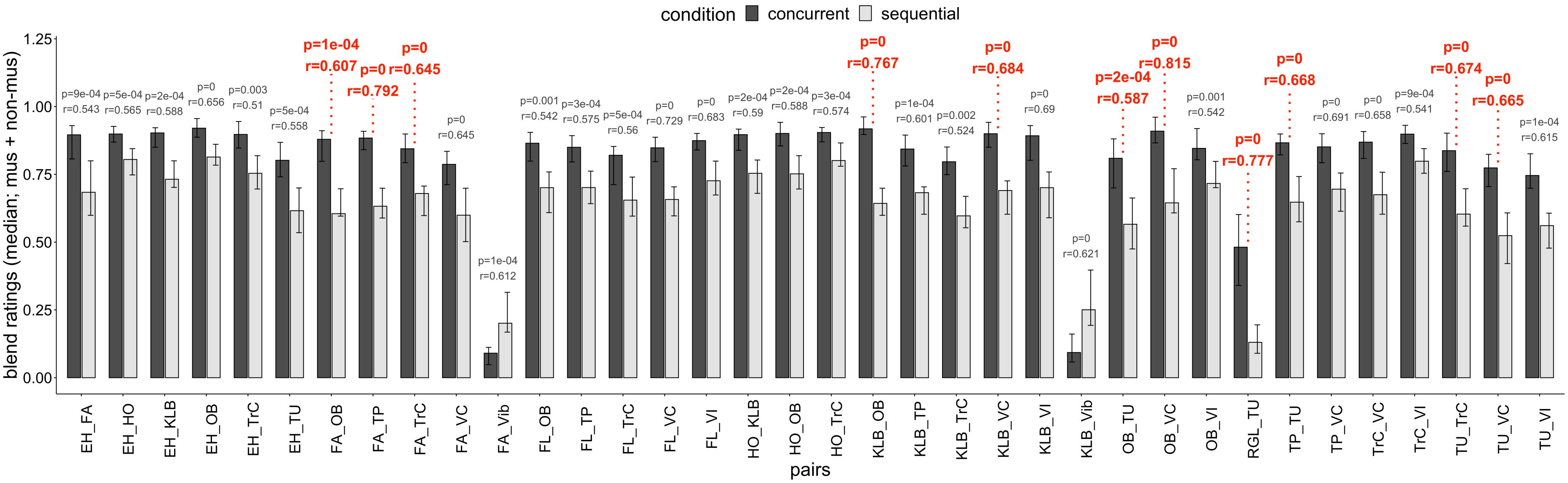


Figure 3-8. Median blend ratings (with 95% CI) in the two blending conditions for pairs having significantly different conditional blend differences with large effect sizes, annotated with corresponding p-values and effect sizes of the differences. Pairs with the largest conditional blend differences are highlighted with red annotations.

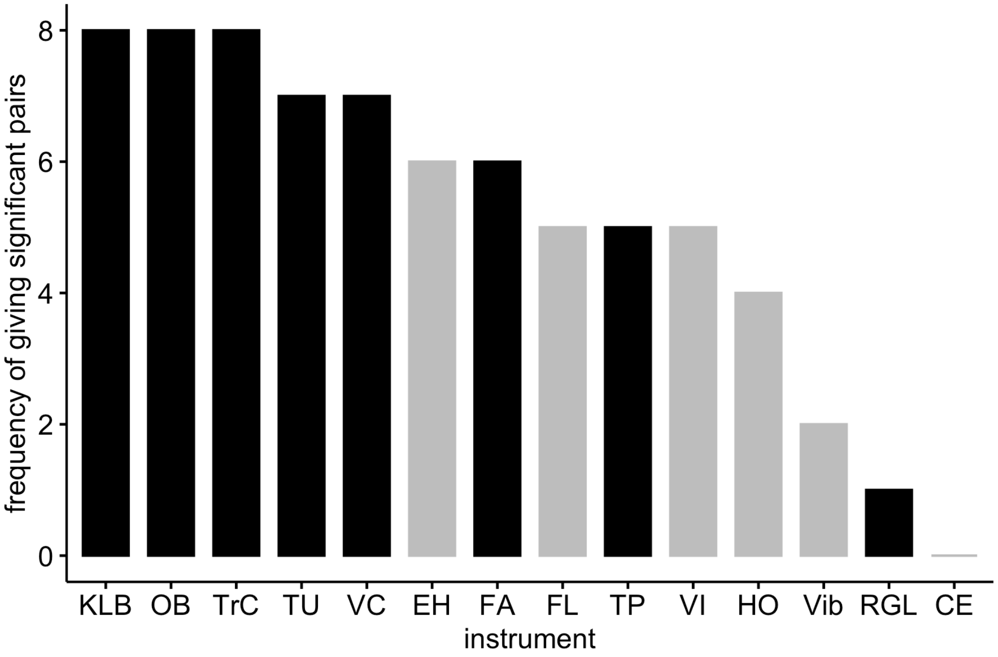


Figure 3-9. Frequencies of instruments forming pairs with significant conditional blend differences with large effect sizes. Instruments belonging to pairs with larger conditional differences are colored in black.

This positive threshold was set to sift out pairs that may have differences that are significantly different from zero but whose conditional blend differences are not large enough. The Holm correction was applied for multiple comparisons. Significantly different pairs in this test (subsequently called "pairs with larger conditional blend differences") were given red annotations in Fig. 3-8. To summarize the pairs shown in Fig. 3-8, the frequencies of individual instruments appearing in these pairs were plotted in Fig. 3-9, with instruments involved in pairs with larger conditional blend differences colored in black. These frequencies were correlated with the thirteen single-instrument acoustic features described in the previous chapter to see if there is a potential acoustic explanation for the instruments' tendency to form pairs that blend substantially differently when being heard and being imagined. Three features showed significant correlation ($p < .05$): for attack slope, $r = –.66$, $p = .010$; for the median of RMS energy, $r = .65$, $p = .011$; for the median of the second tristimulus value, $r = .60$, $p = .022$. These seem to suggest that when paired with other instruments, instrument sounds with slower attack, higher RMS energy, or higher harmonic energy within the second to fourth partial region are more likely to yield blends that are perceived substantially differently when they are heard compared with when they are imagined.

3.2.1.3 Three-way interaction effect

The three-way interaction between blending condition, instrument pair and musical background was significant, *F*(90, 11222) = 1.36, *p* = .014. As with the two-way interaction between blending condition and instrument pair, here the interpretation of this three-way interaction will focus on the same simple effects of blending condition within each instrument pair, additionally taking into the account of musical background (i.e., does musical background affect the conditional blend differences for a given instrument pair?). To do this, difference scores between the two blending conditions were calculated for each instrument pair and each participant. Multiple comparisons were made for all 91 pairs of instruments between the two groups of musicians and non-musicians, testing if the conditional differences were significantly different between the two groups. As Shapiro-Wilk tests suggested that many difference scores deviated significantly from normality[9] and several extreme outliers were found in the data, the non-parametric Wilcoxon rank sum test was used with Holm correction for multiple comparisons. Only two pairs gave significant results, EH_RGL (*W* = 242, *p* = .027, *r* = .45) and RGL_TP (*W* = 231, *p* = .015, *r* = .47).

It should be noted that the p-value correction method used might be so conservative that some significant differences of small effect sizes cannot be detected. Thus, a linear mixed model approach was also attempted to test these comparisons by setting appropriate contrasts to all the factor levels to be compared, bypassing the need to conduct multiple comparisons. As with the main ANOVA analysis, random intercepts were included in the model and were allowed to vary across participants. Degrees of freedom and t-statistics of model coefficients were calculated using Satterthwaites's method (Satterthwaite, 1946). Results showed that musical background had significant effects on the conditional blend differences for seven instrument pairs:[10] **EH_RGL** [*b*

[9] In total there are 2*91 = 182 groups of difference scores to be tested (91 instrument pairs; for each pair, the two musicianship groups were compared in terms of the conditional blend differences). 67 sample groups were statistically significant (*p* < .05).

[10] It is worth pointing out that differing results found with the linear mixed model vs. overall non-significant results with multiple Wilcoxon tests is also partly due to their different focuses on the data: the former tests differences between sample groups by using the mean as the representative statistic, whereas the latter focuses on the median. The non-normality of difference scores likely resulted in the contrasting results observed here. The presence of frequent outliers in the difference scores plausibly justifies using the median instead of the mean as the summarizing statistic. The Q-Q plot of the residuals of the fitted linear mixed model shows that distribution deviates from normality.

= –0.120, *t*(11220) = –3.906, *p* < .0001], **FA_RGL** [*b* = –0.068, *t*(11220) = –2.211, *p* = .027], **OB_RGL** [*b* = –0.089, *t*(11220) = –2.904, *p* = .004], **RGL_TP** [*b* = –0.146, *t*(11220) = –4.753, *p* < .0001], **RGL_TrC** [*b* = –0.069, *t*(11220) = –2.239, *p* = .025], **RGL_TU** [*b* = –0.132, *t*(11220) = –4.305, *p* < .0001], **TU_VC** [*b* = –0.064, *t*(11220) = –2.103, *p* = .035]. Fig. 3-10 shows the means of conditional blend differences of the seven significant pairs for musicians and non-musicians. All seven pairs documented larger conditional blend differences for non-musicians than for musicians, suggesting that for these specific pairs, musicians seem to be somewhat better at imagining blends that match the perception of actual heard blends as these pairs showed smaller conditional blend differences for them. Note the persistence of tubular bell (RGL), which appears in six of these pairs, including its pairing with English horn (EH_RGL), oboe (OB_RGL), and

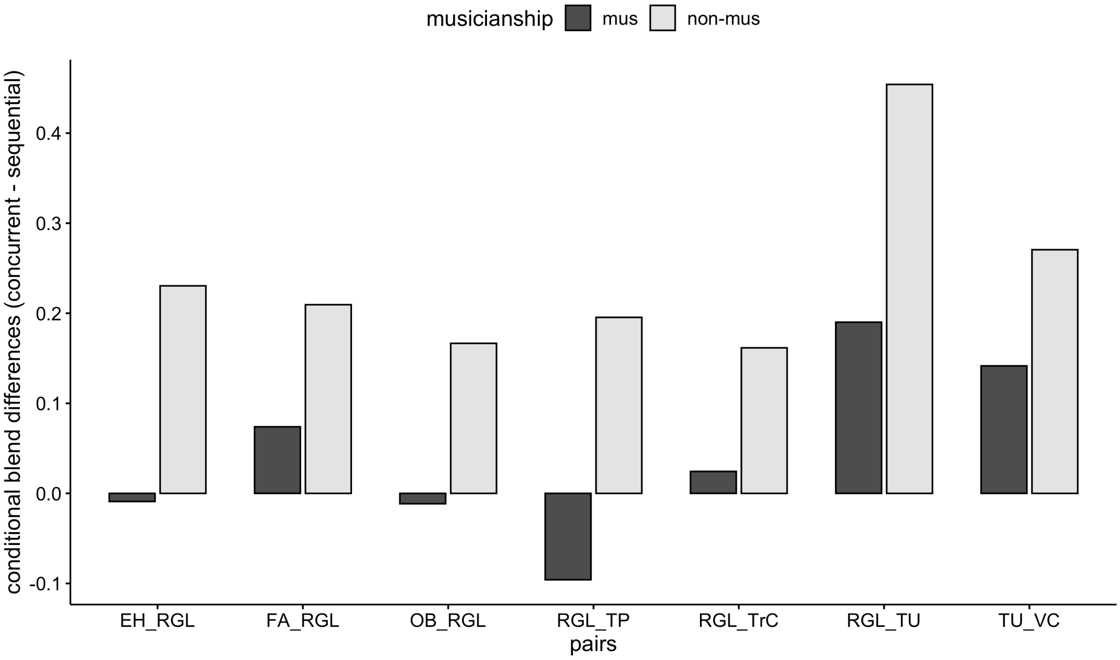


Figure 3-10. Means of conditional blend differences plotted for musicians and non-musicians separately. Pairs plotted in the graph are the ones having significantly different conditional blend differences between musicians and non-musicians.

Plotting the model residuals against fitted values also shows violation of homoscedasticity of error. Meanwhile, it has been shown that except for when influential outliers are present, violation of normality is not as serious as people had thought, and slight deviations from homoscedasticity are also manageable for hypothesis testing in linear models (Knief & Forstmeier, 2021). For the current data, as there are several outliers contributed randomly by many participants throughout different instrument pairs, and the deviation from homoscedasticity is quite prominent, the results of linear mixed model should be largely valid but also interpreted with caution.

trombone (RGL_TP), which are the only three that show some degree of opposite differences between musicians and non-musicians.

As a final concluding point, the validity of ANOVA analysis of the three factors using a linear mixed model will be briefly discussed. As with the model used above to decompose the three-way interaction, the mixed model used at the beginning of this section for omnibus tests of main effects and interaction effects had both non-normally distributed and (prominently) heteroscedastic residuals. The presence of outliers is almost consistent for blend ratings across all levels of the three factors, contributed by different participants, which makes it hard to eliminate these data points while keeping the size of the samples suitable for the analyses. When viewed together with the evidence of robustness of linear models against violation of normality and homoscedasticity assumptions (see footnote 10), the results of the linear mixed model used in the analyses seem to be reasonably acceptable, but the problems of multiple outliers and large deviation from homoscedasticity of model error should also be noted.

3.2.2 Family-wise blends

As mentioned at the beginning of section 3.2, the repeated measure of instrument pair can be abstracted into a higher level within-subject factor of family combination, which provides an alternative angle of investigating the instrumental factors in heard and imagined blends. In total, the fourteen instruments used in the experiment spanned four different families: woodwinds (WW), brass (BR), percussion (PERC), and strings (ST). Since all the individual instruments were paired with each other in the experiment, all four families were also completely paired with each other, giving ten family combinations as the different levels of this new factor. For each participant and blending condition, blend ratings corresponding to the same family combination type were averaged, giving a single rating for this family combination. Under this new organization of rating data, there is one between-subjects factor of musicianship (musicians vs. non-musicians) and two repeated measures of family combination and blending condition (concurrent vs. sequential). The standard ANOVA analysis for mixed design was used.

Before presenting the results, it is worth mentioning the normality assumption of ANOVA. The Q-Q plot of the ANOVA model's residuals showed deviation from normality at both the lowest and highest extremes. Some outliers were also found in blend ratings for combinations of the three factors from different participants; five of them were extreme outliers. At the same time, it has been shown that the Gaussian model is quite robust against the violation of normality assumption (Knief & Forstmeier, 2021). Given the magnitude of non-normality and outliers in the current model, the results of ANOVA should be safe to interpret.

3.2.2.1 Main effects

Levene's tests conducted between the two musicianship groups for all combinations of the levels of the two repeated measures showed that homogeneity of variance was violated for only one such combination ($p < .05$). Given that only one factorial combination violated this assumption and that ANOVA is fairly robust against this type of violation when sample sizes are equal (Field et al., 2012, p. 413), the results of ANOVA should still be considered valid. Similar to the analysis of instrument-wise blends, there was no main effect of musical background on blend ratings, $F(1, 62) < 1$.

For the main effect of family combination, Mauchly's test indicated that the sphericity assumption was violated, $W = 0.00004$, $p < .0001$. Therefore, degrees of freedom were corrected using Greenhouse–Geisser correction ($\hat{\varepsilon} = 0.306$). The main effect was significant, $F(2.76, 170.98) = 574.83$, $p < .0001$. Fig. 3-11 shows the means of family-wise blend ratings for all family combinations. The results were comparable to those listed in McAdams et al. (2016) (blends studied there were orchestral blends in longer contexts) where the combination of only string

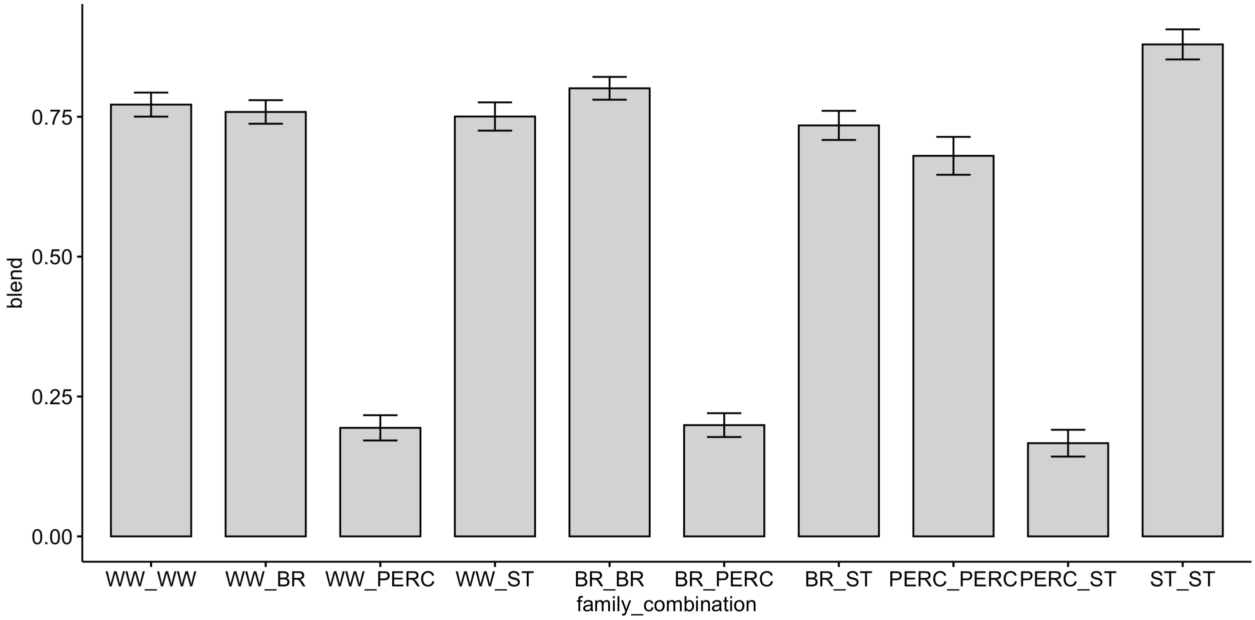


Figure 3-11. Main effect of family combination on family-wise blend ratings (with 95% CI).

instruments yielded the highest blend, and the pairing of percussion instruments with other families gave significantly lower ratings than other family combinations.

The main effect of blending conditions was also significant, $F(1, 62) = 44.66, p < .0001$. As with the results found in instrument-wise blends, blend was rated significantly higher in general when heard than when imagined.

3.2.2.2 Two-way interaction effects

The interaction effect between musical backgrounds and blending condition was only marginally significant, $F(1, 62) = 2.90, p = .093$. This contradicts the result given in section 3.2.1.2 where the interaction between these two factors was found to be significant. Although the two tests address the same effect, the sample sizes are different in the two ANOVA analyses as the blend ratings were averaged across different instrument pairs here. This, along with the fact that the two ANOVA analyses were calculated with different methods, could explain the contradicting results. As the unaveraged data analyzed in section 3.2.1.2 represent the actual responses given by participants in the experiment, it is probably more appropriate to interpret this interaction still as significant.

For the interaction effect between musical background and family combination, Mauchly's test showed a significant violation of sphericity, $W = 0.00004, p < .0001$. Greenhouse–Geisser correction ($\hat{\varepsilon} = 0.306$) was applied. The interaction effect was not significant, $F(2.76, 170.98) = 1.54, p = .209$. Fig. 3-12 shows the interaction graph of this effect, which suggests that musicians and non-musicians gave nearly identical blend ratings (averaged across two blending conditions) for all family combinations. When viewed together with Fig. 3-6 (note how the two lines representing pairwise blend ratings by the two groups intertwine and overlap well with each other), it seems that averaging ratings across the same family combination further eliminates the differences between musicians and non-musicians on the overall perception of blends.

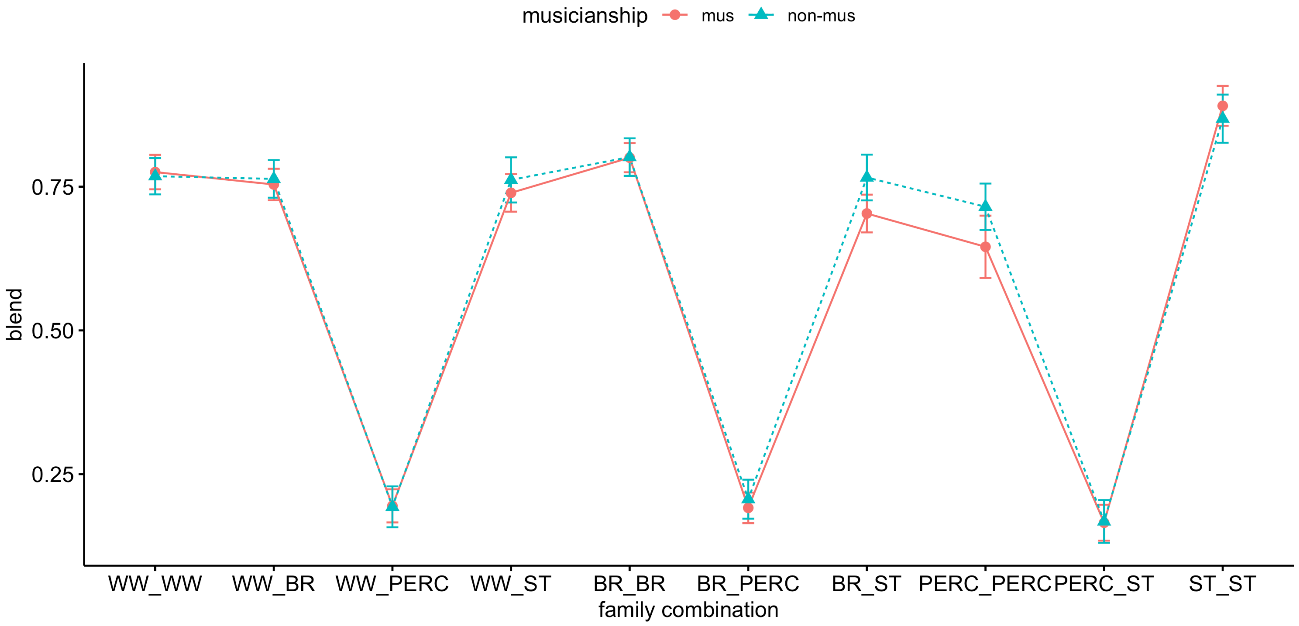


Figure 3-12. Interaction effect between musical background and family combination on family-wise blend ratings (with 95% CI).

For the interaction effect between family combination and blending condition, Mauchly's test showed a significant violation of sphericity, $W = 0.0084$, $p < .0001$. Greenhouse–Geisser correction ($\hat{\varepsilon} = 0.489$) was applied. The interaction effect was significant, $F(4.40, 272.91) = 19.93$, $p < .0001$. As with the analyses done with instrument-wise blends, here the main interest of interpreting this effect is to parse it into simple effects of blending condition on each level of family combination. Shapiro-Wilk tests conducted on the difference scores of blend ratings between the two conditions (across the two musicianship groups) showed that the difference scores in family combinations BR_PERC, BR_ST and ST_ST significantly deviated from normality ($p < .05$). Given the robustness of t-tests against non-normality when sample sizes are relatively large ($n > 30$) (Gravetter & Wallnau, 2008, p. 301), paired t-tests were used for multiple comparisons with Holm correction. Results are annotated in Fig. 3-13 with p-values and effect sizes. Significant results are marked in red. The combinations among brass, string and woodwind families, including their self-pairings (except for ST_ST), were all perceived to blend significantly higher when they were heard compared with in imagination. For combinations involving percussion, there were no significant conditional blend differences.

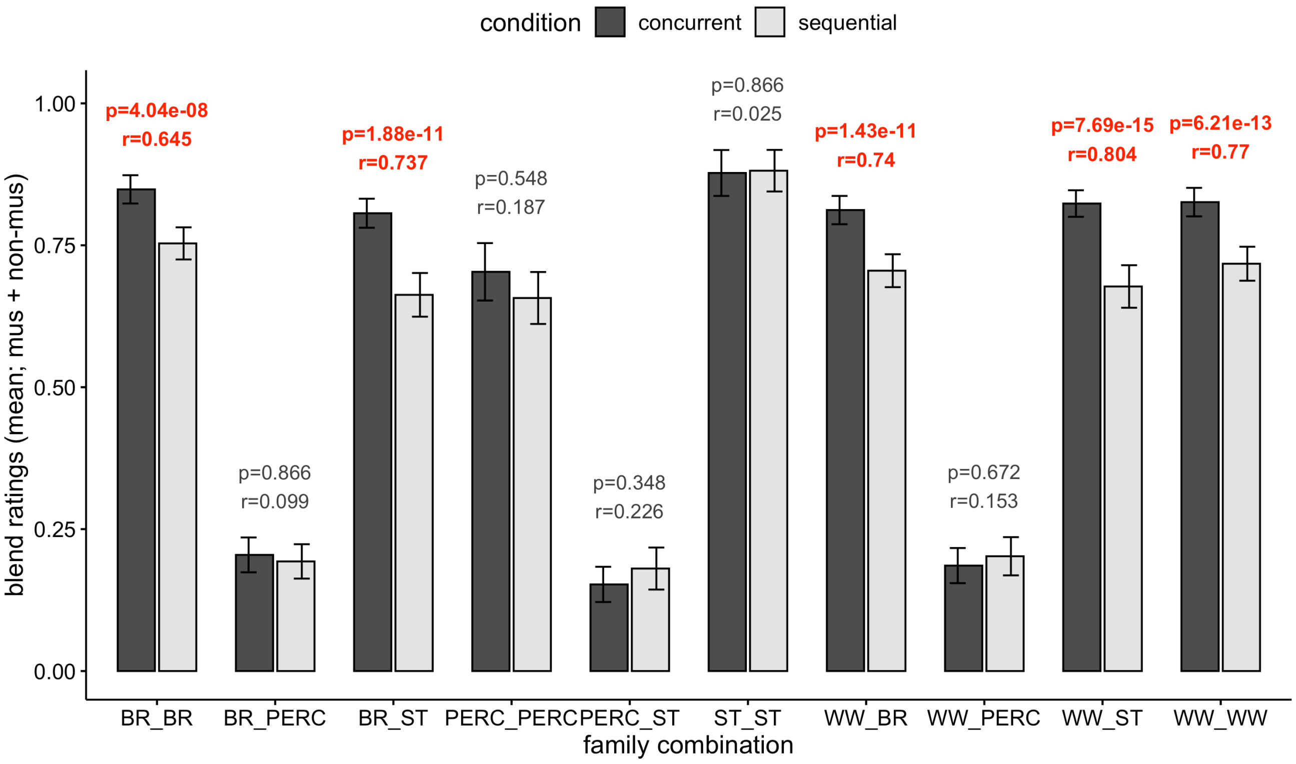


Figure 3-13. Mean blend ratings (with 95% CI) in the two blending conditions for all family combinations, annotated with p-values and effect sizes of the differences. Family combinations having significantly different conditional blend differences are highlighted with red annotations.

3.2.2.3 Three-way interaction effect

Mauchly's test was significant for the three-way interaction, W = 0.0084, p < .0001. Therefore, Greenhouse–Geisser correction ($\hat{\varepsilon}$ = 0.489) was applied. The effect was not significant, F(4.40, 272.91) = 1.78, p = .126. This suggests that the perception of heard vs. imagined blends generalized by instrument family combinations is similar between musicians and non-musicians.

3.3 Regression modeling of blend ratings

To understand how acoustic features of constituent instruments might contribute to blend ratings and, more importantly, how this picture might be different for the two blending conditions, stepwise multiple regression was conducted on the two sets of conditional blend ratings separately with 25 pairwise comparisons of features (listed in Table 2-1) as the entire scope of independent variables. The median of blend ratings for each instrument pair across the two musicianship groups was chosen as the dependent variable. Both independent and dependent variables were standardized before running the analyses, which gave standardized beta coefficients allowing easier comparisons between relative contributions of different features. Three types of stepwise regression directions were tried for the analyses: for the "forward" direction, the model was built from an intercept-only baseline by adding one predictor at a time that accounts for the most yet-unexplained variance in the outcome variables at each stage and improves the predictive power of the model indicated by a decrease in the Akaike information criterion (AIC). The selection process stops when AIC cannot be further improved; for the "backward" direction, a full model was built from the entire scope of predictors and redundant predictors were sequentially removed if the corresponding removal improved the AIC; for the "both" direction, the "forward" and "backward" were alternated, whereby each predictor was added and then any possibly redundant predictor was subsequently removed. For the modeling of each blending condition, the three models given by the three stepwise methods were compared using a 5-fold cross-validation. The model with the smallest root mean squared error (RMSE) was chosen as the final model for that blending condition. Results of the obtained models are presented below.

3.3.1 Blends in the concurrent condition

Table 3-1 lists the model coefficients, their associated p-values, and overall fit of the model, including the adjusted R^2, which is a metric of how well the regression model generalizes beyond the current sample. Plotting the model residuals showed that the normality assumption was met (as confirmed by the Q-Q plot), albeit with deviation from homoscedasticity. Thus, to obtain a more robust estimate of the significance of the model predictors, bias-corrected-and-accelerated (BCa) bootstrap confidence intervals (95%; 2000 bootstrap samples) were also calculated. If the confidence interval didn't cross zero, then the significance of the corresponding predictor could be further confirmed. Predictors with bootstrap-confirmed significance are in boldface in the table.

Overall, the model accounted for around 87% of the variance in the blend ratings. Seven pairwise predictors had significant beta coefficients as confirmed by both the original regression and subsequent bootstrap confidence intervals (listed according to descending standardized beta estimates): **TriCor**, **Tri2_Med_diff**, **Tri2_Med_sum**, **Noisiness_Med_diff**, **Tri3_Med_diff**, **RMSErg_Med_diff, AttSlope_diff**. All significant predictors are negatively correlated with the blend ratings except for Tri2_Med_sum.

Table 3-1. Final model obtained with concurrent blend ratings. Beta estimates are standardized. Standard errors and p-values of beta estimates are attached, along with the overall fit of the model. Significant predictors whose bootstrap confidence intervals (95%) don't cross zero are in boldface.

	Beta estimate	**SE**	***p***
TriCor	–0.528	0.079	< .0001
Tri2_Med_diff	–0.463	0.062	< .0001
Tri2_Med_sum	0.311	0.059	< .0001
Noisiness_Med_diff	–0.249	0.055	< .0001
Tri3_Med_diff	–0.242	0.074	.002
RMSErg_Med_diff	–0.200	0.048	< .0001
Tri1_Med_diff	–0.199	0.066	.004
AttSlope_diff	–0.150	0.057	.009
InHarm_Med_sum	–0.103	0.060	.090
Model multiple R^2: 0.8705	Adjusted R^2: 0.8561		

3.3.2 Blends in the sequential condition

Table 3-2 summarizes the model derived for blend ratings in the sequential condition. Similar to the case of modeling concurrent blends, bootstrap confidence intervals were calculated for the model coefficients as residuals showed both deviation from normality and homoscedasticity. The final model explains around 88% of the variance in the blend ratings. Eight pairwise predictors had significant contributions as confirmed by the original regression and bootstrap confidence intervals (listed according to descending standardized beta estimates): **TriCor**, **Tri2_Med_diff**, **Tri2_Med_sum, Tri3_Med_diff, Noisiness_Med_diff, AttSlope_diff, InHarm_Med_sum, RMSErg_Med_diff**. The selected predictors are the same with those obtained in the model for concurrent blends, except for InHarm_Med_sum which is only present in the sequential model. All predictors are negatively correlated with the blend ratings except for Tri2_Med_sum.

Table 3-2. Final model obtained with sequential blend ratings. Beta estimates are standardized. Standard errors and p-values of beta estimates are attached, along with the overall fit of the model. Significant predictors whose bootstrap confidence intervals (95%) don't cross zero are in boldface.

	Beta estimate	**SE**	***p***
TriCor	–0.375	0.082	< .0001
Tri2_Med_diff	–0.346	0.063	< .0001
Tri2_Med_sum	0.338	0.067	< .0001
Tri3_Med_diff	–0.305	0.073	< .0001
Noisiness_Med_diff	–0.266	0.055	< .0001
AttSlope_diff	–0.258	0.101	.013
AttSlope_sum	0.251	0.110	.025
InHarm_Med_sum	–0.247	0.067	.0004
RMSErg_Med_diff	–0.244	0.055	< .0001
Tri1_Med_diff	–0.198	0.077	.012
SpecCent_IQR_sum	0.140	0.071	.052
SpecCent_Med_sum	–0.093	0.068	.172
Model multiple R^2: 0.8805	Adjusted R^2: 0.8622		

3.4 Multidimensional scaling of blend ratings

Multidimensional scaling (MDS) is a useful tool for visualizing proximity data and uncovering latent dimensions of judgement (Borg et al., 2018). Blend ratings can also be conceived as a specific type of similarity data with higher blend ratings intuitively suggesting a kind of affinity between the associated instruments, as done in studies by Sandell (1989; 1991). The derived multidimensional space, therefore, can be understood as a "blend space", where closely spaced instruments are perceived to blend better than instruments that are farther apart. For the current study, such spaces offer a straightforward way of comparing how instruments are perceived to blend in different conditions. Correlation between the MDS configurations and acoustic features can also help interpret the underlying behaviors of blend ratings in the heard and imagined conditions.

For each condition, the median of blend ratings for each instrument pair across the two musicianship groups was treated as the dissimilarity measure between the associated instruments. Similarity measures were obtained by subtracting dissimilarity data from one, which were then modeled using MDS algorithms implemented by the R package "smacof" (de Leeuw & Mair, 2009). Interval MDS was chosen as the model for the analyses which attempts to preserve the differences among the dissimilarity data.[11] Results of the MDS solutions are presented below.

3.4.1 MDS of blends in the concurrent condition

The classical Torgerson solution was chosen as the default initial configuration for the MDS program. The resulting configuration has a Stress value of 0.068. The significance of the Stress was tested using a permutation test (Mair et al., 2016) where the original dissimilarity data were permuted randomly many times (in the current analysis, $n = 100$) and subjected to MDS. The stress value of the target MDS solution was then located within the distribution of Stress values obtained with permuted data and its p-value can then be calculated as a measure of the significance of the target Stress value. For the Stress (0.068) of the current MDS solution, $p < .001$. The suitability of the current MDS solution was also confirmed by refitting the data using different random initial configurations ($n = 1000$) and comparing the configuration of the best solution with the current

[11] The less stringent MDS variant, the ordinal MDS (which only attempts to preserve the ordering of dissimilarities data in the scaling configuration), was also tried in the analyses. However, they almost always yielded degenerate solutions, with percussion instruments and other instruments being clustered together separately. Thus, the interval MDS was adopted in the analyses.

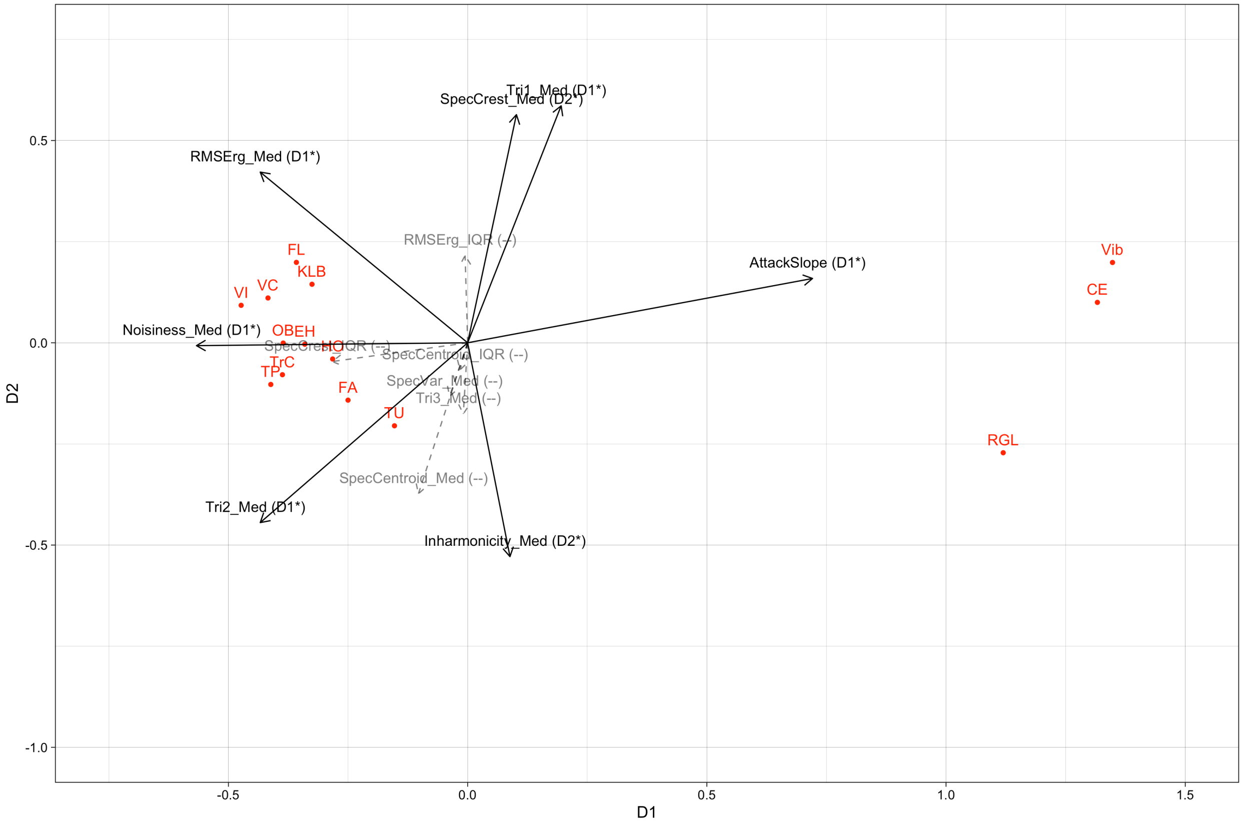


Figure 3-14. MDS solution obtained with concurrent blend ratings, along with projections of acoustic correlates. Non-significant projections are drawn with dashed grey lines. For correlates only having one MDS dimension with significant contribution to the projections, the significant dimension is appended to correlates' names with "*". In the case of no significant contributions from either dimension, "(--)" are appended to the correlates' names.

MDS solution. After matching the two with Procrustean transformation[12], they essentially showed the same configuration. Hence, the default MDS solution was adopted and plotted in Fig. 3-14. A prominent feature is that instruments were separated in two groups based on whether they are percussion instruments or not.

The correlation between dissimilarities data and distances in the MDS configuration is very large, $r = .995$, $p < .0001$, demonstrating a good fit of the MDS solution. Fig. 3-15 visualizes the stress-per-point for each instrument in the MDS solution, quantifying how much incongruence

[12] Procrustean transformations are similarity transformations that preserve the structure of MDS configurations by rotation, reflection, translation, and size scaling, which can be used to match an MDS configuration to another optimally by eliminating their non-structural differences (Borg et al., 2018, p. 84).

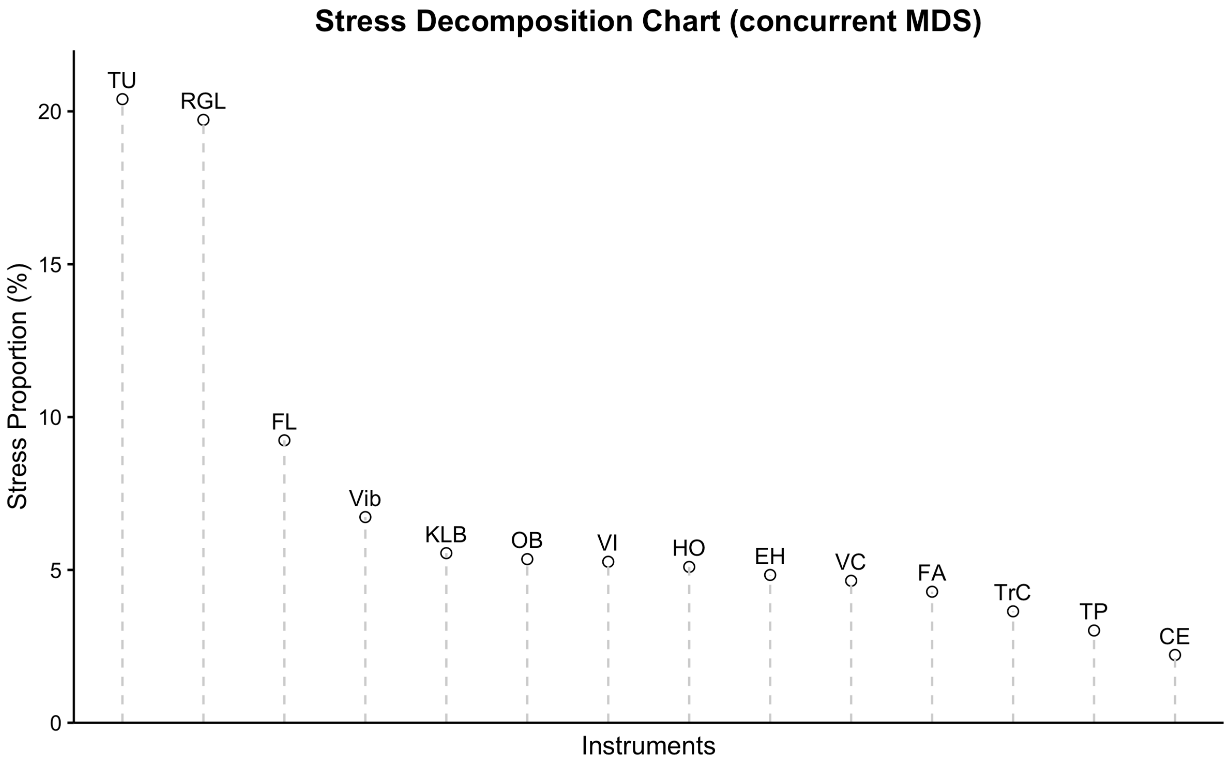


Figure 3-15. Stress contributions from individual instruments for the MDS scaling of concurrent blend ratings.

between fitted distances and corresponding dissimilarities there is for the associated instrument. Instruments with high stress contribution means that participants might have difficulties rating the blends associated with that instrument. For the current case, tuba and tubular bell are the two instruments that seem to elicit such difficulties. To figure out which pairs exactly contributed the most to the stress-per-point of individual instruments, the pairwise representation errors (i.e., the square of difference between dissimilarity rating and corresponding pairwise distance in the MDS configuration) were plotted as heatmap in Fig. 3-16. It is clear from the heatmap that the singularity of tubular bell and tuba is directly associated with their pairing together, that is, participants seemed to have difficulty or inconsistent conception when rating their blend.

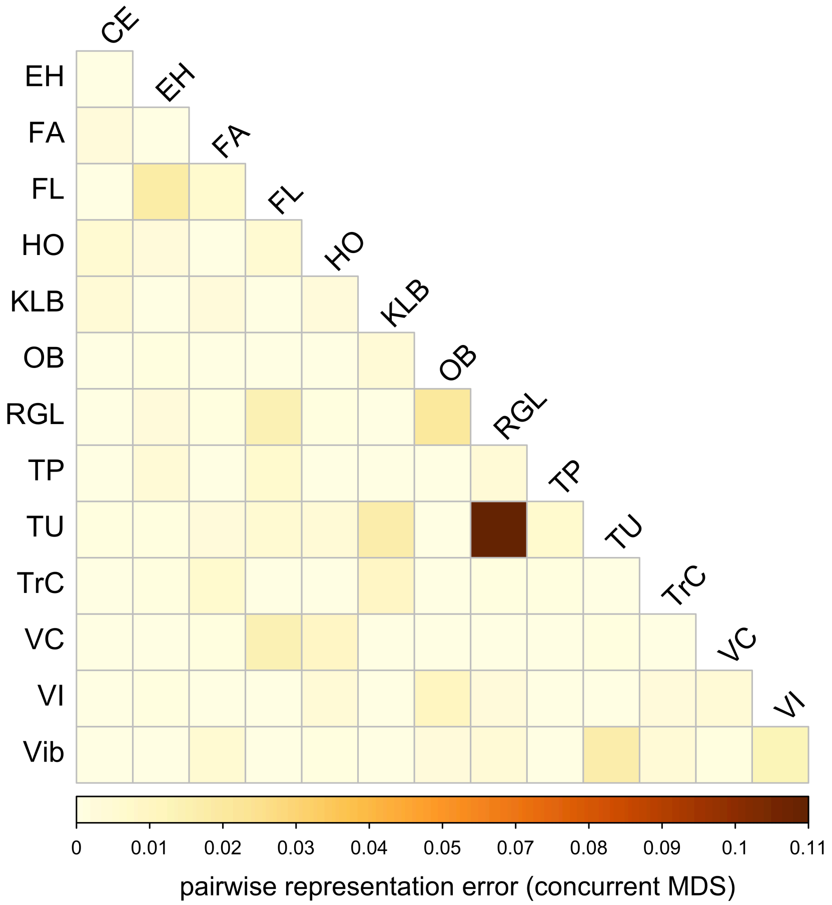


Figure 3-16. Representation errors for each pair of instruments in the MDS solution of concurrent blend ratings. Darker colors correspond to larger representation errors.

3.4.2 MDS of blends in the sequential condition

Similar to the modeling of concurrent blend ratings, the default Torgerson solution was chosen as the initial configuration of the multidimensional scaling of sequential blend data. Alternative initial configurations were also explored by scaling the data with different random configurations (*n* = 1000) and choosing the best solution with the smallest Stress value. The two solutions showed essentially the same configuration after a Procrustean matching process. Thus, the default MDS solution was adopted. The solution has a Stress value of 0.091, which was tested to be significant (*p* < .001) with subsequent permutation test. To allow better comparisons between the MDS results of the two different blending conditions, the MDS configuration of sequential

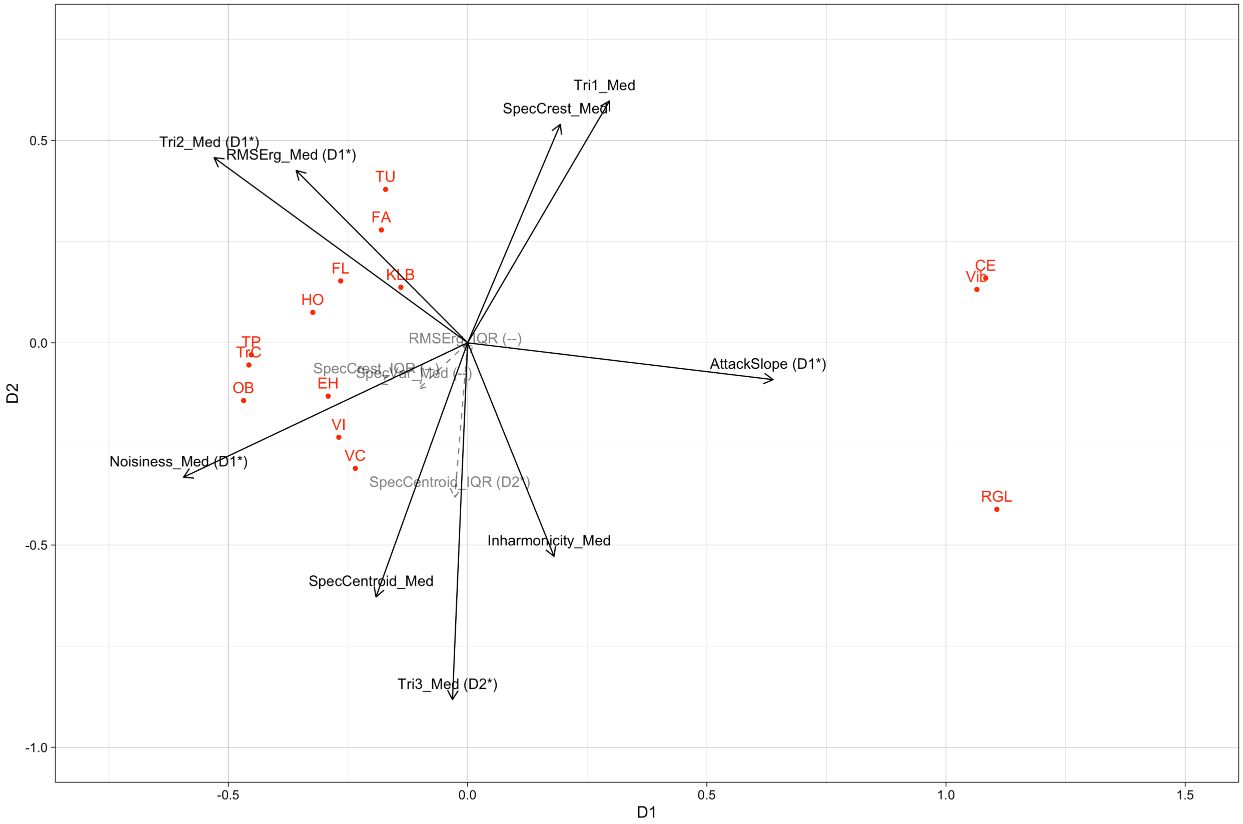


Figure 3-17. MDS solution obtained with sequential blend ratings (optimally matched with the MDS configuration obtained with concurrent blend ratings via Procrustean transformations), along with projections of acoustic correlates.

blends was matched optimally to that of the concurrent blends with Procrustean transformations. The transformed MDS solution is shown in Fig. 3-17, plotted with the same axial limits in Fig. 3-14. Comparing to the "blend space" in the concurrent condition, the same grouping based on percussiveness is observed in the sequential condition. Interestingly, non-percussive instruments seem to occupy a much larger space and are more scattered.

The dissimilarities data correlated very well with corresponding distances in the MDS configuration, $r = .989$, $p < .0001$. Fig. 3-18 plots the stress-per-point for each instrument in the MDS space. Compared with the results in the concurrent MDS, there are no such prominent instruments that contributed singularly to the overall Stress. It seems to suggest that the Stress is more widely distributed across different instruments and different pairs, with trombone, tubular bell, and trumpet contributing relatively more than the other instruments. This can be confirmed in Fig. 3-19 where pairwise representation errors are visualized with a heatmap. The magnitude of

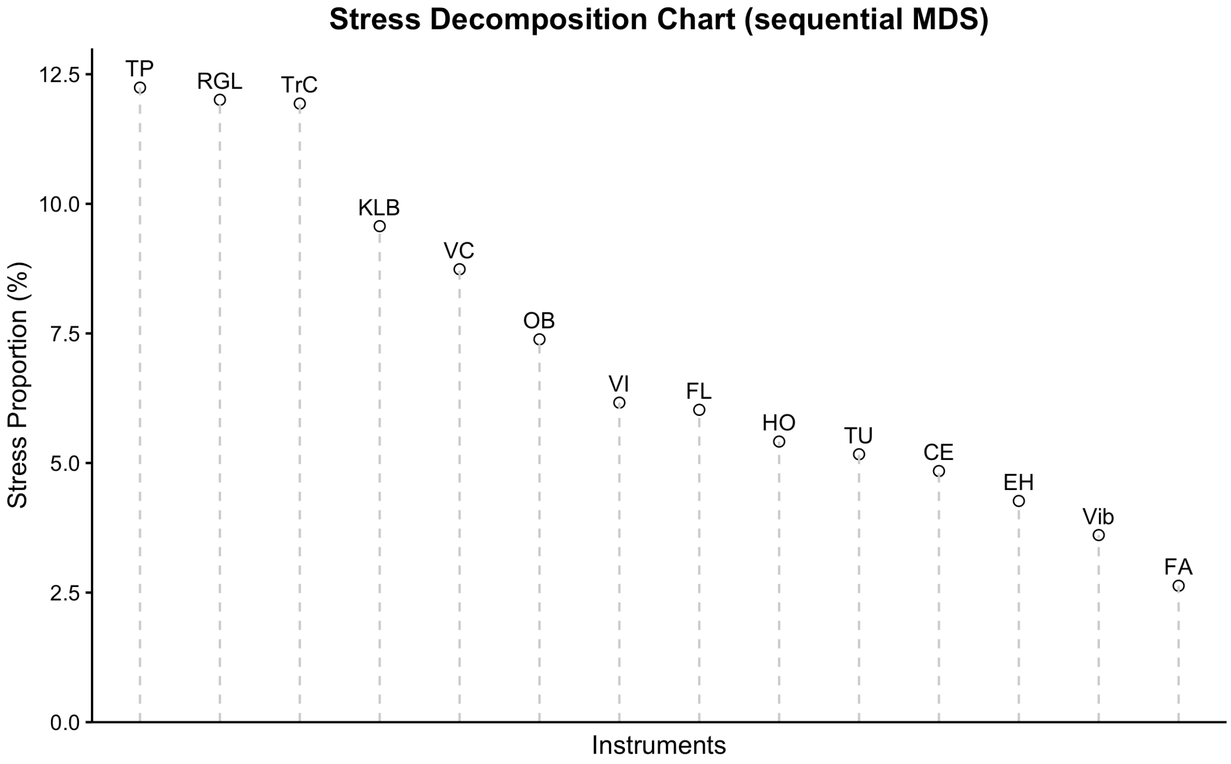


Figure 3-18. Stress contributions from individual instruments for the MDS scaling of sequential blend ratings.

errors is not as big as the singular cases shown in the concurrent MDS, but there are more pairs with relatively large representation errors within the context of MDS of sequential blends. Prominent pairs are: FL_TrC, HO_TP, KLB_TU, KLB_TrC, OB_VC, RGL_TP, RGL_VC, TP_TrC. Overall, it seems that participants had difficulties or inconsistency with rating imagined blends for a wider range of instrument pairs compared with when rating the heard blends. Instruments with high spectral centroid and/or rich higher partials (such as trumpet, trombone, oboe, etc.) or complex harmonic structure (in the case of tubular bell) seem to be especially "problematic" when participants imagined their blends with other instruments.

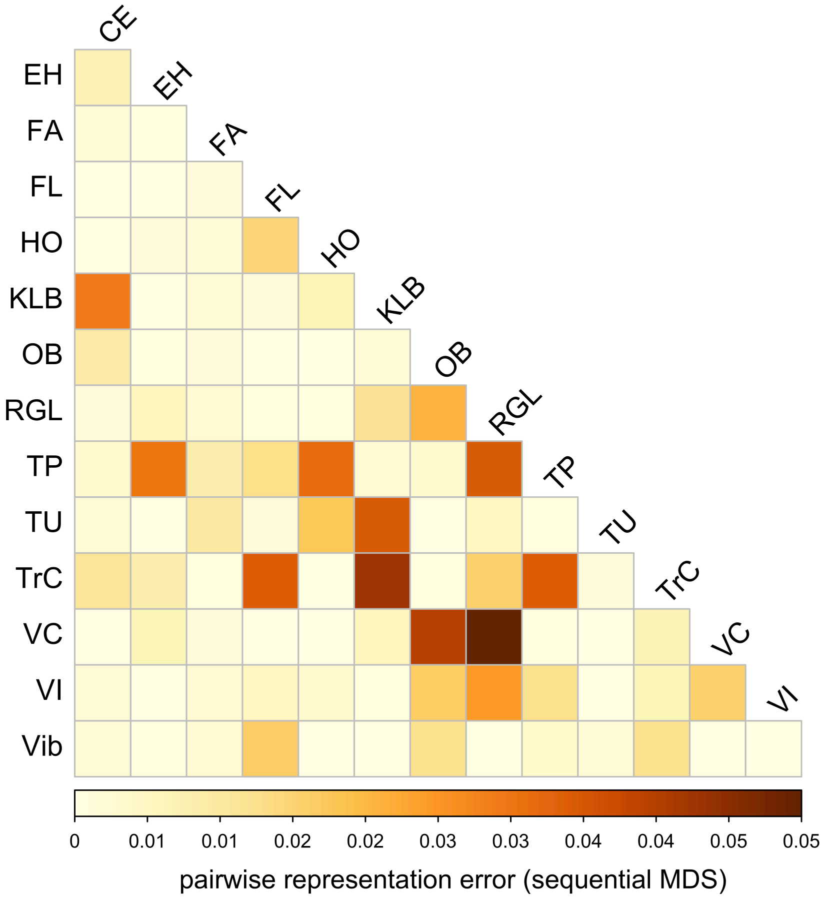


Figure 3-19. Representation errors for each pair of instruments in the MDS solution of sequential blend ratings. Darker colors correspond to larger representation errors.

3.4.3 Projection of acoustic correlates

Each of the thirteen acoustic features of single instruments described in section 2.2.1 was correlated using linear multiple regression with the coordinates of instruments in the MDS configuration for the concurrent and sequential conditions separately (i.e., using X and Y coordinates as independent variables, and acoustic correlates as dependent variables). Acoustic features were all standardized before regression so that the distribution of an acoustic feature was centered around zero with a standard deviation of one. Figs. 3-14 and 3-17 also visualize the correlation results, showing the projection of acoustic correlates onto the "blend space". The relative lengths of projection arrows correspond to the multiple R^2 of the respective regression results. Non-significant regression results are colored in light grey with dashed lines. The significance of individual contribution from the first and second MDS dimensions to the projection fit is indicated following the acoustic correlates' names (see the caption of Fig. 3-14).

3.4.3.1 Acoustic correlates in the concurrent condition

Table 3-3 shows some of the regression results of acoustic correlates in the concurrent MDS space. Correlates with significant regression results are in boldface. Overall, seven acoustic correlates showed significant correlation with the MDS configuration (listed according to descending R^2): **AttSlope**, **Tri2_Med**, **Tri1_Med**, **RMSErg_Med**, **SpecCrest_Med**, **Noisiness_Med**, **InHarm_Med**. Out of these correlates, AttSlope, Tri2_Med, RMSErg_Med and Noisiness_Med also appeared significant as pairwise features in the regression model of raw blend ratings in the concurrent condition. Attack slope very prominently explains the two groups of instruments separated across the first dimension shown in the MDS. There are a few other meaningful projections that might help to explain the distribution of instruments in the MDS space from different perspectives. Almost perpendicular to the projection of attack slope, InHarm_Med seems to be a meaningful direction that explains some amount of variance of the non-percussive instruments, though overall the fit of this projection is not as good as the other projections. Compared to the first dimension, the non-percussive instruments have a slightly larger spread on the second dimension of the MDS space, which contributed significantly only to the projection of InHarm_Med and SpecCrest_Med. It is tempting to ascribe this dimension to the "tonality" of the sound as inharmonicity and spectral crest both correspond more or less to the degree of harmonicity from two opposite directions (which is the case in the projection plot). Nevertheless, as the fit is only moderate, further explanation of this portion of variance may still require other unexplored features.

Table 3-3. Fit of projections and p-values of acoustic correlates in the MDS space obtained with concurrent blend ratings. Significant correlates are in boldface.

	$\boldsymbol{R^2}$	**Adjusted** $\boldsymbol{R^2}$	$\boldsymbol{p}$
AttSlope	0.671	0.611	0.0022
Tri2_Med	0.564	0.485	0.0103
Tri1_Med	0.561	0.482	0.0107
RMSErg_Med	0.550	0.468	0.0124
SpecCrest_Med	0.521	0.434	0.0175
Noisiness_Med	0.515	0.427	0.0186
InHarm_Med	0.487	0.394	0.0254
SpecCent_Med	0.352	0.234	0.0920
SpecCrest_IQR	0.261	0.126	0.1899
RMSErg_IQR	0.196	0.050	0.3008
Tri3_Med	0.160	0.007	0.3835
SpecVar_Med	0.125	–0.034	0.4794
SpecCent_IQR	0.063	–0.108	0.7008

3.4.3.2 Acoustic correlates in the sequential condition

Table 3-4 summarizes the regression results of projecting acoustic correlates onto the sequential MDS space. Nine correlates showed significant correlation with the MDS configuration (listed according to descending R^2): **Tri3_Med**, **Tri2_Med**, **Noisiness_Med**, **Tri1_Med**, **SpecCent_Med**, **AttSlope**, **SpecCrest_Med**, **InHarm_Med**, **RMSErg_Med**.

Tri3_Med, Tri2_Med, Noisiness_Med, AttSlope, InHarm_Med and RMSErg_Med also appeared significant as pairwise features in the regression model of raw blend ratings in the sequential condition. Comparing these results to those obtained with concurrent blend ratings, SpecCent_Med and Tri3_Med emerge as having a significant correlation with the sequential MDS space but not for the concurrent space. Both correlates appear to be meaningful directions that explain the large variance of instruments (primarily along the second dimension). Some of the other correlates that are also significant in the concurrent MDS space remain largely invariant in terms of their projected directions in the sequential MDS space, for example SpecCrest_Med and Tri1_Med, which show better projection fit here and also seem to explain well the spatial variance along the second dimension. Attack slope still pertains to the segregation between percussions and non-percussive instruments. Interestingly, Tri2_Med and Noisiness_Med swap their directions when comparing the two MDS spaces.

Table 3-4. Fit of projections and p-values of acoustic correlates in the MDS space obtained with sequential blend ratings. Significant correlates are in boldface.

	$\boldsymbol{R^2}$	**Adjusted** $\boldsymbol{R^2}$	$\boldsymbol{p}$
Tri3_Med	0.802	0.766	0.0001
Tri2_Med	0.637	0.570	0.0038
Noisiness_Med	0.619	0.549	0.0050
Tri1_Med	0.607	0.535	0.0059
SpecCent_Med	0.597	0.524	0.0067
AttSlope	0.586	0.511	0.0078
SpecCrest_Med	0.521	0.434	0.0174
InHarm_Med	0.507	0.418	0.0204
RMSErg_Med	0.506	0.416	0.0207
SpecCent_IQR	0.349	0.230	0.0946
SpecCrest_IQR	0.189	0.041	0.3166
SpecVar_Med	0.139	–0.018	0.4393
RMSErg_IQR	0.026	–0.151	0.8654

Chapter 4
Discussion

The current study aims to uncover how blends are perceived in imagination compared with when the blends are actually heard, whether musical backgrounds have an effect on this, and finally how acoustic features might contribute to the two different types of blends differently. The four sets of analyses were set to answer these questions, sometimes with overlapping focuses but from different perspectives: 1) correlation between blend ratings in the two conditions was conducted to show the general similarities between imagined and heard blends; 2) ANOVA was conducted with two different organizations of data, i.e., instrument-wise and family-wise blends, to decompose the effects of musical background, instrumental specificity, and blending condition. The two organizations of data provide alternative views at the instrumental aspect of blends, with family-wise organization abstracting the pairwise choices into family-wise combinations that give a more generalized account of the instrumental factor. Compared with the overall correlation done in the first step, ANOVA also provides a more zoomed-in description of how specific instrument pairs or family combinations are perceived to blend differently under the two conditions by people with different musical backgrounds; 3) regression modeling of blends with acoustic correlates was conducted to show how the perception of blends are affected by the interaction of different acoustic features of the paired instruments, and how these features might be used differently in the two blending conditions; 4) multidimensional scaling was used to visualize the "blend spaces" of instruments in the two blending conditions where acoustic correlates were projected, which provides another perspective to re-examine the acoustic modeling done in the third step (as MDS provides a more generic abstraction of the blend ratings). Discussion of the results of these analyses will be formed around the two research questions proposed in section 1.3.3, within which more detailed aspects will be listed pertaining to the specific analyses.

4.1 Congruence and incongruence between heard and imagined blends

Despite local differences observed in specific instrument pairs, imagined and heard blends are largely parallel to each other (see Fig. 3-1 and 3-2 for the overall contours of two types of blend ratings) for both musicians and non-musicians, which can be confirmed also by the large correlation between blend ratings in the two conditions. Thus, it seems that the mental image of blending two instruments is to some extent comparable to that of hearing the instruments, which should be a reasonable extrapolation of previously mentioned evidence of the authenticity of imagining single instrument's timbre. The large picture holds true for both types of blends: pairs containing one percussive and one sustained instrument blend universally worse than all the other pairs.

Within this overall parallelism, different instrument pairs very clearly exhibit different properties that induce various degrees of difference between the two types of perceived blends, including the sign of the difference. Regarding the general difference between heard and imagined blends, almost all participants reported that many pairs blended much better when they actually heard them being played compared to what they previously imagined, which is confirmed by the significant main effect of blending condition. This global pattern holds true when inspecting the data from musicians and non-musicians separately. It seems plausible that despite the ability for people to imagine two instruments blending together, the mental image of this interaction still deviates from the real perception of actual sounding blends, and in the case of the current experiment the imagination seems to be on average more "conservative" than the real blends[13]. A possible explanation of this, which has been confirmed by some participants' feedback, is partly related to the "constructive" nature of mentally blending two existing instruments: operating on the mental images of two distinct instruments itself assumes *a priori* the "twoness" of the composite sound[14].

[13] It should be emphasized that this "conservativeness" is an averaged description. A more meaningful interpretation will be discussed as we consider the specificities of instrument pairs. Nevertheless, it seems to echo participants' general feedback well.

[14] Even the operation of superimposing two instruments in mind turned out to be a bit problematic for some participants as some of them reported a certain level of general difficulty when imagining blends.

The exact nature of differences between blends in the two experimental conditions is largely dependent on the specific instrument pairs in question. As shown in Fig. 3-7, blends of two sustained instruments were generally “underestimated” in imagination, and blends of one percussive and one sustained instrument were generally “overestimated” in imagination. As these two categories of blends also generally received ratings that are somewhat clustered at the higher and lower regions, respectively, on the rating scale (i.e., pairs of sustained instruments are generally better blenders than pairs mixing percussive and sustained instruments), it seems to suggest that imagined blends in general tend to be rated more conservatively. Note in Fig. 3-8 that most of the pairs with significant conditional blend differences are pairs of two sustained instruments, which were perceived to blend better in reality. Only two pairs of sustained-percussive instruments have significant conditional blend differences with better blends in imagination. None of the two pairs were considered to have “larger” conditional blend differences as tested in section 3.2.1.2. It seems plausible that for instruments with sharp attack contrasts, blends were relatively easy to be conceived in imagination and rated because of the stark difference of attack profiles as a signifier of non-blend. Thus, they were in general given low ratings, which were not too different from the perception of hearing the blends. For instruments without big attack contrasts, blends were harder to conceive and evaluate in imagination, possibly because of the generic difficulty of superimposing two instruments’ images in imagination and the intrinsic “twoness” of the imagination task, as mentioned previously.

Regarding the innate difficulty of imagining blends, as hypothesized in section 1.3.3, the action of retaining two instruments in memory and subsequently superimposing them internally requires a certain amount of attention resources which might detract from the accuracy of the blend’s representation. Some participants reported the “blurry” nature of imagined blends, where it was possible to imagine the blend of two instruments but usually only a rough image (compared with the direct perception of blends in the concurrent condition). Thus, participants were less confident to rate the imagined blends as either really well-blended or badly blended. This is also reflected in the distribution of ratings: for the concurrent condition, blends are more populated near both ends of the scale; for the sequential condition, blends are more spread across the entire scale. Another possible explanation of concurrent blend ratings being more polarized than sequential blend ratings is linked to the ordering of the experimental conditions. The imagination task (and its associated difficulty and “indirectness”) in the first part likely conditioned and heightened participants' awareness and expectation of what an actual good or bad blend sounds like. The exposure to real blends in the second part thus confronted participants with fresh stimuli that were easily accessible and amenable to evaluation (compared with the more indirect "path" to imagined blends). This augmented state of realization of aspects of blends (e.g., "I didn't expect this pair to blend that well/badly until I heard it", which was common feedback from participants) possibly encouraged participants to rate with more confidence that led to more polarized ratings.

The previous observation of pairs involving percussion instruments, however, is not the case for tubular bell, as all its pairs were perceived to blend somewhat better in reality than in imagination (whether the differences are significant or not). The inharmonic sound of tubular bell seems to suggest to participants' its inability to blend with other instruments, as all the other instruments have relatively much more harmonic sounds with clearly defined pitches. In reality, the sound of tubular bell was actually able to blend with others to a better degree, possibly due to masking and complex interactions between frequencies of the paired sounds. On the one hand, this again points to the fact that blends in imagination are probably not able to preserve and account for the actual acoustic interactions happening within real blends. On the other hand, it provides some evidence that the evaluation of blends in imagination might be biased on simply evaluating the similarity between constituent instruments, as the immediacy of tubular bell's uniqueness possibly explains its low blend ratings when paired with other instruments in imagination. These points will be re-discussed in the following sections.

The tendencies of single instruments to form pairs with significant conditional blend differences, as summarized in Fig. 3-9, showed some degree of correlation with instruments' attack slopes, RMS energy, and the second tristimulus value. Especially, the factor of attack profile seems consistent with the observation mentioned above that it might be easier to conceive and evaluate blends containing percussive and sustained instruments than those made of instruments with similar attack qualities. The other two factors are a bit difficult to explain, and are thus in need of further investigation.

Analyses with family-wise blends essentially suggest very similar findings. Combinations between non-percussive families were all perceived to blend significantly higher in reality than in imagination, except for the self-combination of string family which were perceived to be essentially the same in the two conditions. The differences between the two blending conditions for combinations involving percussions are all non-significant. These findings are in line with the observations obtained with instrument-wise blends that instruments with contrasting attack behaviors are more immediate to blend evaluation in imagination than other pairs.

4.2 The effect of musical backgrounds

Plotting the distribution of blend ratings by musicians and non-musicians separately (Fig. 4-1) suggests that non-musicians tended to give more polarized blend ratings than musicians. As shown in Fig. 4-1, non-musicians tended to use the extremities of the rating scale more frequently than musicians did. Verbal feedback from some non-musician participants suggests that the idea of blend exists on a continuum was understandable for them but somehow still a bit difficult to operate on themselves. As musicians supposedly engage with instrumental practices and using blends much more frequently, it seems reasonable that they are more confident in and capable of describing blends with small differences.

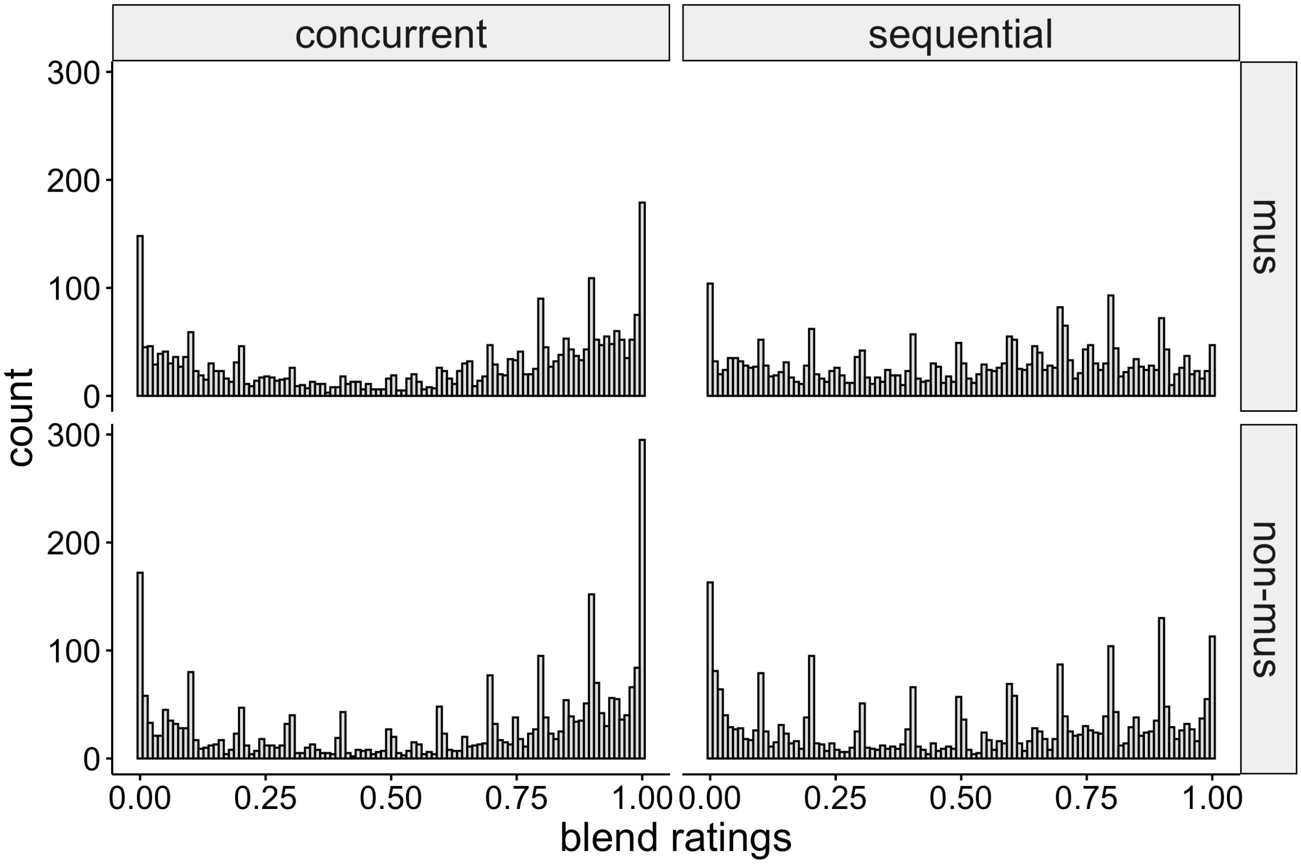


Figure 4-1. Distributions of blend ratings by musicians (upper panels) and non-musicians (lower panels) separated by the two blending conditions.

The significant three-way interaction between musical background, instrument pair, and blending condition suggests that musical background may play a role in how blends are imagined vs. heard for different instrument pairs. Fig. 3-10 suggests that for instrument pairs that do significantly distinguish between the two groups of participants in terms of conditional blend differences, musicians' imagination of blend seem to be overall more similar to their perception of heard blends compared to non-musicians. For non-musicians, the potential for these pairs to blend are more prominently "underestimated" than with musicians. The dominance of tubular bell in the pairs shown in Fig. 3-10, especially with the pair RGL_TP (also the case for EH_RGL and OB_RGL, but less prominent) showing different directions of conditional differences, suggests that its blends were probably perceived or "understood" differently by musicians and non-musicians, possibly due to its inharmonic nature. The other possibility is related to the reported ambiguity of blends involving one percussive and one sustained instrument. Similar to the results found in Tardieu & McAdams (2012), participants sometimes reported very different ideas about how such mixed pairs blend in perception: while some stated that they didn't think such pairs blended at all (whether in reality or imagination) because of the immediate contrasts between the two instruments' attacks, some said they saw such pairs as "chimeric" blends with the attack coming from the impulsive instrument and the tail (sustained part) from the sustained instrument, creating a plausible "new" instrument as a single source in perception. This ambiguity was seen in both musicians and non-musicians and probably caused the sometimes-contrasting perceptual results around these pairs.

4.3 Acoustic correlates of blend

Overall, acoustic modeling of blends in the concurrent and sequential conditions suggests a similar set of features that may explain the blend ratings. The model for concurrent blend ratings shares a few common acoustic predictors with the one obtained in Sandell (1991) for unison blends: in Sandell's work, the absolute difference of perceptual attack time and the third value (energy of the fifth harmonic and above) of tristimulus of the difference spectrum between the paired instruments were included in the final model. Perceptual attack time difference describes the contrasting behaviors of attacks of the paired instruments, which functions similarly to the feature AttSlope_diff in the current model. The tristimulus feature in Sandell's model is conceptually the same as Tri3_Med_diff in the current model, with some minor difference in calculation[15]. For the concurrent blend model, a few other tristimulus related features also turned out to be significant. The model coefficients suggest that as the difference between instruments' tristimulus values become larger (for the higher harmonic region above the fundamental), their blend becomes worse. The positive coefficient of Tri2_Med_sum seems to suggest that overall higher energy in the middle harmonic regions (spanning from the second to the fourth harmonics) also promotes blend. In line with Sandell's model, the negative coefficient of AttSlope_diff means that large difference between attack behaviors leads to worse blend. The prominence of tristimulus related features found in models of both blending conditions suggests an important role of features related to contrasts of spectral envelopes between the blending sounds. Further investigation of more refined features like formant shapes and locations might give better explanatory power.

Interestingly, the coefficient of TriCor is negative, which suggests that when other predictors are held constant, higher positive correlation between tristimulus of the paired instruments (i.e., instruments with similar harmonic energy distribution of the three harmonic regions) leads to worse blend. This seems to go against some of the previous evidence that concurrent sounds with insufficiently overlapping spectra tend be heard independently (Reuter, 1997). The same negative contribution from TriCor is also observed for sequential blend ratings. A possible explanation (at least in the case of concurrent blends) for this counterintuitive behavior might be that all the concurrent pairs were synthesized by simply superimposing two instrument samples that were recorded independently, which were not intended for an ideal blending scenario. A problematic technical detail found during the preparation stage of the experiment was that playing these individual sounds together created some unnatural artefacts for certain pairs due to phase cancellation especially for the high partial regions[16]. This potentially complicated the masking behaviors between the spectra of sounds. An ideal scenario would be to record blends as-is in a natural manner à la Kendall & Carterette (1991) where mixing and position of microphones can

[15] The tristimulus feature used in Sandell's work was calculated first by taking a difference spectrum between average spectra of the two paired instruments. Tristimulus values were then calculated on the difference spectrum.

[16] This was also the underlying reason for manually adjusting the time shift values for synchronizing instruments when preparing the concurrent stimuli (see section 2.1.2), as a way to preserve the naturalness of blends while also ensuring perceptual synchrony between the instruments.

be adjusted to best reflect how instrumental blends are usually produced and heard, which would probably lead to different results than the current findings.

On the other hand, as tristimulus doesn't fully reflect the distribution of formants, the contribution of negatively correlated tristimulus values might not necessarily negate the presence of overlapping formants, but rather simply suggests an additional favoring of dissimilar harmonic energy distribution as a blend contributor when other factors are held constant. The persistence of this contributor in the sequential condition also seems to suggest a universal preference for such regional harmonic energy contrasts regardless of whether the blends are heard or imagined. It is not difficult to see that TriCor can be uniquely determined from pairwise composite and difference values for the three tristimulus descriptors (i.e., Tri1_Med_sum, Tri1_Med_diff, Tri2_Med_sum, …, if neglecting the small error resulting from using median as the summary statistic), though non-linearly. The significance of TriCor and a few specific composite and difference tristimulus features found in regression suggests that the relationship between blend and tristimulus is quite convoluted. It might help to design a more focused experiment in which a set of sounds are synthesized, sharing basic acoustic features such as formant locations, attack profiles, etc. At the same time, the relative distribution of harmonic energy in different frequency regions is allowed to vary for different sounds, providing various tristimulus profiles. After listening to pairs of these sounds, participants would then be asked to rate their blends, following a similar concurrent vs. sequential conditional setup. Such an experiment, where the difference between sounds is largely limited to only tristimulus, could hopefully give a clearer picture of how tristimulus affects blend specifically.

The commonality of features retained by models of concurrent and sequential blends seems to support the qualitative observation in section 4.1 that heard and imagined blends are overall parallel. Surprisingly, a factor that is missing from both models is spectral centroid, which has usually made a significant contribution to modeling perceived blend. A possible explanation of this could be found in Lembke et al. (2019), where they included the blends of impulsive and sustained sounds in their stimuli. They observed that the temporal features were more influential in modeling blends while spectral features became secondary or even tertiary. They suggested that "In perceptual tasks comparable to those employed in these experiments, participants may focus their attention on the dominant distinctions across stimuli at the cost of perceptual resolution for the less pronounced differences", which seems applicable to the current study not only because of the inclusion of impulsive instruments that might draw the attention of participants disproportionately, but also because of the contrasts between the two experimental conditions that could have biased the perception of participants as mentioned before, which similarly might lead to the reduced "perceptual resolution for the less pronounced differences".

When comparing the two "blend spaces" obtained with MDS, a few interesting observations can be made. Very prominently, both spaces showed the same segregation between percussive and non-percussive instruments, which again supports the universal role of attack contrasts in shaping blends. When looking at the projections of acoustic descriptors, certain significant correlates (e.g., Tri1_Med and SpecCrest_Med) remained almost constant in terms of their directions, suggesting some invariant aspects when judging heard and imagined blends. The absence of significant projections of Tri3_Med and SpecCent_Med in the space of heard blends, which, however, appear to be significant in the space of imagined blends, deserves a closer look. Fig. 4-2 shows the two spaces together, with a few "pairs with larger conditional blend differences" (see section 3.2.1.2) being emphasized by connecting the paired instruments with a colored line. First of all, note that despite being non-significant in the concurrent space, the projections of Tri3_Med and SpecCent_Med retain almost the same directions there as in the sequential space. Secondly, note the changes of orientation of the highlighted pairs between the two spaces. While these pairs are relatively close together on the projections of Tri3_Med and SpecCent_Med in the concurrent space (the connected lines are close to perpendicular to these projections), they are farther apart (i.e., less blended) and much more stretched out along the projections of Tri3_Med and SpecCent_Med (the connected lines are comparatively closer to parallel to these projections).

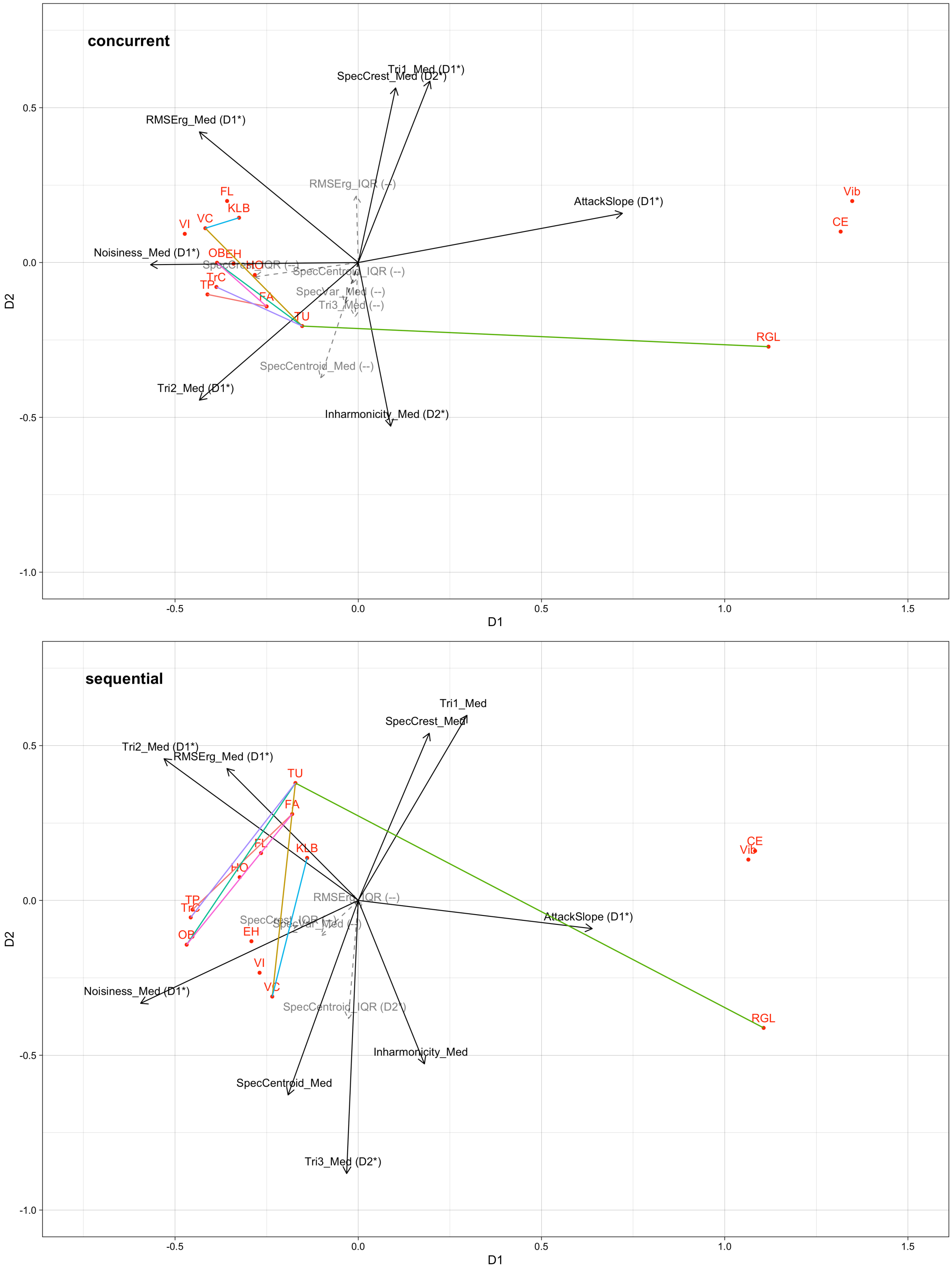


Figure 4-2. MDS spaces obtained with concurrent and sequential blend ratings. A few pairs that were tested to have large conditional blend differences are connected by colored lines (colors are shared by the two graphs).

These behaviors seem to suggest that (at least for the cases of pairs that have significantly large conditional blend differences), when participants imagine blends, a primary focus is on the differences between the high frequency contents of the two sounds. If the sounds are very different in terms of their brightness and/or relative energy in the high partial regions, then they are less likely to blend well in imagination because of their uniqueness on these aspects. On the other hand, these two features seem to be more irrelevant to evaluation of heard blends, possibly because of, as already discussed above, the complex masking and interaction between frequencies happening in real blends that were not faithfully accounted for in imagination but led to better perceived blends in real sounds, as well as the generally reduced perceptual resolution for these frequential features. The absence of these two projections in the concurrent "blend space" plausibly collapses the instruments into a smaller space (thus better blends in general), where some of the individual differences between instruments that are deemed to be prohibitive to blends in imagination become less influential for good blending to happen with real sounds.

In the end, one might also surmise that evaluating blends in imagination resorts more or less to judging the similarity between the involved instruments, instead of a complete mental recreation of two instruments sounding together. Based on these comparisons and some participants' feedback, it seems that there is an underlying analytical or "comparative" inclination when imagining and evaluating blends, where certain differences and contrasts between two sounds are readily understood (or intuitively interpreted) as a signifier of non-blends, without necessarily undergoing the supposed "superimpose-and-analyze" mental process. A comparative study using the same stimuli with participants rating the perceived similarity between instruments might provide more evidence for this hypothesis.

As a concluding mark, it is worth mentioning that the interpretation of acoustic correlates based on MDS spaces might be compared with the previous results obtained from direct regression modeling on blend data as complementary explanations. Their divergence (e.g., spectral centroid appears to be meaningful within "blend spaces" but not in the regression models) is expected, as MDS already re-models and abstracts the blend data itself. Thus, the two approaches essentially provide two alternative angles in tandem for interpreting and explaining the observed blends.

Chapter 5
Conclusion

The current study assessed how instrumental blends are perceived in imagination in comparison with the perception of blend in physically presented sound dyads. The results show that macroscopically imagined blends are largely consistent with heard blends, suggesting that mental images of blends can be constructed from isolated sources, which still match the perception of real blends to a certain degree. On a microscopic level, how well the imagined images of blends correspond to the heard blends is contingent on the complex interaction between the specific instruments involved and the participants' musical backgrounds. For pairs involving impulsive instruments, their blends can be readily conceived and evaluated in imagination relying on the contrasting attack profiles of the instruments, therefore most closely matching the perception of heard blends. Without this cue, blends can be harder to be properly conceived in imagination, and the degree of blend is usually underestimated.

Acoustic modeling shows similar sets of acoustic features meaningful for explaining the two types of blends. This invariance is also reflected in multidimensional scaling of the two sets of blend data, where certain acoustic correlates show similar projections in the two "blend spaces". Attack slope and tristimulus appear to have consistent explanatory power for both heard blend and imagined blend ratings, where greater difference between the respective values of the two instruments leads to worse blend. Interestingly, features related to the amount of energy present in the higher frequency regions of sounds appear to be meaningful when explaining instrument pairs that received significantly different ratings between the concurrent and sequential conditions. Instruments with rich high partials and prominent energy in the high frequency regions (such as trumpet and tubular bell) possibly signal an innate disinclination to blend with others due to their uniqueness, which functions as an important cue for evaluating imagined blends. In reality, due to the complex masking and interactions between the spectra, the degree of blends for these instruments is actually improved to different extents, as reflected by their higher concurrent blend ratings. It seems plausible that it is difficult for the mental faculty to fully account for the complex sensory interactions occurring between sounds that may affect the actual blends. A "comparative" strategy, by which sounds are analyzed according to their differences and uniqueness, thus seems to be what is underlying the evaluation of imagined blends besides an authentic mental image of the blends as superimposed timbral composites.

The factor of musical backgrounds is complex and intertwined with specific instruments in question. Non-musicians tended to give more extreme ratings to a given pair than musicians, probably because musicians are more sensitive to various degrees of timbral changes and are able to evaluate blends on a finer scale. The greatest divergence seems to be associated with the perception of pairs involving one sustained and impulsive instrument, as the idea of how such pairs blend is rather blurry and different across individuals.

As mentioned in the previous chapters, there are several limitations of the current study due to time and resource constraints. The stimuli used in the current studies were individual samples that didn't take into account the "musical" scenario of blending with other instruments. Instrumental blends in a real-world scenario entail coordination between musicians that are specific to the instruments, and ideally the reception of the blends would also be optimized in terms of the spacing between instruments and listeners, as well as room acoustics. Technical problems as such unwanted cancellation between sounds would be greatly alleviated with blends being recorded by pairs of performers. This would also spare the effort of manually synchronizing pairs of instruments with greater precision and naturalness.

Regarding the analytical methods, as mentioned in section 3.2.1, including random slopes for the within-subjects factors should improve the validity of the analyses with linear mixed models. Acoustic features for modeling blends were chosen primarily based on prior knowledge and generic representativeness, which are far from comprehensive. Approaches allowing for selecting meaningful regressors out of a large number of potential predictors, such as partial least squares or lasso regression, might improve the interpretability of the current results with other useful acoustic correlates. Local spectral descriptors pertaining to the formant structures of instruments (Lembke & McAdams, 2015) also seem promising as additional features to explore for explaining the blend ratings. Finally, as mentioned in section 4.1, the ordering of the two conditions in the current experiment seems to have a potentially great effect on how participants evaluate blends. While presenting the sequential condition first can help bypass the priming effect with presenting concurrent blends to participants first, it probably brought in other biases as mentioned in the previous discussion. Randomizing the order of presentation as a comparative follow-up study might be able to show how great the biases are.

As imagination of timbral combinations are most often encountered by composers or conductors when dealing with real-world compositions, it seems logical that a more musically meaningful examination of the properties of timbral imagination should allow for a larger musical context. Composers rarely focus on the blends of two static single notes. Phrasal, melodic, or even textural organizations thus may serve as meaningful musical contexts for future explorations on timbral imagination, allowing a more organic and ecologically valid investigation of other possible musical factors that affect blends in imagination.

On the other hand, the yet-unanswered question of what is actually happening when people actively imagine blends calls for more detailed neurological approaches to elucidate its underlying mechanism. Some of the participants recognized after the experiment that they didn’t always or were not always sure that they followed the instruction of imagining the two sounds playing simultaneously (as it was not always easy), but rather switched to other methods momentarily, such as holding the first note in memory and superimposing this image onto the second note when it was being played where evaluations were made on this “half-imagined” blend. Other methods reported include simply judging the similarity between instruments as a “shortcut” to rate how they would blend in imagination. Methods such as functional MRI (fMRI), as adopted in Halpern et al. (2004), might shed light on how comparable imagining blends is with hearing blends or simply comparing two sounds on a neurological level.

Appendix 1: Loudness match adjustment

Table A. Loudness adjustment applied to instrumental samples based on the results of loudness matching pre-experiment.

Filenames	Loudness adjustment (dB)
CE_ES_f_D#4.wav	2.73
EH_pA_sus_f_D#4.wav	-4.14
FA_nA_sus_f_D#4.wav	-0.55
FL1_oV_pA_sus_f_D#4.wav	-1.55
HO_oV_nA_sus_f_D#4.wav	0.4
KLB_nA_1-105_f_D#4.wav	-1.81
OB_pA_sus_f_D#4.wav	0.14
RGL_led_ff_D#4.wav	0.93
TP_oV_nA_sus_f_D#4.wav	-1.54
TU_oV_nA_sus_f_D#4.wav	-1.05
TrC_oV_nA_sus_f_D#4.wav	-1.11
VC_sus_mf_D#4.wav	-0.43
VI_sus_mf_D#4.wav	-3.56
Vib_ES_Me_sp-0_mf1_D#4.wav	3.27

Appendix 2: Synchrony adjustment

Table B. Time offsets between instrumental samples applied to ensure pairwise perceptual synchrony. Values are times in milliseconds that row's stimuli should be delayed to be in synchrony with column's stimuli.

	CE	EH	FA	FL	HO	KLB	OB	RGL	TP	TU	TrC	VC	VI	Vib
CE	-	-	-	-	-	-	-	-	-	-	-	-	-	-
EH	-2.9	-	-	-	-	-	-	-	-	-	-	-	-	-
FA	-30.5	-17.4	-	-	-	-	-	-	-	-	-	-	-	-
FL	-108.8	-117.6	-107.4	-	-	-	-	-	-	-	-	-	-	-
HO	-11.6	2.9	23.2	98.7	-	-	-	-	-	-	-	-	-	-
KLB	-40.6	-36.3	-29	58	-42.1	-	-	-	-	-	-	-	-	-
OB	-5.8	-8.7	17.4	95.8	-1.5	26.1	-	-	-	-	-	-	-	-
RGL	-2.9	17.4	34.8	119	2.9	74	-1.5	-	-	-	-	-	-	-
TP	-37.7	-20.3	10.2	97.2	-24.7	33.4	-16	-36.3	-	-	-	-	-	-
TU	-4.4	-1.5	20.3	121.9	-7.3	47.9	-1.5	-5.8	11.6	-	-	-	-	-
TrC	-23.2	-2.9	17.4	127.7	-7.3	53.7	-21.8	-4.4	2.9	-2.9	-	-	-	-
VC	-36.3	-13.1	11.6	98.7	-5.8	23.2	-11.6	-16	4.4	-7.3	-1.5	-	-	-
VI	-65.3	-4.4	-5.8	90	-24.7	13.1	-40.6	-40.6	-13.1	-14.5	-7.3	-10.2	-	-
Vib	1.5	18.9	40.6	105.9	7.3	59.5	18.9	2.9	42.1	10.2	13.1	53.7	37.7	-

(Note: One may observe that time offsets associated with the flute sound are almost always much larger than with other instruments. This can be attributed to the characteristically slow and airy attack of the flute, which calls for bigger temporal adjustments for it to be in sync with other instruments.)

References

Adler, S. (2002). *The Study of Orchestration* (3rd ed.). W. W. Norton and Company.

Atkinson, R. C., & Shiffrin, R. M. (1968). Human Memory: A Proposed System and its Control Processes. In K. W. Spence & J. T. Spence (Eds.), *Psychology of Learning and Motivation* (Vol. 2, pp. 89–195). Academic Press. https://doi.org/10.1016/S0079-7421(08)60422-3

Baddeley, A. (2012). Working memory: Theories, models, and controversies. *Annual Review of Psychology*, *63*, 1–29. https://doi.org/10.1146/annurev-psych-120710-100422

Berlioz, H., & Strauss, R. (1948). *Treatise on instrumentation* (T. Front, Trans.). Edwin F. Kalmus.

Blatter, A. (1997). *Instrumentation and Orchestration* (2nd ed.). Schirmer Books.

Borg, I., Groenen, P. J. F., & Mair, P. (2018). *Applied Multidimensional Scaling and Unfolding* (2nd ed.). Springer International Publishing. https://doi.org/10.1007/978-3-319-73471-2

Bregman, A. S. (1990). *Auditory scene analysis: The perceptual organization of sound*. MIT Press.

Camos, V., Lagner, P., & Barrouillet, P. (2009). Two maintenance mechanisms of verbal information in working memory. *Journal of Memory and Language*, *61*(3), 457–469. https://doi.org/10.1016/j.jml.2009.06.002

Cowan, N. (1984). On short and long auditory stores. *Psychological Bulletin*, *96*(2), 341–370. https://doi.org/10.1037/0033-2909.96.2.341

Crowder, R. G. (1989). Imagery for musical timbre. *Journal of Experimental Psychology: Human Perception and Performance*, *15*(3), 472–478. https://doi.org/10.1037/0096-1523.15.3.472

Crowder, R. G. (1993). Auditory memory. In S. McAdams & E. Bigand (Eds.), *Thinking in Sound: The Cognitive Psychology of Human Audition*. Oxford University Press. https://doi.org/10.1093/acprof:oso/9780198522577.003.0005 de Leeuw, J., & Mair, P. (2009). Multidimensional Scaling Using Majorization: SMACOF in R. *Journal of Statistical Software*, *31*(3), 1–30. https://doi.org/10.18637/jss.v031.i03

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics using R*. Sage.

Goodchild, M., & McAdams, S. (2018). Perceptual processes in orchestration. In E. Dolan & A. Rehding (Eds.), *The Oxford handbook of timbre*. Oxford University Press. https://doi.org/10.1093/oxfordhb/9780190637224.013.10

Goodwin, A. W. (1980). An acoustical study of individual voices in choral blend. *Journal of Research in Music Education*, *28*(2), 119–128. https://doi.org/10.1177/002242948002800205

Gordon, J. W. (1987). The perceptual attack time of musical tones. *Journal of the Acoustical Society of America*, *82*(1), 88–105. https://doi.org/10.1121/1.395441

Gravetter, F. J., & Wallnau, L. B. (2008). *Essentials of statistics for the behavioral sciences* (6th ed). Thomson/Wadsworth. http://catdir.loc.gov/catdir/toc/fy0801/2007921807.html

Grey, J. M. (1975). *An Exploration of Musical Timbre* [Thesis, Stanford University]. https://ccrma.stanford.edu/files/papers/stanm2.pdf

Halikia, M. H. (1985). *The perceptual segregation of simultaneous sounds* [Thesis, McGill University]. https://escholarship.mcgill.ca/concern/theses/1g05fc384

Halpern, A. R., Zatorre, R. J., Bouffard, M., & Johnson, J. A. (2004). Behavioral and neural correlates of perceived and imagined musical timbre. *Neuropsychologia*, *42*, 1281–1292. https://doi.org/10.1016/j.neuropsychologia.2003.12.017

Heisig, J. P., & Schaeffer, M. (2019). Why You Should *Always* Include a Random Slope for the Lower-Level Variable Involved in a Cross-Level Interaction. *European Sociological Review*, *35*(2), 258–279. https://doi.org/10.1093/esr/jcy053

ISO 389–8. (2004). *Acoustics – Reference zero for the calibration of audiometric equipment – Part 8: Reference equivalent threshold sound pressure levels for pure tones and circumaural earphones* (Tech. Rep.). Geneva, Switzerland: International Organization for Standardization.

Kazazis, S., Depalle, P., & McAdams, S. (2021). *The Timbre Toolbox version R2021a, User's manual*. https://github.com/MPCL-McGill/TimbreToolbox-R2021a.

Kendall, R. A., & Carterette, E. C. (1991). Perceptual scaling of simultaneous wind instrument timbres. *Music Perception*, *8*(4), 369–404. https://doi.org/10.2307/40285519

Kendall, R. A., & Carterette, E. C. (1993a). Identification and blend of timbres as a basis for orchestration. *Contemporary Music Review*, *9*(1–2), 51–67. https://doi.org/10.1080/07494469300640341

Kendall, R. A., & Carterette, E. C. (1993b). Verbal Attributes of Simultaneous Wind Instrument Timbres: I. von Bismarck's Adjectives. *Music Perception: An Interdisciplinary Journal*, *10*(4), 445–467. https://doi.org/10.2307/40285583

Kendall, R. A., & Carterette, E. C. (1993c). Verbal Attributes of Simultaneous Wind Instrument Timbres: II. Adjectives Induced from Piston's "Orchestration." *Music Perception: An Interdisciplinary Journal*, *10*(4), 469–501. https://doi.org/10.2307/40285584

Knief, U., & Forstmeier, W. (2021). Violating the normality assumption may be the lesser of two evils. *Behavior Research Methods*, *53*(6), 2576–2590. https://doi.org/10.3758/s13428-021-01587-5

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in Linear Mixed Effects Models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13

Lembke, S.-A., Levine, S., & McAdams, S. (2017). Blending between bassoon and horn players: An analysis of timbral adjustments during musical performance. *Music Perception*, *35*(2), 144–164. https://doi.org/10.1525/MP.2017.35.2.144

Lembke, S.-A., & McAdams, S. (2015). The role of spectral-envelope characteristics in perceptual blending of wind-instrument sounds. *Acta Acustica United with Acustica*, *101*(5), 1039–1051. https://doi.org/10.3813/AAA.918898

Lembke, S.-A., Parker, K., Narmour, E., & McAdams, S. (2019). Acoustical correlates of perceptual blend in timbre dyads and triads. *Musicae Scientiae*, *23*(2), 250–274. https://doi.org/10.1177/1029864917731806

Mair, P., Borg, I., & Rusch, T. (2016). Goodness-of-Fit Assessment in Multidimensional Scaling and Unfolding. *Multivariate Behavioral Research*, *51*(6), 772–789.

Martin, F. N., & Champlin, C. A. (2000). Reconsidering the limits of normal hearing. *Journal of the American Academy of Audiology*, *11*(2), 64–66.

McAdams, S. (1984a). *Spectral Fusion, Spectral Parsing, and the Formation of Auditory Images* [Thesis]. Stanford University.

McAdams, S. (1984b). The auditory image: A metaphor for musical and psychological research on auditory organization. In W. R. Crozier & A. J. Chapman (Eds.), *Cognitive Processes in the Perception of Art* (pp. 289–323). North-Holland. https://doi.org/10.1016/S0166-4115(08)62356-0

McAdams, S. (2013). Musical timbre perception. In D. Deutsch (Ed.), *The Psychology of Music (3rd Edition)* (pp. 35–67). New York: Academic Press. https://doi.org/10.1016/B978-0-12-381460-9.00002-X

McAdams, S. (2019a). The perceptual representation of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 23–57). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_2

McAdams, S. (2019b). Timbre as a structuring force in music. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 211–243). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_8

McAdams, S., Gianferrara, P., Soden, K., & Goodchild, M. (2016). *Factors influencing instrument blend in orchestral excerpts*. 14th Biennial Meeting of the International Conference on Music Perception and Cognition, San Francisco.

McAdams, S., Goodchild, M. & Soden, K. (in press). A taxonomy of orchestral grouping effects derived from principles of auditory perception. *Music Theory Online*.

Nees, M. A., Corrini, E., Leong, P., & Harris, J. (2017). Maintenance of memory for melodies: Articulation or attentional refreshing? *Psychonomic Bulletin & Review*, *24*(6), 1964–1970. https://doi.org/10.3758/s13423-017-1269-9

Peeters, G., Giordano, B. L., Susini, P., Misdariis, N., & McAdams, S. (2011). The Timbre Toolbox: Extracting audio descriptors from musical signals. *The Journal of the Acoustical Society of America*, *130*(5), 2902–2916. https://doi.org/10.1121/1.3642604

Piston, W. (1955). *Orchestration*. W. W. Norton.

Rasch, R. (1978). The perception of simultaneous notes as in polyphonic music. *Acustica*, *40*, 21–33.

Reuter, C. (1997). Karl Erich Schumann’s principles of timbre as a helpful tool in stream segregation research. In M. Leman (Ed.), *Music, Gestalt, and computing: Studies in cognitive and systematic musicology* (pp. 362–374). Springer Verlag. https://doi.org/10.1007/BFb0034126

Rimsky-Korsakov, N. (1964). *Principles of orchestration: With musical examples drawn from his own works* (M. Steinberg, Ed.; E. Agate, Trans.). Dover.

Sandell, G. J. (1989). Perception of concurrent timbres and implications for orchestration. *Proceedings of the 1989 International Computer Music Conference* (pp. 268-272). San Francisco, CA: International Computer Music Association.

Sandell, G. J. (1991). *Concurrent timbres in orchestration: A perceptual study of factors determining “blend”* [Thesis]. Northwestern University.

Sandell, G. J. (1995). Roles for spectral centroid and other factors in determining “blended” instrument pairing in orchestration. *Music Perception*, *13*(2), 209–246. https://doi.org/10.2307/40285694

Satterthwaite, F. E. (1946). An Approximate Distribution of Estimates of Variance Components. *Biometrics Bulletin*, *2*(6), 110–114. https://doi.org/10.2307/3002019

Scheffers, M. T. M. (1983). Sifting Vowels: Auditory Pitch Analysis and Sound Integration [Thesis]. University of Groningen.

Schulze, K., & Tillmann, B. (2013). Working memory for pitch, timbre, and words. *Memory*, *21*(3), 377–395. https://doi.org/10.1080/09658211.2012.731070

Siedenburg, K., & McAdams, S. (2017a). Four distinctions for the auditory “wastebasket” of timbre. *Frontiers in Psychology*, *8*, 1747. https://doi.org/10.3389/fpsyg.2017.01747

Siedenburg, K., & McAdams, S. (2017b). The role of long-term familiarity and attentional maintenance in short-term memory for timbre. *Memory*, *25*(4), 550–564. https://doi.org/10.1080/09658211.2016.1197945

Siedenburg, K., & Müllensiefen, D. (2019). Memory for timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, Perception, and Cognition* (pp. 87–118). Springer International Publishing. https://doi.org/10.1007/978-3-030-14832-4_4

Smith, B. K. (1995). PsiExp: An environment for psychoacoustic experimentation using the IRCAM musical workstation. In *Society for Music Perception and Cognition Conference'95*. Berkeley, CA: University of California, Berkeley.

Soemer, A., & Saito, S. (2015). Maintenance of auditory-nonverbal information in working memory. *Psychonomic Bulletin & Review*, *22*(6), 1777–1783. https://doi.org/10.3758/s13423-015-0854-z

Stiller, A. (1985). *Handbook of instrumentation*. University of California Press.

Tardieu, D., & McAdams, S. (2012). Perception of dyads of impulsive and sustained instrument sounds. *Music Perception*, *30*(2), 117–128. https://doi.org/10.1525/Mp.2012.30.2.117

Zatorre, R. J., & Halpern, A. R. (2005). Mental Concerts: Musical Imagery and Auditory Cortex. *Neuron*, *47*(1), 9–12. https://doi.org/10.1016/j.neuron.2005.06.013