Jahrbuch Musikpsychologie
Yearbook of Music Psychology

PsychOpen GOLD

Short Papers

Detection of Clash of Keys in a Non-Dichotomous Task: Effects of Musical Genres, Instruments and Participants' Expertise

Erkennen von Tonartenkonflikten mittels nicht-dichotomer Aufgabe: Effekte von Musikgenres, Instrumenten und Teilnehmer*innen-Expertise

Anna Wolf[1], Bastian Wüst[1]

[1] *Department of Music Education and Church Music, University of Music FRANZ LISZT Weimar, Weimar, Germany.*

Jahrbuch Musikpsychologie, 2026, Vol. 34, Article e225, https://doi.org/10.5964/jbdgm.225

Received: 2025-04-04 • **Accepted:** 2026-01-19 • **Published (VoR):** 2026-02-25

Reviewed by: Kai Siedenburg; Elke Lange.

Corresponding Author: Anna Wolf, Department of Music Education and Church Music, University of Music FRANZ LISZT Weimar, Platz der Demokratie 2/3, 99423 Weimar, Germany. E-mail: anna.wolf@hfm-weimar.de

Supplementary Materials: Code, Data, Materials [see Index of Supplementary Materials]

Abstract

When listening to music, people are able to separate the various parts or streams that the respective piece is made of, depending on their musical experience and aural training. Specific musical expertise affects these analytical listening skills, but different studies have shown that differences between experts and amateurs are smaller than often anticipated and depend heavily on the adequacy of the task for amateurs. The present replication study of Kopiez and Platz (2009) has investigated in a convenience sample (N = 97) whether a presumably obvious clash of keys between solo and accompaniment can be detected. In an incomplete study design, participants listened to two pieces of music (jazz and classical) with solos each played by one of two instruments (trumpet and saxophone) in either a clashing or fitting condition. Participants' musical training and perceptual abilities showed a medium correlation with the harmoniousness ratings difference between the clashing and fitting stimuli. Overall, the clashing version of the classical piece was rated as less harmonious than the jazz piece. These results are in line with similar research and raise questions concerning the appropriate research method to investigate the perception of participants with various degrees of expertise.

Keywords

analytical listening, bitonality, harmoniousness, perceptual abilities, replication study

Zusammenfassung

Beim Hören von Musik sind Menschen, je nach musikalischer Erfahrung und Gehörbildung, unterschiedlich gut in der Lage die verschiedenen Stimmen oder Ströme im jeweiligen Stück zu trennen. Spezifische musikalische Expertise wirkt sich auf diese analytischen Hörfähigkeiten aus, wobei verschiedene Studien zeigen, dass die Unterschiede zwischen Experten und Amateuren oft geringer sind als angenommen und wiederum stark von der Eignung der Aufgabe für die Amateure abhängen. In der vorliegenden Replikationsstudie von Kopiez und Platz (2009) wurde untersucht, ob ein vermeintlich offensichtlicher Tonartenclash zwischen Solo und Begleitung in zwei Musikstücken von Teilnehmer*innen einer Gelegenheitsstichprobe (N = 97) erkannt wurde. Im Rahmen eines unvollständigen Designs haben die Teilnehmer*innen zwei Musikstücke gehört (Jazz und Klassik), wobei die Soli jeweils von einem von zwei Instrumenten (Trompete und einem Saxofon) in entweder passenden oder „kollidierenden" Tonarten gespielt wurden. Die Expertise und die musikalischen Wahrnehmungsfähigkeiten der Teilnehmer korrelierten mittelstark mit der Differenz der Bewertungen für kollidierende und passende Stimuli. Insgesamt wurde die kollidierende Version des klassischen Stücks als weniger stimmig eingestuft als das Jazzstück. Diese Ergebnisse entsprechen den Erkenntnissen ähnlicher Untersuchungen und werfen Fragen zur passenden Forschungsmethodik für die Wahrnehmung bei Personen mit unterschiedlicher musikalischer Expertise auf.

Schlüsselwörter

Analytisches Hören, Bitonalität, Stimmigkeit, Musikalische Wahrnehmungsfähigkeiten, Replikationsstudie

"Write some ditty in one key, write the accompaniment in another, and *voilà*— something that sounds as 'bad' as the most studiously atonal utterance of a real, hard-working composer" (Harrison, 1997, pp. 393–394). Such reported ridicule of bitonality is omnipresent in the music theoretical discourse. As pointedly condensed in the introductory quote, it is to a large degree rooted in the fact that bitonality kept all features of tonal music and only changed the uniformity of one musical key in one musical piece to two—or more, in the case of polytonal music (Whittall, 2001).

Famous composers who employed bitonality as compositional techniques were, for example, Charles Ives, Benjamin Britten, Darius Milhaud, and Igor Stravinsky. Within their compositions they used various degrees of bitonality, such as a single melodic line in a different key, two keys for the two voices in a duet (e.g., in Peter Grimes by Britten), or as a continuous stylistic feature for a whole, albeit often rather short piece (e.g., the Saudades do Brazil by Milhaud).

Following the Gestalt laws to the concept of bitonality, there are several indications that might suggest a one-to-one mapping from the compositional idea using two keys to the listener's capacity to hear those two keys. The law of common fate—regarding the tonic and the harmonic and melodic development within one voice as a common fate—and the law of the "good Gestalt" (and maybe even the laws of proximity and similarity) should, certainly depending on the exact piece of music, predict the possibility

of listeners to perceive and appreciate the two voices in their different keys (Wertheimer, 1923, translation by Ellis, 1938). However, more recent findings have shown that, in comparison to visual perception, for which the Gestalt laws have primarily been developed, it is only rarely possible to listen to and follow two separate auditory streams at the same time (Oehler, 2014; Bregman, 1990, pp. 524–525). Therefore, we assume a rather intertwined percept as a predominant mode of perception, where the scene analysis skills are dependent, e.g., on one's musical experience and specific aural skills (e.g., Hake et al., 2024).

Aside from the different notions on bitonality as either a "distinctly mechanical" (Whittall, 2001) or skillful compositional technique, there are many findings focusing on the disparity between ideas embedded in musical concepts (composed music) and ideas experienced in musical percepts (heard music). Investigations into the latter have surfaced a variety of unexpected findings: Concerning the tonal center of a piece, listeners are not influenced in their rating whether this piece remains in the originally tonal center or modulates towards another key (Cook, 1987). Maybe even more surprising, our preference for classical music does not change when form parts of this piece are presented in the order intended by the composer of randomly rearranged into a nonsensical order (Karno & Konečni, 1992). And with regards to manipulations in a piece's tonality, Lalitte et al. (2009) found in their study on tonal (i.e., original) and atonal (i.e., artificially created) versions of Beethoven sonatas that participants followed criteria of musical style and presence/absence of tonality and therefore, respectively, rating the tonal excerpts on the one hand and the atonal excerpts on the other hand as very similar. However, when segmenting the piece and reporting their arousal, these implicit ratings followed the similar structure within the dyad of a tonal and the thereof created atonal piece. Music training was no explaining factor for participants' ratings here.

More specific to the perception of bi- and polytonality, Thompson and Mor (1992) adapted the probe tone method and captured the tonal perception in bitonal pieces: Responses matched the combined key profiles of the utilized keys, the prevalent keys in the compositions were simultaneously conveyed to participants and the authors were even able to rate which key stronger influenced participants' key perception. This study is connected to a previous study by Krumhansl and Schmuckler (1986), who, largely unsuccessfully, tried to manipulate the perception of the key profile by introducing them dichotically using the Petroushka passage. Their results also point to a perceptual fusion of both keys, which is so predominant in the perception that it was impossible to only pay attention to one key and ignore the other. Both studies were conducted with musically trained participants with a usual lower limit of at least five years of music training.

In addition to these studies on music theoretical concepts and psychological percepts, one can find a variety of findings with surprising results on human music perception. On a basic level, Fricke (2014) offers a concise and numbers-based review of the lis-

tener's tolerance for intonational deviations in music. Listeners tend to tolerate (or "zurechthören"[1]) deviations of ~20 cents in most situations, which can increase to up to 45 cents, depending on the circumstances (length of tone, melodic tension, or passing vs. target note). Larrouy-Maestri and Morsomme (2014) concur with these findings from a production perspective: Even professional singers deviate about 30 cents (and more, when singing with lyrical technique) in singing production tasks. A following study taking perception into account quantifies listener's tolerance at roughly 25 cents with a high retest reliability of r = .88 (Larrouy-Maestri, 2018). Differences between experts and amateurs were notably large in this study and are further corroborated by the results from the mistuning perception test. The test relies on stimuli with mistunings between the vocal part and its accompaniment of 10 and 100 cents in either direction (Larrouy-Maestri et al., 2019). The hereby measured mistuning perception skills correlated with self-reported musical training and perceptual abilities with about r = .4. These results are especially well founded due to the careful test development, which not only addressed the instrument's reliability, but also its concurrent and discriminatory construct validity. In contrast, this mistuning perception test correlated with the PROMS tuning test (Law & Zentner, 2012) only with r = .25. Larrouy-Maestri et al. explain this divergence with a dissimilar item design, where Law and Zentner compared a mistuned chord with a correctly tuned reference chord enabling participants to use their echoic memory and thereby testing a partly different ability.

The effect of enculturation and training is also addressed in a review by Omigie et al. (2014): Considering music exposure and deliberate music training, the authors state that listeners' tolerance and liking for dissonant music must be separated because a higher tolerance for dissonance, induced by mere exposure and/or training, does not directly lead to a higher liking of such music.

Differences in music perception and analytical listening skills between experts and amateurs are by far not as clear-cut as a first assumption might seem. The seminal paper by Bigand and Poulin-Charronnat (2006) especially focuses on declarative language, which is learnt by musicians during their musical training, engagement with music notation and even more deliberate in music theory and ear training classes. "Untrained listeners", as labeled by Bigand and Poulin-Charronat, rarely possess such knowledge and are destined to fail at such tasks; however, when the tasks use fair language, especially using same-different-paradigms, we cannot observe skill or perceptual differences between trained and untrained listeners. In their investigation of melodic symmetry in melodies, Mongoven and Carbon (2017) found accuracy levels of about 36–50% (compared to a chance level of 33%), depending on the stimuli, and registered no differences in

1) Various translations exist from „harmonizing in hearing" (Kopiez & Platz, 2009, p. 329), "shift[ing] toward the intended pitch of each tone" (Parncutt & Hair, 2018, p.491), or "straightened out hearing" (Mazzola, 2017, p. 1353).

symmetry detection for participants with varying degrees of musical training, albeit in a small sample size and with a very rough measure of musical training.

Investigating the specific perception—or rather detection—of bitonality in music, the results by Wolpert (2000) were rather clear: While musicians reliably detect bitonality in music without any instruction or hint (overall detection of 100%), non-musicians recognize it to a much lower degree (overall detection of 40%). In a first replication study, Kopiez and Platz (2009) also took listening expertise, focus of attention (hinting at the fit between melody and accompaniment—or not) and musical style into account. They replicated the overall large difference between experts and non-experts—but at a much lower overall level (experts: 58% vs. amateurs: 18%)—and found style-specific disparities with highest detection rates in classical music (50%) and lowest detection rates in jazz (23%). Participants who were directly instructed to listen to the fit between melody and accompaniment scored higher (62%) than those who were not (24%). Discussing these results, the authors emphasize their results and state, "that hearing is not simply 'inverse acoustics' or 'inverse music theory'" and more "a kind of 'harmonizing in hearing' (Zurechthören)" (p. 329).

Hamamoto et al. (2010) expanded their study to ecological stimuli from the *Saudades do Brasil* by Darius Milhaud (interestingly, also in recomposed 'monotonal' versions as stimuli) and employed, among other tasks, an indirect rating task using scales for likeability, correctness and pleasantness. Monotonal stimuli were higher rated in correctness and pleasantness, although not in likeability, with larger differences in rating for musicians than non-musicians. Furthermore, after completing a rather fast training task to learn how to detect bitonality within their study (as well as the exposure from the previous listening tasks), non-musicians improved markedly and performed similarly to musicians, "once they have acquired the appropriate vocabulary" (p. 442). Their free response tasks, on the other hand, showed group differences between musicians (74%) and non-musicians (20%) in line with previous results.

The following replication study was a conceptual replication and introduced some changes compared to Kopiez and Platz (2009): We collected data online, adapted the rating task by Hamamoto et al. (2010), and investigated an effect of musical instrument/timbre in addition to genre. Our research questions and hypotheses were the following:

1. Is the clash of keys perceivable in a nondirected listening task using adjective rating scales addressing the stimuli's harmoniousness? We hypothesize that this study will replicate previous findings and that the clash of keys will lead to a much lower rating than the monotonal version.
2. We hypothesize that we will replicate differences in rating between experts and amateurs found; however, those could be large effects as observed by Wolpert (2000), Kopiez and Platz (2009) and the free response task by Hamamoto et al. (2010) or possibly smaller as found in the rating data by Hamamoto et al. (2010).

3. We hypothesize that the participants' self-reported music perceptual abilities as measured by the respective subscale of the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014; Schaal et al., 2014) are associated with the participants' ratings: on the one hand simply due to connections between this subscale and the various other constructs measured by the Gold-MSI (such as musical training), on the other hand because a self-report of one's perceptual abilities should theoretically correlate with a task indirectly measuring music perception skills.

Method

Design

The study followed a 2 × 2 × 2 study design with missing values by design. Participants listened to *2 pieces*, a classical and a jazz piece. The solo parts were played by *2 instruments*, a trumpet and an alto saxophone. We produced *2 versions* where the solo and the accompaniment were in the same key ("fitting") or where the accompaniment was a whole tone (200 cents) too low ("clashing"). To avoid carryover effects, participants listened to four versions overall: They listened to both pieces played by both instruments. We counterbalanced the design with regards to the fitting/clashing version so that participants were only presented with either the fitting or the clashing version (and not both) to avoid speculations when hearing the same piece played by the same instrument for a second time. Information on the musical stimuli can be found in the Data Availability section.

Participants

A total of $N = 97$ participants took part in the experiment (57 female, 40 male, 0 non-binary). Their mean age was $M = 34.1$ years ($SD = 15.7$) and most of them were students (40%) or had an employment relationship (45%). Overall, two thirds of the students (an absolute 27% of the 40%) were studying music, 43% were amateur musicians and 6% were working in the music business; none of the above applied to 24%.

The overall score in the Perceptual Abilities subscale (nine items, theoretical range = [9, 63]) of the Gold-MSI (Müllensiefen et al., 2014) was $M = 50.1$ with a $Mdn = 52$ and an $SD = 8.18$. Compared to the German speaking sample (Schaal et al., 2014), our sample's mean was four scale points above the 50th percentile and ranged in the 70th percentile, the self-assessed perceptual abilities were clearly above average.

Stimuli and Questionnaire

The stimuli originated from two pieces: 1) the Haydn trumpet concerto (Hob. Vlle/1, second movement "Andante", bars 8–16) and 2) the jazz standard "Take Five" by the Dave

Brubeck quartet (bars 1–12). Both solo parts were played by a trumpet as well as an alto saxophone in the original key. The chosen instruments are commonly used in either musical style while clearly differing from each other with respect to their timbre. The accompaniment for both pieces was produced on a Yamaha Clavinova CLP-370 with the sound of a Grand Piano, which was easily pitched from the correct key to a key a whole tone (200 cents) lower (Classical piece: A flat major & G flat major; Jazz piece: E flat minor & D flat minor).

After listening to each musical piece, participants rated their impression using the items elated ("beschwingt"), harmonious ("harmonisch"), solemn ("getragen"), off-key ("schief"), soft ("sanft"), cheerful ("heiter"), coherent ("stimmig") and stressed ("gestresst") on a 5-point scale (1 = *does not apply at all* to 5 = *applies completely*). The three items harmonious ("harmonisch"), off-key ("schief", as an inverted item) and coherent ("stimmig") were the a priori target items for the further statistical analysis, which we examined in a exploratory factor analysis (see Results).

Procedure

This study was conducted online using the platform soscisurvey.de (Leiner, 2024). Participants were welcomed and referred to the anonymity and voluntariness of the study. Using an audio example, participants could adjust the volume to a comfortable and normal level when listening to music. They were then presented with the four stimuli and rated their impression after each presentation. We asked for their demographic information on the second last page as well as whether they knew either of the musical pieces. On the last page participants replied to the Perceptual Abilities subscale from the Gold-MSI and specified the kind of speaker they used (tablet/phone loudspeaker, computer speaker, headphones or stereo equipment). They indicated how well they were able to hear the stimuli with regards to the audio examples' quality and the circumstances during their participation regarding their attention and distractions (M = 4.2 and M = 4.4; SDs = 0.7; range = [1; 5]). Finally, participants were thanked for their participation.

Results

Exploratory Factor Analysis

Compared to Wolpert (2000) and Kopiez and Platz (2009) we approached the participants' perception using a rating task with eight adjective ratings. Five items (*elated, solemn, soft, cheerful* and *stressed*) served as presumed distractor items and were assumed to be irrelevant for further analyses, three items (*harmonious, off-key* as an inverted item and *coherent*) were our presumed target items. Using an exploratory factor analysis, we investigated the factor structure within the full item set.

The Kaiser-Meyer-Olkin criterion for this data set was *KMO* = .80 and therefore adequate or meritorious (Kaiser & Rice, 1974) for a factor analysis. Both a scree plot as well as the Kaiser-Guttman criterion suggest a solution with two factors.

We employed an exploratory factor analysis using a maximum-likelihood estimation. The axes were rotated with a promax rotation, which allowed for correlations between factors and tried to reach near-zero correlations for every item with all but one factor. This option favors simplicity, but allows for relations between factors, which is plausible content-wise with this specific set of items that should all function as descriptors for musical stimuli (Abell et al., 2009). The found model does not differ from the empirical data, $\chi^2(28) = 22.04$, $p = .17$, and explains 64% of the variance in the data.

All three target items (*harmonious, off-key* as an inverted item and *coherent*) load on the first factor with loadings between .89 and .91. The next-highest loading is .55 for the item *stressed*, however, due to its much lower loading, its different content and since it was not part of the presumed a priori target items, we decided not to include it in the mean score for further analyses. The internal consistency of the three target items was $\alpha = .94$ and would decrease to .90 or .91 if one of the items were dropped. The mean score of these items (range 1 to 5) constituted the dependent variable and described the perceived harmoniousness in the stimuli.

Detection of the Clash of Keys: Descriptive Results

Participants' ratings differed largely between the clashing and fitting stimuli. The distributions of participants' responses for the eight conditions are displayed in Table 1, more detailed significance tests and effect sizes are reported with the ANOVA.

Table 1

Descriptive Statistics for the Analysis of Variance

	Fitting		Clashing	
	M	***SD***	***M***	***SD***
Classical: Saxophone	4.09	0.880	2.05	0.991
Classical: Trumpet	4.23	0.730	1.92	1.003
Jazz: Saxophone	4.63	0.438	3.13	1.415
Jazz: Trumpet	4.26	0.736	3.08	1.264

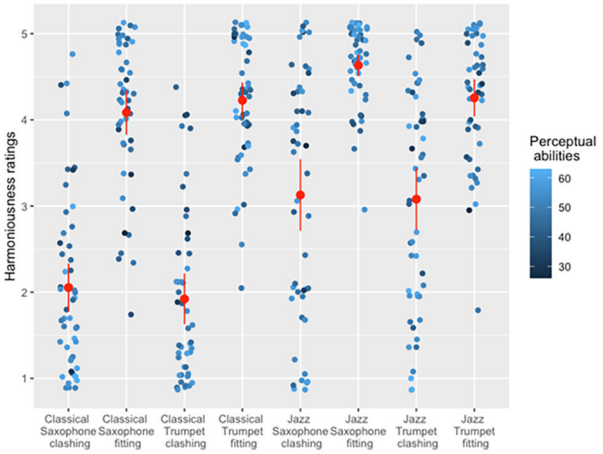
Analysis of Variance

We investigated the difference in harmoniousness perception between the eight conditions (2 pieces × 2 instruments × 2 versions) using an analysis of variance with two- and three-way interactions. Two relevant and significant stimuli-related effects appeared, a large difference between the fitting/clashing versions, $F(1, 95) = 71.3$, $p < .001$, $\eta_p^2 = .16$,

and a medium difference between the two pieces, $F(1, 95) = 24.8$, $p < .001$, $\eta_p^2 = .06$ (see Figure 1). There were no relevant differences with regards to the instrument or for any of the interactions between the independent variables.

Figure 1

Distribution of Harmoniousness Ratings Within the Eight Conditions



Correlational Analysis

Finally, we derived a new dependent variable to estimate whether a participant's perceptual ability and years of musical practice were related to their difference in rating between fitting and clashing stimuli. Due to an incomplete study design, every participant listened to two fitting and clashing stimuli each, for example, classical/trumpet/clashing and jazz/saxophone/fitting as well as classical/saxophone/fitting and jazz/trumpet/clashing. This was done to avoid carryover effects, such as participants speculating on the purpose of the study, when hearing the same piece with the same instrument for a second time.

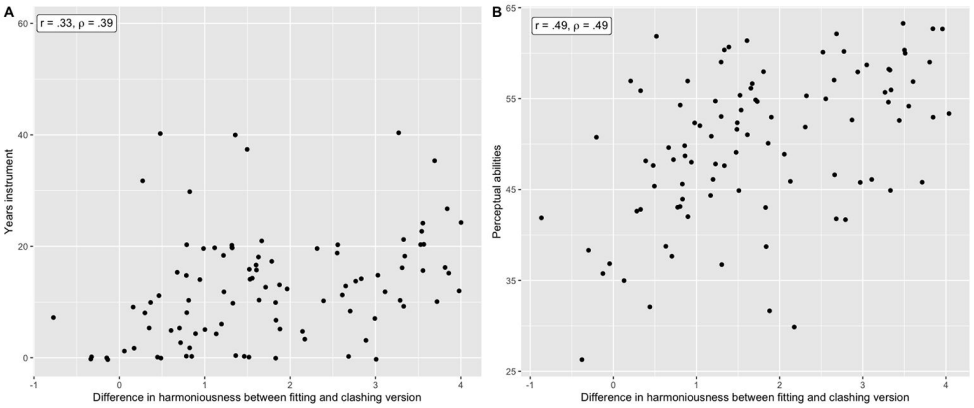
We therefore calculated the following difference score with a mean of two stimuli rating scores each, where genres and instruments were counterbalanced:

mean(fitting stimuli) – mean(clashing stimuli)

This difference score correlated positively with the participants' years of instrumental or voice practice and their perceptual abilities with a medium effect size each, both for parametric (years practice: $r = .33$, $p < .001$; perceptual abilities: $r = .49$, $p < .001$) as well as non-parametric tests (years practice: Spearman's $\rho = .39$, $p < .001$; perceptual abilities: $\rho = .49$, $p < .001$; see also Figure 2; the non-parametric correlation coefficient was included due to the visually detected outliers in Figure 2A).

Figure 2

Distribution of Harmoniousness Ratings With Regards to the Participants' Perceptual Abilities



Discussion

The present study investigated, using an expertise-unrelated task, the impression of participants when listening to bitonal pieces music with fitting or clashing keys for melody and accompaniment. The dependent variable consisted of three items to be judged on rating scales that represent an overall rating in harmoniousness. We observed a large difference in rating between the fitting vs. clashing stimuli (Hypothesis 1 confirmed) and a medium difference between the classical and the jazz piece. The years of accumulated daily music practice and the perceptual abilities were correlated with the responses to the extent of a medium effect size each (Hypotheses 2 and 3 confirmed).

We clearly anticipated differences in ratings between stimuli with a clash of keys compared to fitting keys from the results of previous research and can clearly corroborate this finding in our study. More specifically concerning genre difference, the results from Kopiez and Platz (2009) in either genre were replicated in our study: The rated harmoniousness in the clashing versions was much lower for the classical piece than the jazz piece (d = 0.96), while the difference between the fitting versions was smaller (d = 0.41). While there was no overall effect of instrument, the latter effect originated mostly from different ratings for the saxophone versions, however, these post-hoc explanations are only a numerical tendency and must be interpreted with caution. Whereas the saxophone was originally developed for use in classical orchestras and marching bands, our current listening habits link it much more to jazz than classical music, which might serve as a probable explanation for its higher ratings in the jazz stimulus compared to the classical stimulus. While both selected stimuli are representations of their respective genre, the generalizability of results stays debatable, especially since each

genre was only represented by one musical piece. Further studies might use different stimuli or, even better, multiple stimuli from several genres to increase the stability of the measurement and better grasp the diversity of musical culture.

This study found a rating difference between musicians and non-musicians similar to the outcomes in the clash of keys rating task by Hamamoto et al. (2010) and the correlations between several musical sophistication scales and the mistuning perception test by Larrouy-Maestri et al. (2019). As hypothesized, the self-reported perceptual abilities showed a relevant correlation with the harmoniousness rating and this was, in comparison, significantly larger than the correlation with someone's years of musical training, which was the sole criterion for musical expertise in the previous clash of keys studies. Evidently, music-specific perceptual abilities, which can be cultivated, among others, by playing an instrument, but also by other musical activities, seem more relevant than music training to detect various features in a piece of music.

As mentioned in the introduction, in the study by Lalitte et al. (2009) music training did not explain participant behaviour in the segmentation task or the observed arousal. Furthermore, Cook even argues that when instructing musicians and non-musicians with vague tasks, they are "engaged in different tasks" (Cook, 1994, p. 69), most obvious in experts' active and frequently-used technical language in comparison to amateurs' lower skills in such a technical jargon, that is, possessing a probably less active (and more passive) vocabulary and, lastly, very different habits and degrees of freely describing and—even more—criticizing music. Studies on the perception of bitonality using the probe tone paradigm cannot add to the body of evidence here, since Krumhansl and Schmuckler (1986) and Thompson and Mor (1992) only collected data from musically trained participants.

Grouping the evidence on the perception of clash of keys so far, there seem to be three major approaches for operationalisation using a) a free response task, possibly varied with a directed or non-directed instruction and focusing on a declarative detection of a clash of keys, b) a procedure derived from the probe tone paradigm, i.e., a twelve-fold presentation of each stimulus with a potential for carry-over effects, and c) continuous ratings using adjective scales focusing on a general and non-dichotomous, but continuously scalable impression, which lead to less distinct results and do not show "what people do hear" (Wolpert, 2000, p. 225). Concerning the participants' strategies when replying to a free response task, it might be insightful to investigate the occurring thoughts in musicians and non-musicians in a qualitative study, for example whether the detection of a clash of keys is a sudden realisation or the result of an incremental analysis forming over time. In either scenario, musicians tend to be at an advantage by possessing instruments for analysis—even considering basic knowledge about keys and typically dissonant intervals, such as the major second—and by holding more confidence in voicing sudden and rather implausible suspicions, such as the audacity of music researchers in producing an experimental stimulus employing a concept as peculiar

as bitonality. An operationalisation using a non-dichotomous rating procedure might introduce a fairer study setting for participants with less musical expertise, as has been shown here and by Hamamoto et al., 2010.

At the same time, missing clashes of keys and not being able to verbalise such a phenomenon, as shown in studies with a free response task, might even be commonplace: Anglada-Tort and Müllensiefen (2017) presented the identical interpretation of a section of a Bruckner symphony three times paired with three different conductor portraits and asked for various ratings, e.g., liking, technical aspects, and emotional quality, as well as free-text replies for a description of performances and the participant's enjoyment. In 85% of cases, participants did not report the sameness of the musical pieces nor reply in a uniform way to the rating scales, they most often missed the fact that all three stimuli were identical. This study is an example for the influence of a convincing piece of auxiliary information and corroborates, along with studies mentioned in the literature review by Anglada-Tort and Müllensiefen (2017), the need to study figures or "situations" of authority, such as the mere setting of an empirical study, and whether this has the capacity to mute the perception or detection of unexpected stimulus features.

This study did not offer clear-cut detection rates for musicians and non-musicians but rather approximated the detection of a clash of keys with a response possibility that needs less conviction in participants. Due to self-selection of participants and recruiting from a rather music-savvy (but not necessarily musically trained) population, it is quite likely that even participants with little or no musical training were above average interested in music and could therefore not be classified as true non-experts in music. Second, as this study was an online study we had no control over the actual sound quality produced for participants. Self-reported items on the quality of the stimuli and the listening condition were rated very well, which, however, cannot be equated with a controlled listening situation in a laboratory study. Further research can focus on the range of methodical approaches (a free response task to quantify detection rates versus a rating task to quantify metrical differences, possibly accompanied by a qualitative survey) and how results from such operationalization differences interact with groups of different expertise.

Funding: The authors have no funding to report.

Acknowledgments: The authors have no additional (i.e., non-financial) support to report.

Competing Interests: The authors have declared that no competing interests exist.

Author Contributions: *Anna Wolf*—Conceptualization | Data curation | Formal analysis | Methodology | Project administration | Supervision | Visualization | Writing – original draft | Writing – review & editing. *Bastian Wüst*—Conceptualization | Investigation | Methodology | Resources.

Ethics Statement: The present study was conducted in accordance with ethical principles and standards according to the guidelines of the German Society for Psychology. According to German law, no ethics approval has been required. Informed consent was provided by all participants, and they had the option to cease participating in the study at any time without any negative consequences.

Data Availability: The research data and stimuli for this article are available on the OSF (see Wolf, 2025).

Supplementary Materials

For this article, research data and stimuli are available (see Wolf, 2025).

Index of Supplementary Materials

Wolf, A. (2025). *Detection of clash of keys* [Data, code, stimuli]. OSF. https://osf.io/sj3da

References

Abell, N., Springer, D. W., & Kamata, A. (2009). *Developing and validating rapid assessment instruments*. Oxford University Press.

Anglada-Tort, M., & Müllensiefen, D. (2017). The repeated recording Illusion. *Music Perception, 35*(1), 94–117. https://doi.org/10.1525/mp.2017.35.1.94

Bigand, E., & Poulin-Charronnat, B. (2006). Are we experienced listeners? A review of the musical capacities that do not depend on formal musical training. *Cognition, 100*(1), 100–130. https://doi.org/10.1016/j.cognition.2005.11.007

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

Cook, N. (1994). Perception: A perspective from music theory. In R. Aiello & J. A. Sloboda (Eds.), *Musical perceptions* (pp. 64–95). Oxford University Press.

Cook, N. (1987). The perception of large-scale tonal closure. *Music Perception, 5*(2), 197–205. https://doi.org/10.2307/40285392

Ellis, W. D. (1938). *A source book of Gestalt psychology*. Routledge. https://archive.org/details/in.ernet.dli.2015.198039

Fricke, J. P. (2014). Intonation in der abendländischen Musik [Intonation in Western music]. In C. Reuter & W. Auhagen (Eds.), *Kompendien Musik. Musikalische Akustik* (pp. 126–132). Laaber Verlag.

Hake, R., Bürgel, M., Nguyen, N. K., Greasley, A., Müllensiefen, D., & Siedenburg, K. (2024). Development of an adaptive test of musical scene analysis abilities for normal-hearing and hearing-impaired listeners. *Behavior Research Methods, 56*(6), 5456–5481. https://doi.org/10.3758/s13428-023-02279-y

Hamamoto, M., Botelho, M., & Munger, M. P. (2010). Non-musicians' and musicians' perception of bitonality. *Psychology of Music, 38*(4), 423–445. https://doi.org/10.1177/0305735609351917

Harrison, D. (1997). Bitonality, pentatonicism, and diatonicism in a work by Milhaud. In J. M. Baker, B. D. W, & J. W. Bernard (Eds.), *Music theory in concept and practice* (pp. 393–408). University of Rochester Press.

Kaiser, H. F., & Rice, J. (1974). Little jiffy, mark IV. *Educational and Psychological Measurement, 34*(1), 111–117. https://doi.org/10.1177/001316447403400115

Karno, M., & Konečni, V. J. (1992). The effects of structural interventions in the first movement of Mozart's symphony in G minor K. 550 on aesthetic preference. *Music Perception, 10*(1), 63–72. https://doi.org/10.2307/40285538

Kopiez, R., & Platz, F. (2009). The role of listening expertise, attention, and musical style in the perception of clash of keys. *Music Perception, 26*(4), 321–334. https://doi.org/10.1525/mp.2009.26.4.321

Krumhansl, C. L., & Schmuckler, M. A. (1986). The Petroushka chord: A perceptual investigation. *Music Perception, 4*(2), 153–184. https://doi.org/10.2307/40285359

Lalitte, P., Bigand, E., Kantor-Martynuska, J., & Delbé, C. (2009). On listening to atonal variants of two piano sonatas by Beethoven. *Music Perception, 26*(3), 223–234. https://doi.org/10.1525/mp.2009.26.3.223

Larrouy-Maestri, P. (2018). "I know it when I hear it": On listeners' perception of mistuning. *Music & Science, 1*, 1. 17. https://doi.org/10.1177/2059204318784582

Larrouy-Maestri, P., & Morsomme, D. (2014). Criteria and tools for objectively analysing the vocal accuracy of a popular song. *Logopedics, Phoniatrics, Vocology, 39*(1), 11–18. https://doi.org/10.3109/14015439.2012.696139

Larrouy-Maestri, P., Harrison, P. M. C., & Müllensiefen, D. (2019). The mistuning perception test: A new measurement instrument. *Behavior Research Methods, 51*(2), 663–675. https://doi.org/10.3758/s13428-019-01225-1

Law, L. N. C., & Zentner, M. (2012). Assessing musical abilities objectively: Construction and validation of the Profile of Music Perception Skills. *PLoS One, 7*(12), Article e52508. https://doi.org/10.1371/journal.pone.0052508

Leiner, D. J. (2024). *SoSci Survey* (Version 3.7.00) [Computer software]. https://www.soscisurvey.de

Mazzola, G. (2017). Auditory physiology and psychology. In G. Mazzola (Ed.), *The topos of music IV: Roots* (pp. 1353–1368). Springer. https://doi.org/10.1007/978-3-319-64495-0_2

Mongoven, C., & Carbon, C.-C. (2017). Acoustic Gestalt: On the perceptibility of melodic symmetry. *Musicae Scientiae, 21*(1), 41–59. https://doi.org/10.1177/1029864916637116

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

Oehler, M. (2014). Auditorische Szenenanalyse [Auditory scene analysis]. In C. Reuter & W. Auhagen (Eds.), *Kompendien Musik. Musikalische Akustik* (pp. 195–217). Laaber.

Omigie, D., Dellacherie, D., & Samson, S. (2014). Effects of learning on dissonance judgments. *Journal of Interdisciplinary Music Studies, 8*, 11–28.

Parncutt, R., & Hair, G. (2018). A psychocultural theory of musical interval: Bye bye Pythagoras. *Music Perception, 35*(4), 475–501. https://doi.org/10.1525/mp.2018.35.4.475

Schaal, N. K., Bauer, A.-K. R., & Müllensiefen, D. (2014). Der Gold-MSI: Replikation und Validierung eines Fragebogeninstrumentes zur Messung Musikalischer Erfahrenheit anhand einer deutschen Stichprobe [The Gold-MSI: Replication and validation of a questionnaire instrument for measuring musical sophistication, based on a German sample]. *Musicae Scientiae, 18*(4), 423–447. https://doi.org/10.1177/1029864914541851

Thompson, W. F., & Mor, S. (1992). A perceptual investigation of polytonality. *Psychological Research, 54*(2), 60–71. https://doi.org/10.1007/BF00937134

Wertheimer, M. (1923). Untersuchungen zur Lehre von der Gestalt. II [Research into Gestalt theory]. *Psychologische Forschung, 4*(1), 301–350. https://doi.org/10.1007/BF00410640

Whittall, A. (2001). *Bitonality*. Oxford University Press. https://doi.org/10.1093/gmo/9781561592630.article.03161

Wolpert, R. S. (2000). Attention to key in a nondirected music listening task: Musicians vs. nonmusicians. *Music Perception, 18*, 225–230. https://doi.org/10.2307/40285910

Jahrbuch Musikpsychologie (JBDGM) is the official journal of the German Society for Music Psychology (DGM).

PsychOpen GOLD is a publishing service provided by the Leibniz Institute for Psychology (ZPID), Germany.