Repository of Hochschule Hannover

HOCHSCHULE
HANNOVER
UNIVERSITY OF
APPLIED SCIENCES
AND ARTS

Confusingly Similar: Discerning between Hardware Guitar Amplifier Sounds and Simulations with the Kemper Profiling Amp

Nina Düvel, Reinhard Kopiez, Anna Wolf, Peter Weihe

Suggested citation:

Düvel, Nina, Reinhard Kopiez, Anna Wolf, and Peter Weihe. 2020. "Confusingly Similar: Discerning between Hardware Guitar Amplifier Sounds and Simulations with the Kemper Profiling Amp." *Music & Science* 3. https://doi.org/10.25968/opus-3081.

Abstract

Over the last decades, the simulation of musical instruments by digital means has become an important part of modern music production and live performance. Since the first release of the Kemper Profiling Amplifier (KPA) in 2011, guitarists have been able to create and store a nearly unlimited number of "digital fingerprints" of amplifier and cabinet setups for live performances and studio productions. However, whether listeners can discriminate between the sounds of the KPA and the original amplifier remains unclear. Thus, we constructed a listening test based on musical examples from both sound sources. In a first approach, the psychoacoustic analysis using mel-frequency cepstrum coefficients (MFCCs) revealed a high degree of timbre similarity between the two sound sources. In a second step, a listening test with N = 177 showed that the overall discrimination performance was d' = .34, which was a rather small difference ($0.0 \leq d' \leq 0.74$). A weak relationship between the degree of general musical sophistication and discrimination performance was found. Overall, we suggest that listeners are rarely able to assign audio examples to the correct condition. We conclude that, at least on a perceptual level, our results give no support for a commonly accepted pessimistic attitude toward digital simulations of hardware sounds.

Terms of use **CC BY-NC 4.0**

This document is made available under these conditions:
Creative Commons - CC BY-NC - Namensnennung - Nicht kommerziell 4.0 International
For more information see:
https://creativecommons.org/licenses/by-nc/4.0/deed.de

Check for updates

sempre:
Society for Education, Music and Psychology Research
MUSIC & SCIENCE

Research Article

Music & Science
Volume 3: 1–16
© The Author(s) 2020
Article reuse guidelines:
sagepub.com/journals-permissions
DOI: 10.1177/2059204320901952
journals.sagepub.com/home/mns
SAGE

Confusingly Similar: Discerning between Hardware Guitar Amplifier Sounds and Simulations with the Kemper Profiling Amp

Nina Düvel[1], Reinhard Kopiez[1], Anna Wolf[2] and Peter Weihe[1]

Abstract

Over the last decades, the simulation of musical instruments by digital means has become an important part of modern music production and live performance. Since the first release of the Kemper Profiling Amplifier (KPA) in 2011, guitarists have been able to create and store a nearly unlimited number of "digital fingerprints" of amplifier and cabinet setups for live performances and studio productions. However, whether listeners can discriminate between the sounds of the KPA and the original amplifier remains unclear. Thus, we constructed a listening test based on musical examples from both sound sources. In a first approach, the psychoacoustic analysis using *mel-frequency cepstrum coefficients* (MFCCs) revealed a high degree of timbre similarity between the two sound sources. In a second step, a listening test with $N = 177$ showed that the overall discrimination performance was $d' = .34$, which was a rather small difference ($0.0 \leq d' \leq 0.74$). A weak relationship between the degree of general musical sophistication and discrimination performance was found. Overall, we suggest that listeners are rarely able to assign audio examples to the correct condition. We conclude that, at least on a perceptual level, our results give no support for a commonly accepted pessimistic attitude toward digital simulations of hardware sounds.

Keywords

Kemper profiling amp, guitar amplifier, timbre, sound synthesis

Submission date: 29 July 2019; Acceptance date: 24 December 2019

As a widely accepted premise, popular music in the 21st century is unthinkable without electronic technology (see, for example, Théberge, 2001). The development of new technologies has been particularly relevant for the electric guitar. Starting in the 1990s, the development of powerful digital audio processing tools has fostered the extension of sound design for rock guitarists and freed their sound from a particular combination of amplifiers and cabinets. Three main technological approaches can be described. First, collections of sampled sounds (e.g., the *Ample Sound* series or the *Virtual Guitarist*) are based on thousands of samples of a selected guitar model and can be played by any MIDI instrument input. Second, the modeling approach (e.g., the *Tech21* by SansAmp or the *Line 6 POD* by Roland) processes the instrumental input through convolution algorithms of a selection of fixed built-in amplifiers, cabinets, and microphones (for technical details, see Eichas & Zölzer, 2018; for a historical overview of modeling amps, see Herbst et al., 2018). The manufacturers and engineers created these presets using physical modeling in that they, roughly speaking, found mathematical equations to describe the change from the input to the output signal. Details of this modeling procedure are described by Pakarinen and Yeh (2009). The third and most recent

[1] Hanover Music Lab, Hanover University of Music, Drama and Media, Hanover, Germany
[2] Institute for Systematic Musicology, Universität Hamburg, Hamburg, Germany

Corresponding author:
Reinhard Kopiez, Hanover Music Lab, Hanover University of Music, Drama and Media, Neues Haus 1, 30175 Hanover, Germany.
Email: reinhard.kopiez@hmtm-hannover.de

Creative Commons Non Commercial CC BY-NC: This article is distributed under the terms of the Creative Commons Attribution-NonCommercial 4.0 License (https://creativecommons.org/licenses/by-nc/4.0/) which permits non-commercial use, reproduction and distribution of the work without further permission provided the original work is attributed as specified on the SAGE and Open Access pages (https://us.sagepub.com/en-us/nam/open-access-at-sage).

development comes from so-called *profiling*, which uses the acoustical characteristics of the original amplifier for the processing of guitar or bass input signals. Contrary to sampling and modeling approaches, profiling is not restricted to pre-configured guitar sounds but open to the generation of customized guitar sounds using the entire signal path of amplifiers, effects, speakers, and microphone combinations. In other words, a profiling amp can capture the sound of any combination of components. The sound profiling is realized by tools such as the Kemper Profiling Amp (KPA), which was introduced to the market in 2011 and described by its manufacturer as:

> the first digital guitar amp to *really nail* [*sic*] the full and dynamic sound of a guitar or bass amp. This is made possible by a radical, patented technology and concept which we call "PROFILING™". It offers an extensive preset of sound profiles of amplifiers and cabinets to the user but also enables the recording of own [*sic*] sound profiles. This collection of profiles is continuously extended by the KPA user community. (Kemper GmbH, 2019)[1]

The difference between modeling and profiling can be explained in the words of the KPA inventor, Christoph Kemper:

> Modelling, in the technical sense, is bringing the physics of the real world into a virtual world, mostly by defining formulas for the real world and letting them calculate on a real-time computer. [. . .] In contrast, profiling is an automated approach for reaching a result that is probably too complex and multi-dimensional to achieve by ear, or by capturing the behaviour of individual components in isolation. (Greeves, 2012)

During the profiling process, the KPA

> sends various tones and signals into the reference amp — it will sound like warbles and static at various pitches and intensities [. . .] These dynamically changing sounds allow the Profiler to learn about the non-linear behavior of the tube architecture, and the dimensions of the passive components in the original amp. The Profiler then listens to how the reference amp reproduces these sounds, and analyzes the results. These characteristics are then recreated in the virtual signal flow of the Profiler. (*Kemper Profilier: Profiling guide*, 2015, p. 10)

Christoph Kemper protected his invention by several patents which explain the purpose and the operating principle of the KPA (Kemper, 2009, 2015a, 2015b, 2015c, 2017). The testing stimulus for the profiling procedure is being mentioned in the patents but not described in detail. Therefore, the exact operating principle of the profiling process remains a company secret, and more insight into the technical details of the built-in profiling algorithms of the KPA is not available.

The profiling approach elicited enthusiastic reactions in the community of rock guitarists. For example, in an early product review, Vinnicombe (2012) wrote about the KPA: "A truly revolutionary piece of kit for serious recording guitarists and producers (. . .) you can grab the (hungry man's) lunchbox-sized KPA and jet off to a studio or gig halfway around the world with your signature sound in your carry-on luggage." The author concluded, "We've heard various approaches to digital modelling sound good in the studio before but this is as close to a 'real' mic'd valve amp sound as to be indistinguishable." In another statement, Andy Sneap (2012, 3:40, 10:11), music producer and member of *Sabbat*, stated: "We actually could not believe what this thing did, how accurate it was [. . .] I can't tell the difference." Another statement shows that the KPA has become an indispensable device not only for studio work but also for musicians on tour, such as Nathan Spicer, a guitarist who toured with Katy Perry. This view on new technologies is in line with those of other authors who emphasize that the most important challenge for popular musicians is the transition of the produced work from studio to the live stage (Herbst et al., 2018, p. 499), but with the profiling technology, any studio sound can easily be replicated onstage. Spicer stated, "I never had to use my backup, totally reliable" (Spicer, 2015, 4:15–4:20). In general, musicians are convinced of the benefits of the KPA in studio music production, which include producing reproducible sounds, finding sounds instantly without a new setup, being adequate for home recording in limited room sizes environments, and significantly improving the workflow.

However, as reported by Herbst (2019), not all rock music producers embrace the new modeling technique, and a certain degree of skepticism remains. For example, some producers subjectively feel that digitally produced sounds might be less transparent in the mix (for a short summary of the ongoing discourse on discussion boards, see Herbst et al., 2018). Mynett (2017, p. 57) noted that although the current level of amp simulation is very high and can thus provide a very close emulation of amps, cabinets, and microphones with credible equivalence, critics disapprove of those hardware simulations due to a lack of "natural air" in the profiled sound.

As mentioned by Herbst et al. (2018), in contrast to the high practical relevance of the profiling technology for music production, there is currently no academic debate on the potential influence of this technique on musical practice and perceptual aspects; it remains unclear whether listeners can discriminate between the sounds of the KPA and the original amplifier.[2] Our search in relevant databases and journals, such as the *Journal of the Audio Engineering Society*, the *Journal of the Acoustical Society of America*, PsycINFO, and ProQuest, showed very limited results: Only two relevant publications with a research focus on perception and acoustics of profiled sounds from the KPA were found. In an exploratory approach, Majewski

and Malecki (2015) compared sounds produced by an original amplifier with those based on the KPA emulation of the same amp. The authors used re-amping technology, the recording of a clean guitar signal using direct in (DI) and playback of the cleanly recorded track through a miked guitar amp/speaker (see Huber & Runstein, 2010, p. 147) to keep the input signal for both amplifiers constant. Musical examples were based on three short arrangements with electric guitar, bass guitar, and drums with the guitar playing a lead passage, a rhythm, and a solo. Acoustical analysis of the final recordings of both amplifiers (RMS energy, waveform, and spectrum) revealed small differences. An additional listening experiment with eight guitar experts rating the sound characteristics (e.g., naturalness or roughness) was conducted. Majewski and Malecki (2015, p. 4) concluded that "the emulated sound is not the same as the recorded, sampled one." No quantitative indicators (i.e., perceptual or acoustical) for the dissimilarity were given; thus, findings remained vague. In their pioneering study, Herbst et al. (2018) used a set of 71 low-level timbre descriptors (Siedenburg et al., 2016) for the analysis and comparison of a comprehensive set of sound examples both from original hardware amplifiers and profiled sounds: 16 guitar amps were recorded (re-amping method) in combination with 14 stimuli (e.g., single chords, single notes) and various sounds (clean, overdriven, and distorted), resulting in a total of 1,344 recordings and 95,424 test values. In total, between-groups analyses (original vs. profiled sounds) of timbral differences revealed a high degree of timbral similarity: Only 8% of 71 timbre descriptors showed significant differences. For example, profiled sounds showed more loudness (RMS Gammatone 1, partial $\eta^2 = .088$, medium effect size) and were characterized by more non-periodic noise (Herbst et al., 2018, p. 491). Although no systematic interactions were found, the quality of the profiles also seemed to depend on the particular amplifier model. Even though the quality of the profiles was very good, the authors concluded that future studies should consider perceptual tests with direct comparison of sound sources and stimuli played in a musical context. This research desideratum mentioned by Herbst et al. was one point of reference for our study.

Although Mynett (2017, p. 57) stated that the current level of amp simulation is very high and can thus provide a very close emulation of amps, cabinets, and microphones with credible equivalence, we cannot exclude that expert listeners might be able to discriminate between the two sound sources of an original amp versus a profiled amp. Previous research has demonstrated that the identification of sources of orchestra music produced by live orchestras versus virtual realizations based on pre-recorded orchestra sample libraries had an average success rate of 72.5%. Subgroups of listeners with high sound-discrimination expertise (e.g., conductors or recording engineers) reached correct identification rates of 80% on average (Kopiez et al., 2016). In other words, the discrimination between sound sources might be a difficult, but possible task, even in subgroups of experts.

Research Aims

This study aimed to contribute to the field of research regarding the distinguishability between (a) original or live instruments and (b) simulations or models of their sound. This field of research has an ongoing history and has examined, for example, whether listeners can distinguish a Stradivarius from a new violin (Fritz et al., 2017; Levitin, 2014) or a recording from a live band (Sharples, 2017). The previously outlined study by Kopiez et al. (2016) also falls into this field of research.

In our study, the research aims were threefold. First, we investigated the perceptual discrimination and sensitivity for original guitar amplifier sounds versus profiled sounds. We assumed that a participant's discrimination performance would be influenced by domain-specific expertise in the field of rock guitar performance and music production. Second, as suggested by Herbst et al. (2018), we used a more holistic context: The stimuli were selected from various musical contexts and musical genres (instead of isolated chords or notes). Third, although the statistical analysis of our research was guided by a classical null-hypothesis significance testing (Cumming & Calin-Jageman, 2017), we also used methods to allow the unbiased test of the null hypothesis of non-discriminability between the original and the profiled sounds as enabled by Bayesian statistics (Kruschke, 2015).

Method

We conducted an online study in which participants had to classify individual audio examples that were created either by the original amp (OA) or the simulation of the Kemper Profiling Amp (KPA).

Materials and Stimuli

The audio examples used in this study were created by the fourth author and had previously been used for a product review (Weihe, 2012). The stimuli stemmed from two different rock guitar setups: The first one was a 1968 Marshall JMP "Super Lead" 100 Watt amp in combination with two 4×12 cabinets (one from the mid-1960s Pinstripe and one from the early 1970s) connected to a Fender Stratocaster guitar (1962 Candy Apple Red Strat). The microphones were Electro-Voice RE 20, Neumann U89, Royer R121, and Sennheiser MD441 in combination with a V72 preamp. In the second setup, a Vox AC30 was combined with a 1953 Fender Telecaster guitar. Here, the microphones were a Neumann U87i, a Royer R121, and a Sennheiser MD441 in combination with a V72 preamp (for technical details of the OA recording setup, see Figure S1, and for a list of the

used hardware, microphone positions, and pickup combinations, see Supplement A in the online supplement).

Next, simulations (profiles) of both setups were created with the KPA. In contrast to the previous studies by Majewski and Malecki (2015) and Herbst et al. (2018), we decided not to use a re-amping procedure. The decision for this atypical recording setup (no re-amping) was motivated by the results of careful listening pre-tests: When the guitar signal was looped through the KPA into guitar amplifier, a significant deterioration of the sound became salient. This sound distortion was due to multiple conversions of impedance between the mixing console and guitar amplifier as well as the internal buffer and A-D converter of the KPA. The negative effects of a missing direct connection between guitar and amplifier have been confirmed by modern physical approaches to the sound-influencing parameters of the electric guitar: As explained by Zollner (2007, pp. 10-4–10-6), there is a complex electrical feedback and non-linear interaction between the input impedance of the amplifier tubes and the electrical output signal of the guitar pickup with a strong impact on the characteristic attack of a guitar tone. For example, the guitar is galvanically coupled to the grid of the first tube, and the pickup transmission behavior is influenced by this coupling. However, in the case of re-amping, this coupling gets lost, and the typical feedback loop with its dampening effects on the pickup resonance is omitted (Zollner, 2011–2013, p. 10-318).

After careful consideration, we thus decided not to re-amp and not to use the internal A-D converter of the KPA (for the signal path of the KPA recording procedure, see Figure S3 in the online supplement). Instead, all examples were performed twice with the OA and KPA recording setup (P. Weihe, personal communication, February 19, 2018; for the profiling signal path, see Figure S2 in the online supplement). All digital conversions were processed by the Lavry Gold AD122 A-D converter. The guitarist performed and heard the miked sounds in the control room via the studio monitor at the end of the signal chain, which was the same listening position as for the subjects in our experiment. The amplifiers were located in the soundproof control room. Thus, the direct sound radiation of the OA and KPA could not influence the player.

The entire device setup and profiling was supervised by a product manager from the Kemper GmbH. Six musical examples featuring various musical styles and sounds (within popular music) were first recorded with the original hardware (this condition is called *original amp* [OA] in the following). For the second condition (*Kemper Profiling Amp* [KPA]), the same six musical examples were performed again using the respective digital profile in the KPA. Detailed diagrams of the various recording setups can be found in Figures S1, S2, and S3 in the online supplement. In addition to these 12 (= 6 * 2) test stimuli, two stimuli pairs were created for practicing purposes (see sound examples available from https://osf.io/3jsx6/?view_only=9495be4ffb664a26bc149edbf6039506).

The 12 stimuli had a mean length (excluding silence in the beginning and end of the audio file) of 13.45 sec (SD = 5.93), the OA stimuli (M = 13.87, SD = 7.07) being on average slightly longer than the KPA stimuli (M = 13.02, SD = 4.47). To objectively measure the timbral differences between the two matched stimuli, we conducted a psychoacoustic feature analysis. We decided to use *mel-frequency cepstrum coefficients* (MFCCs) as they seem to be able to sum up characteristic aspects of musical timbre. The MFCCs were developed for automatic speech recognition by Davis and Mermelstein (1980) and have been used for a variety of applications in the music domain. For example, their relevance for timbre analysis of musical stimuli was examined by Logan (2000), Loughran et al. (2008), and Baniya, Lee, and Li (2014). For a comprehensive explanation of the algorithms used for MFCC analysis and further information, see Thiesen et al. (2019).

Due to the noisy and inharmonic spectra of distorted guitar sounds, additional sound descriptors were considered for a more comprehensive characterisation of stimuli (for more details on the audio descriptors, see Caetano et al., 2019):

1. Spectral Entropy. This spectral descriptor measures the relative Shannon entropy of the spectrum and indicates whether the spectrum is dominated by pronounced, isolated peaks or is rather flat and uniform. On the one hand, the greater the value is (with 1 as a maximum), the flatter (similar to noise) the curve is. On the other hand, if the signal is dominated by a single prominent peak (as in the case of a pure tone), it will result in a lower entropy value. The spectral entropy was calculated by means of the software MIRtoolbox V 1.7.2 (Lartillot, 2019). Previous research has shown that parameters from the flux-related dimensions, such as spectral flux, frequency modulation, and spectral entropy, can make sounds more distinct from each other so that they can be better retained in working memory (Golubock & Janata, 2013).
2. Inharmonicity. This spectral descriptor quantifies that instrument sounds are imperfectly periodic. It measures the deviation of the frequencies of the partials from pure harmonics. The inharmonicity was calculated by means of the software MIRtoolbox V 1.7.2 (Lartillot, 2019).
3. Roughness. This spectral descriptor quantifies the sensation of amplitude fluctuation. It starts at a modulation frequency of about 15 Hz and reaches its maximum near modulation frequencies of about 70 Hz (Fastl & Zwicker, 2010, p. 257). As roughness is calculated in arbitrary values in the MIRtoolbox and no time series output is available, we preferred calculation in standardized units (centi

Asper), which is offered by the software dBSONIC V 4.501 ("dBSONIC," 2012; Lartillot, 2019).

4. Loudness. This spectral descriptor quantifies intensity sensations. In contrast to the measurement of sound pressure level (SPL), loudness measurement is based on a psychoacoustic (frequency-dependent) ear model of sound perception (Fastl & Zwicker, 2010, p. 203). Psychoacoustic analyses were conducted in standardized units (sone) by the software dBSONIC V 4.501 ("dBSONIC," 2012).
5. Spectral flux. This spectro-temporal descriptor captures local spectral changes over time and measures the spectral differences of the current frame relative to the previous frame. The parameter is used in timbre space studies and refers to the dimension of spectral irregularity (e.g., McAdams et al., 1995). This type of spectral change is perceptually highly relevant, and in our study, it was calculated by means of the software MIRtoolbox V 1.7.2 (Lartillot, 2019).
6. Zero crossing rate (ZCR). This temporal descriptor gives information on how many times the waveform crosses the zero axis. Every change of sign from negative to positive is counted. Periodic sounds have smaller rates than noisier sounds. According to the description of the MIRtoolbox manual, the ZCR is a "simple indicator of noisiness" of a sound, as it tends to be smaller for periodic sounds than for white noise. It was calculated by means of the software MIRtoolbox V 1.7.2 (Lartillot, 2019). Although the zero crossing rate should be regarded as a description for "sounds with higher frequency," a perceptual evaluation is still missing. Strictly speaking, from a perceptual perspective, the ZCR has no structural relevance for the auditory system as the cochlear does not register ZCRs of the entire waveform.

Procedure and Design

The online survey was conducted on the research platform SoSci Survey (https://www.soscisurvey.de). It was performed in accordance with relevant institutional and national guidelines and regulations (Deutsche Gesellschaft für Psychologie e.V., 2016; Hanover University of Music Drama and Media, 2017) and with the principles outlined in the Declaration of Helsinki. Formal approval of the study by the Ethics Committee of the Hanover University of Music, Drama and Media was not mandatory as the study adhered to all required regulations. The anonymity of participants and confidentiality of their data were ensured. Participants were informed about the objectives and the procedure of the survey as well as their right to withdraw from the study at any time without adducing reasons or having any negative consequences. All participants gave their informed consent online by ticking a checkbox. The listening experiment started with a short introduction, followed by an explanation of the KPA, its purpose, and its operating principle. To prove that they had understood the instructions, participants had to answer a content-related question: Several statements were given, and correct answers had to be marked in a tick box.[3] In case of incorrect answers, participants were guided back to the explanations of the KPA for the re-reading of the instructions.

The *Headphones and Loudspeaker Test* (HALT; Wycisk et al., 2018) was conducted to obtain more information on the sound transmission properties of the playback devices used by the participants and to guarantee comparability of listening conditions. This test comprised four steps. First, a test sound with a number of sections containing white noise of either high or low sound pressure level was presented for the adjustment of the volume. Second, the lower cut-off frequency was determined by counting tone events in between sections of white noise for the determination of the lower cut-off frequency (several stimuli with pure tones from 20 to 180 Hz in steps of 20 Hz were presented). Third, participants located a stimulus either in their heads or around them. The stimulus was created by a dummy head recording. Participants wearing headphones were expected to locate the sound in their heads, whereas participants using a loudspeaker would identify the sound source in the direction of their speakers. Fourth, a counting task with white noise and pure tones was constructed to be different on the right and left channel. The result of the counting task provided insight into whether participants were listening monophonic or stereophonic.

The discrimination task started with two practice trials in which practice pairs of audio examples (the same musical example recorded via KPA and OA) was presented (see sound examples available from https://osf.io/3jsx6/?view_only=9495be4ffb664a26bc149edbf6039506). Participants could listen to both stimuli repeatedly and in arbitrary order. Their task was to determine which audio example was created by which sound source. Immediate feedback was given. In case of a false response, the practice trial could be repeated. Then, a second practice pair of audio examples was presented. Subsequently, individual presentations of the 12 test stimuli followed. Participants listened to each stimulus in isolation and as often as they wanted without feedback. For the control of retest reliability, two stimuli (one from each condition) were presented twice. The stimuli were presented in a fully randomized order, and the same musical excerpt was never presented successively within the 14 (= 6 pairs and 2 retest) stimuli. Participants rated each single stimulus on an 8-point scale, indicating their categorization of the stimulus as produced with the OA (1 to 4) or the KPA (5 to 8) as well as the confidence of their answers (endpoints meaning high confidence; see Figure S4 in the online supplement).

To control for the influence of musical sophistication, participants completed the German version (Schaal et al.,

Table 1. Classification of response patterns for each pair of stimuli; $N = 1062$ is the total number of responses to pairs of stimuli.

Stimulus is…	Participant says…	OA		Sums of rows
		„OA"	„KPA"	
KPA	„OA"	$p_{11} = \frac{a}{N} = \frac{236}{1062} = .222$	$p_{12} = \frac{b}{N} = \frac{192}{1062} = .181$	$\frac{a+b}{N} = \frac{428}{1062} = .403 = prop(FA)$
	„KPA"	$p_{21} = \frac{c}{N} = \frac{323}{1062} = .304$	$p_{22} = \frac{d}{N} = \frac{311}{1062} = .293$	$\frac{c+d}{N} = \frac{634}{1062} = .597 = prop(CR)$
	Sums of columns	$\frac{a+c}{N} = \frac{559}{1062} = .526 = prop(Hits)$	$\frac{b+d}{N} = \frac{503}{1062} = .474 = prop(Misses)$	$1 = 100\%$

2014) of the Goldsmiths Musical Sophistication Index (Gold-MSI; Müllensiefen et al., 2014). Items from the General Factor, Factor 1 (Active engagement), Factor 2 (Perceptual abilities), and Factor 3 (Musical training) were considered.

Finally, we inquired about the demographics and musical background of the participants. Questions on their musical background focused on their musical profession, their electric guitar expertise, and whether they knew about the KPA before participating in this study ("knowledge of the KPA"). The exact wording of the questions and possible answers can be found in Supplement B.

Statistical Power and Required Sample Size

To reduce the probabilities of Type I and Type II errors (false positive and false negative findings), we had to consider the significance level and the statistical power. Thus, the thresholds for the Type I error was set to $\alpha \leq .05$ and for the Type II error to $\beta \leq .20$, resulting in a statistical power of $1–\beta > .80$. As we always presented one stimulus at a time, an A–Not A method (also called one-interval design or yes–no task; Bi, 2015, p. 5; Macmillan & Creelman, 2005, p. 1; Wickens, 2002, pp. 3–6) was used. Because participants evaluated both, KPA as well as OA sound examples, our A–Not A design was a paired design and as participants not only listened to one but to six pairs of stimuli, the design was also called "replicated" (Bi & Ennis, 2001, p. 216).

For a paired design, the effective sample size $\tilde{N}$ is determined by the number of participants N and the number of replications k as well as an adjustment factor C (for calculation see equation 11.2.3 in Bi, 2015, p. 305) by the following equation (Bi, 2015, p. 321):

$$\tilde{N} = \frac{N * k}{C} \quad (1)$$

Due to missing a priori information from previous studies on adjustment factor C as well as the proportions of response patterns (for classification, see Table 1 in this paper and Table 4.5 in Bi, 2015, p. 76), only post-hoc calculations about the required sample size are presented here. The response patterns classified each pair of responses to the given pair of stimuli (matched pairs, not presented simultaneously). Two answering options for both stimuli of one pair resulted in four possible response patterns. The frequencies of the occurrences of these patterns are called a, b, c, and d (Bi, 2015, p. 76). Specifically, the frequency of responses to the OA stimulus with "OA" as well as to the KPA stimulus with "KPA" is represented in c.

Post-hoc analysis of our data revealed that adjustment factor $C = 1$ and therefore $\tilde{N} = N * k$. Considerations concerning power and required (effective) sample size for a paired A–Not A design using McNemar's test (McNemar, 1947) can be found in Bi (2015, pp. 84–87). McNemar's test examines a contingency table with paired data with two levels of the independent variable ("OA" and "KPA") as well as two measurements (one OA and one KPA stimulus) for significant differences. As the same participant responded to both the KPA and the OA stimuli, the data was paired, so McNemar's test was the appropriate choice to test for differences. Therefore, the proportion p_{12} of response pattern c ($p_{12} = \frac{c}{N}$) and the proportion p_{21} of response pattern b ($p_{21} = \frac{b}{N}$) were needed; post hoc calculation revealed $p_{12} = .304$ and $p_{21} = .181$ (see Table 1). The calculation of the required effective sample size $\tilde{N}$ for McNemar's test can be conducted according to two approaches: one by Miettinen (1968; see equation 4.5.7 in Bi, 2015, p. 85) and one by Bennett and Underwood (1970; see Table 4.10 and code "mcnpows" in Bi, 2015, pp. 85–87). Miettinen's approach estimates a required effective sample size of $\tilde{N} = 193$, Bennet and Underwood's approach of $\tilde{N} = 198$. Therefore, our minimum number of participants N could be calculated because our study involved six replications (six pairs of audio examples, $k = 6$): $N_{Miettinen} = \frac{\tilde{N}}{k} = \frac{193}{6} \approx 33$ (rounded up) and $N_{Bennet\ \&\ Underwood} = \frac{\tilde{N}}{k} = \frac{198}{6} = 33$. Therefore, the collection of data from 33 participants would have been sufficient to ensure a power of at least 80% for McNemar's test. In our study, the sample size and therefore the test power exceeded the required minima.

Participants

The invitation to the online survey was circulated to professors and students of German universities of music. Additionally, the announcement was sent through the social media. In the end, $N = 379$ persons took part, and $N = 183$ (about 50%) completed the questionnaire.

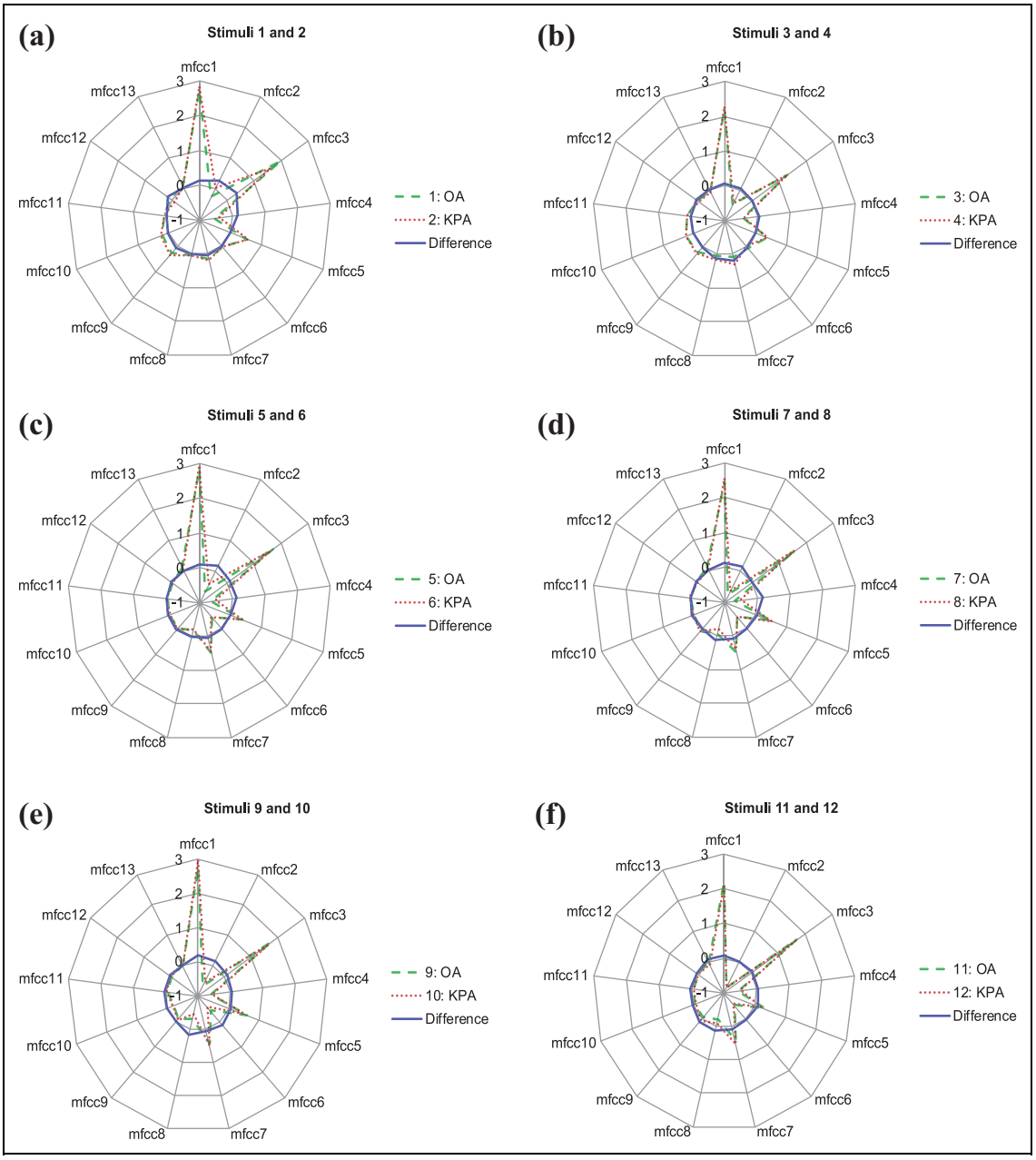

Figure 1a-f. (color online) Radar plot of MFCC analyses (arbitrary values) for the 12 stimuli. The MFCCs of matched stimuli (same piece, different amplifiers) have been integrated into one diagram along with the calculated differences between the values of the stimulus pairs.

Results

Psychoacoustic Properties of the Stimuli

To make the assumed differences in timbre between the OA and the KPA more transparent, we applied the analysis of low-level timbre features to all stimuli. Using the MIR-Toolbox (Lartillot et al., 2008), the first 13 MFCCs as well as spectral entropy, inharmonicity, roughness, loudness, spectral flux (mean and median), and zero crossing rate were calculated for each of the 12 stimuli and the four practice stimuli. The MFCCs as part of the results of the timbre analyses are shown in Figure 1(a–f); analyses of the practice stimuli can be found in Figure S5 in the online supplement.

The differences between the MFCCs for each pair of stimuli were very small. The average difference between all 78 pairs of measurements (= 13 MFCCs * 6 stimulus pairs) was 0.072 units (SD = 0.065 units). As shown in Figure 1, differences between the MFCCs for the OA and the KPA stimuli are close to a circle in the vicinity of zero (which means that the values have almost no difference).

Additional analyses of audio descriptors revealed a similar picture: As shown in Figure S10, the loudness of the audio recordings of the 12 stimuli was very similar as was the psychoacoustic roughness (Figure S11). Although some sections of the recordings showed small perceptual differences between sound sources (e.g., the beginning of Track 11 sounded less distorted when compared to the KPA version of Track 12), the overall differences between the two sound sources OA and KPA were very small for all six psychoacoustic parameters (see Table S10 for details). The descriptive statistics for the relative differences between OA and KPA as shown in Table S11 reveals a high degree of proximity for all six sound descriptors, which is reflected in high relative degrees of similarity (smaller value divided by larger value) close to the maximum of 1 [0.907; 0.994]. However, the interpretation of findings is complex because differences in psychoacoustic features between KPA and OA did not show a consistent pattern: For example, for roughness, the KPA showed higher values in three stimulus pairs (see Table S10), the OA for two pairs, and one stimulus pair (11 and 12) showed nearly equal values. For the parameter loudness, the KPA again showed higher values in three stimulus pairs; however, they were only partially identical with the pairs from the roughness analysis. Additionally, an interpretation of observed differences (see Tables S10 and S11) in terms of JNDs (Just Noticeable Differences) did not apply: First, JNDs are defined for signal qualities of defined characteristics but not for signals with high spectral complexity as in our case (Fastl & Zwicker, 2010). For example, as explained by Daniel (2008, p. 264), "to perceive roughness differences, the degree of modulation of a sinusoidal amplitude-modulated sound has to be changed by about 10% resulting in a just noticeable roughness difference of about 17%." But the production of our stimuli was far away from the production of controlled variations. Second, JNDs depend on the absolute sound pressure level, with higher SPLs resulting in lower JNDs (Fastl & Zwicker, 2010, Ch. 7). This means that a verbatim interpretation of observed differences between KPA and OA would require listening tests under controlled (calibrated) conditions, which is incompatible with an internet experiment. Thus, in our study, differences of psychoacoustic features should be interpreted based only on careful ocular inspection but not in terms of absolute, laboratory-based JNDs. For future research under controlled laboratory conditions, for instance, JNDs of the roughness and loudness for our musical stimuli could be determined in line with the procedure suggested by You and Jeon (2008) for technical sounds.

On the basis of psychoacoustical analyses, we thus concluded that the timbre of the stimuli was very similar. In order to examine whether these differences were of perceptual relevance and sufficient for listeners to discriminate between the two recording conditions and sound sources, we conducted a listening study, which is described in the following sections.

Data Preparation

Six of the 183 participants showed a highly negative sensitivity ($d' \leq -1.4$), which most likely resulted from a misunderstanding of the task or an intentionally wrong assignment of the audio examples to the two conditions (z-test results in $p < .05$). Those outliers were therefore excluded from data analysis. The participants with a highly positive sensitivity were not classified as outliers by the same procedure and remained in the data set for analysis as they were probably experts and had a high actual discrimination ability. See Figure S6 in the online supplement for a histogram of the sensitivity of the whole sample.

On average, the resulting $N = 177$ participants were 38.0 years old ($SD = 12.7$ years, range = 16 to 69 years, male = 94.9%, female = 3.4%, diverse and no response = 1.7%). This imbalance of genders was expected because those participants who were interested in the research topic and participated in our study, especially those playing the electric guitar, are in a strongly male-dominated field. This did not result in problems, however, as we had no hypotheses about gender differences in perceptual differences.

Control Variables

Table S1 in the online supplement shows the electric guitar expertise (89.3% played the electric guitar on an amateur, semi-professional or professional level; 39.0% had heard about the KPA before beginning the study, and 50.8% had even used it before) and the degree of musical profession (which encompassed professional guitarists or other professional musicians specializing in popular music). The degree of musical sophistication was measured by the German version of the Goldsmiths Musical Sophistication Index (Gold-MSI; Schaal et al., 2014). We employed the self-report inventory which considers a wide range of aspects of musical sophistication. The data on musical sophistication from a total of 38 items are best modeled by a General Factor and five group factors. For the present study, we employed only the items for the general factor as well as three group factors (Active Engagement, Perceptual Abilities, and Musical Training: in total 32 items). For further analysis, the sum score of each factor was calculated. Results are displayed in Table S2 in the online supplement. The comparison with the possible range of scores and the quartiles of the norm sample (Schaal et al., 2014, p. 445) revealed that the present sample was highly sophisticated in all four factors (General Factor, Active Engagement, Perceptual Abilities, and Musical Training). Scale scores were mostly located above the upper quartile of the norm distribution.

The sound transmission properties of the playback devices used by the participants were measured by the HALT (Wycisk et al., 2018). The distribution of the lower cut-off frequency is displayed in Figure S7 in the online supplement. Most of the audio devices (85.3%) were able

to transmit a test tone of 40 Hz or lower (median lower cut-off frequency: 40 Hz, lower quartile: 20 Hz, upper quartile: 40 Hz).

Results of the stereo versus mono listening situation are shown in Table S3, with most participants (92.1%) listening in stereo to the audio examples (twisted channels included). Finally, the spatial location task indicated the use of headphones for $N = 79$ participants (44.6%) and loudspeakers for $N = 98$ participants (55.4%). For more details and statistical tests, see the last paragraph in Methods ("Influence of Listening Conditions") and Table S9.

Analysis of General Response Behavior

The participants' ratings of all 14 stimuli (six pairs and two retest stimuli) on an 8-point scale is displayed in Figure S8. Most medians were in the middle of the scale between 3 and 6 (endpoints indicating a high confidence, the middle of the scale [4 and 5] indicating low confidence of the judgment). Therefore, participants were generally unsure about their decisions as to the source of the sound. Ratings of the KPA-stimuli (blue in Figure S8) became generally higher, meaning more correct (since points 4 to 8 on the scale indicated a classification as produced with the KPA) than the OA stimuli (orange). As the lower end of the scale stands for the OA and the upper for the KPA, this tendency among participants suggested that they were able to discriminate between the audio examples above chance level.

To test whether participants responded in a different way to OA stimuli than to KPA stimuli, we used McNemar's test. The alternative hypothesis was that the proportion of hits would be larger than the proportion of false alarms (FA). The null hypothesis was that the two proportions would not differ and participants would therefore classify the stimuli at chance level. Table 1 classifies the response patterns for each pair of stimuli according to Table 4.5 in Bi (2015, p. 76).

With this information, McNemar's test was calculated as follows:

$$X_M^2 = \frac{(b-c)^2}{b+c} = \frac{(192-323)^2}{192+323} = 33.32 \qquad (2)$$

The critical value of the X^2 distribution with one degree of freedom is $X^2 \geq 3.84$; therefore, McNemar's test reached significance ($p < .001$). The proportion of hits was significantly larger than the proportions of false alarms.

An exact McNemar's test (using the binomial distribution) resulted in a significance level of $p < .001$ and an estimation of the so-called odds ratio $o = b/c = .594$; 95% CI [.495, .713] (see equation 4.4.14 in Bi, 2015, p. 78). Thus, we could also conclude that the odds ratio o was not equal to 1 (which would have been the case if participants had classified stimuli according to chance level) and the proportion of hits was larger than the proportion of false alarms. In other words, participants showed a task performance above chance level.

The basic idea of Signal Detection Theory (SDT) is that people's accuracy in discrimination tasks is related to four possible response types: hits, correct rejections, misses, and false alarms (Macmillan & Creelman, 2005). The following coding scheme for the given answers was used (see Table S4 for the coding matrix): If an OA was correctly identified as the sound source, the answer was coded as "hit" whereas if a KPA-based sound was identified as OA, the answer was coded as "false alarm." The corresponding indicator of a person's sensitivity and performance in a discrimination task is the so-called "*d* prime" (*d*') value, which is calculated by Equation 3 on the basis of z transformed relative answer proportions (Macmillan & Creelman, 2005, p. 8):

$$d' = z(Hit\ rate) - z(False\ alarm\ rate) \qquad (3)$$

The second indicator of a person's discrimination performance is the response bias c, which describes a participant's tendency to respond on some basis other than merit, showing a tilt toward one response or the other (Macmillan & Creelman, 2005, p. 27). This bias is measured by the index c according to Equation 4:

$$c = -0.5\Big(\big(z(Hit\ rate) + z(False\ alarm\ rate)\big)\Big) \qquad (4)$$

To avoid infinite *d*' values, we converted answer proportions of 0% and 100% in line with the standard recommendations (Macmillan & Creelman, 2005, p. 8): A response rate of 0% was corrected to the constant value of $1/(2\ k)$ and a rate of 100% to the value of $1\text{-}1/(2\ k)$ with $k = 6$ (number of stimuli pairs). For example, a *Hit rate* of 100% (= 1.0) was corrected to a value of 1–1/12 = .917 and a *Hit rate* of 0% to a value of 1/12 = .083. The distribution of *d'* values and bias c are shown in Figure 2 and the descriptive analyses of the proportions of hits and false alarms, sensitivity, and response bias in Table 2.

The mean sensitivity of *d'* = .34 was in the range of a "rather small difference" (*d'* between 0.0 and .74) and far from being "meaningful" (from *d'* = .74 upwards; Bi, 2015, p. 44). The analysis of correct response rates revealed that listeners identified the OA and KPA sound sources with an overall correct response rate of 56.2%.

Concerning the response bias, participants tended to generally classify the audio examples as created by the KPA (since the bias c was positive) but compared to a sample without answering bias, the distribution shifted by only 0.1 standard deviations.

Confidence Ratings of Responses

The confidence of the participants in the A–Not A task of classifying an audio example under one of the two conditions (OA or KPA) was measured on an 8-point rating scale. The end points (1 and 8) represented a very high confidence in one's judgment whereas scale points in the

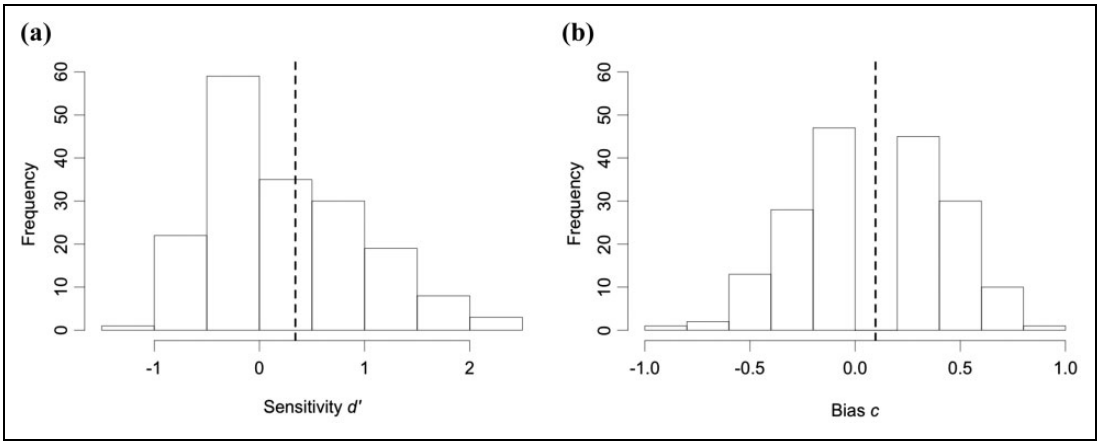

Figure 2a–b. Results from the auditory discrimination task (discrimination sensitivity *d'* and bias *c*) between guitar sounds generated by the original amplifier and simulations via the Kemper Profiling Amp. The vertical dashed lines indicate the respective mean of the distribution.

Table 2. Descriptive analyses of the overall proportions of hits and false alarms, sensitivity, and bias for the total sample.

	M	*SD*	*Min.*	*Max.*
Hit rate (= correct responses to OA stimuli)	.526	.187	.0	1.0
Miss rate (= false responses to OA stimuli)	.474	.187	.0	1.0
Correct rejection (CR) rate (= correct responses to KPA stimuli = CRs = 1 – FAs)	.597	.183	.167	1.0
False alarm (FA) rate (= false responses to KPA stimuli)	.403	.183	.0	.833
Overall proportion of correct responses (OA and KPA)	.562	.143	.250	.917
Sensitivity *d'*	0.343	0.787	-1.383	2.350
Response bias *c*	0.097	0.327	-0.967	0.907

Table 3. Descriptive data of the confidence ratings (1 = high confidence, 4 = low confidence).

	M	*SD*	*Min.*	*Max.*
Confidence rating averaged over all 12 stimuli	2.378	0.698	1.00	3.92
Confidence rating averaged over the 6 KPA stimuli	2.338	0.717	1.00	4.00
Confidence rating averaged over the 6 OA stimuli	2.418	0.755	1.00	4.00

center (4 and 5) showed high uncertainty (see the answer scale in Figure S4).

For data analysis, the right side of the scale (ratings ≥ 5) was mirrored to the lower left side so that 1 equaled high- and 4 low-confidence – regardless of the classification of the sounds as produced by the OA or KPA. The distribution of the mean confidence ratings is shown in Figure S9. The mean overall confidence rating is shown in Table 3. On average, participants tended to be more unsure as to when sounds came from the OA compared to those from the KPA; however, this difference was characterized by a very small effect size (paired *t*-test: $t(176) = -2.26$, $p < .05$; Cohen's $d = -0.17$, 95% CI [–0.318, –0.022]).

It could be predicted that there was a relation between the correct determination of the sound source and the respective confidence rating. As a result, we calculated the correlations between sensitivity *d'* and confidence ratings averaged over all 12 stimuli, the proportion of hits and confidence rating averaged over all 6 OA stimuli, the proportion of false alarms, and the confidence rating averaged over all 6 KPA stimuli. However, correlations were very small ($|r| \leq .1$) and did not reach significance.

Answer Reliability

To control for answer reliability, we tested Stimuli 9 and 12 twice (the retest items were numbered as items 13 and 14). Therefore, correlations between the answers of the same audio-example (rating from 1 to 8 for stimuli 9 and 13, 12 and 14, respectively) were calculated. Overall, the correlation within either stimulus was low; overall correlation: $r(175) = .175$; items 9 and 13: $r = .123$, $p = .052$ (one-tailed); items 12 and 14: $r(175) = .227$, $p < .001$ (one-tailed). Stability of answers is a fundamental question in measuring any construct (Revelle & Condon, 2018). The test–retest reliability depends on the test length, and if only a part of the item is presented twice, the calculated test–retest reliability needs to be corrected according to the *Spearman-Brown prophecy formula* (Equations 5 and 6; Revelle & Condon, 2018, p. 721):

$$r_{xx} = \frac{n \times \bar{r}_{ij}}{1 + (n-1)\bar{r}_{ij}} \quad (5)$$

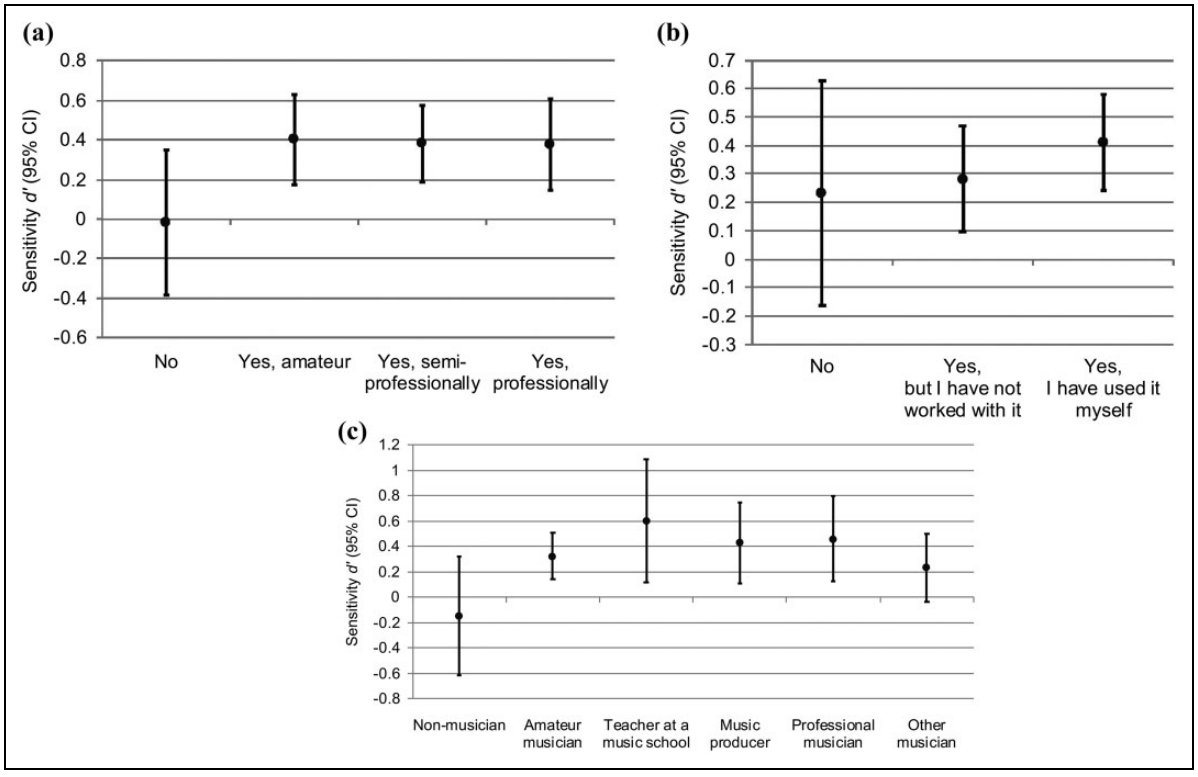

Figure 3a–c. Sensitivity *d'* grouped by electric guitar expertise (a), knowledge of the KPA (b), and musical profession (c).

which means

$$r_{xx\ (6\ \times\ 2\ items)} = \frac{6\ \times\ \bar{r}_{ij}}{1 + 5\ \times\ \bar{r}_{ij}} = \frac{6\ \times\ .175}{1 + 5\ \times\ .175} = .56 \quad (6)$$

Therefore, the "true" overall retest reliability r_{xx} for all 12 items could be estimated as $r_{xx} = .56$. Although the test–retest scores for measurement tools that are used to make decisions about a single individual should have a minimum coefficient of $r \geq .8$, lower reliability is acceptable when individual scale scores are aggregated with groups as units of analysis (Abell et al., 2009, p. 95).

Item Difficulties

To control for the influence of items on response likelihoods, we conducted statistical analyses for all 12 stimuli and 2 retests. As Table S5 shows, the proportion of correct answers ranged between 42.4% (stimulus 11) and 77.4% (stimulus 12).

Differences in Discrimination Sensitivity between Groups of Various Expertise

Relations between a high discrimination performance and the musical background variables were examined using inference and Bayesian statistics. Table S6 and Figure 3(a–c) display the sensitivity *d'* divided into groups of different electric guitar expertise, knowledge of the KPA and musical profession.

To investigate whether the differences between the subgroups occurred by chance or whether these differences were statistically significant, we calculated an ANOVA for each variable. The results are shown in Table S7. The differences between the sensitivity of subgroups according to the three variables – electric guitar expertise, knowledge of the KPA, and musical profession – were not significant; $p > .05$ with $F(3, 173) = 1.504$, $\eta^2 = 0.025$; $F(2, 174) = 0.726$, $\eta^2 = 0.008$; $F(5, 171) = 1.140$, $\eta^2 = 0.032$; respectively.

Although all three analyses of variance showed non-significant differences between subgroups of control variables, null hypothesis significance testing (NHST) offered no direct quantification of the evidence for the null hypothesis (Kruschke, 2015, pp. 335–355; Wagenmakers et al., 2018b, p. 46). Thus, an additional Bayesian approach of statistical analysis was employed. The basic concept hereby is to examine the fit of different models (null and alternative model) to the observed data. The resulting Bayes factor (*BF*) describes how much better a certain model fits the observed data compared to another model. Benchmarks of the BF can be classified as follows: extreme evidence (> 100), very strong evidence (30–100), strong evidence (10–30), moderate evidence (3–10), anecdotal evidence (1–3), and no evidence (0–1) (Wagenmakers et al., 2018a, p. 67) (for further explanation of the underlying logic of Bayesian data analysis, see Kruschke, 2015, pp. 13–32).

Table 4. Correlation of musical sophistication and sensitivity.

Gold-MSI	Pearson's *r*	*p* (one-tailed)
General factor	.143	.029
Factor 1: Active Engagement	.073	.168
Factor 2: Perceptual Abilities	.099	.094
Factor 3: Musical Training	.023	.379

In our case, BF_{01} meant that the probability of the null model was tested against the probability of the alternative model. The three BF_{01} values were as follows: Electric Guitar Expertise: $BF_{01} = 7.96$, Knowledge of the KPA: $BF_{01} = 9.4$, Musical Profession: $BF_{01} = 7.56$ (for details see Table S8). This provided moderate evidence for the validity of the null model for all three independent variables (no differences between subgroups in discrimination performance).

To analyze the relationship between musical sophistication as measured by the Gold-MSI and discrimination performance (sensitivity *d'*), we conducted a correlation analysis (see Table 4). Only the correlation between the degree of general musical sophistication and the sound discrimination sensitivity for the sound sources was found to be significant with a small effect size ($r(175) = .14$, $p = .029$).

Influence of Listening Conditions

To control for the influence of the quality of the playback device (lower cut-off frequency, use of headphones vs. loudspeakers, and listening to monophonic vs. stereo signals) on the discrimination performance of the participant, we conducted a further correlation analysis. However, no evidence was found for a relation between the transmission quality of the audio devices and the discrimination performance (for detailed descriptive data and *t*-tests see Table S9).

Discussion

We investigated participants' ability to discriminate between electric guitar recordings created by an original amplifier and a comparable sound created by the Kemper Profiling Amp (KPA). Our goal was to contribute to the field of research comparing the sound of old or analog instruments against new developments, such as digital recreations. The discrimination ability showed a rather small effect size for this set of stimuli, which was significant due to the high statistical test power of the analysis. Therefore, we conclude that the classification task was quite difficult for the listeners.

Unexpectedly, we found no significant differences between groups of different electric guitar expertise, knowledge of the KPA, or musical profession. The analysis with Bayes statistics revealed moderate evidence for the hypothesis that the groups did not differ in their discrimination ability. The only influence of musical expertise on the discrimination ability could be seen in the Gold-MSI scores: The general musical sophistication showed a small and significant connection to the discrimination ability; however, no causal effect could be stated. No systematic relation between the correctness and the confidence of a rating could be detected. This fact also emphasizes the difficulty of the task: Participants were not able to gauge whether they were giving right or wrong answers although they were performing better than chance.

With our research design, we were able to achieve a higher ecological validity than former studies due to two reasons. First, our stimuli contained musical excerpts that could also occur in "real" music of various styles on a recording or in a live setting. With the choice of the stimuli, we tried to cover possible musical applications (e.g., concerning genre) as well as possible within such a limited number of examples. We employed a holistic approach for the construction of the stimuli and therefore did not aim at either emphasizing the strengths or exploiting the weaknesses of the KPAs profiling procedure. Nevertheless, there were inevitable limitations of the profiling procedure; for example, the stimulus pair 11 and 12 had a string bending sound (continuous pitch change as an expressive means) at the beginning, which was considerably distorted by the KPA. That may be one reason that participants succeeded in classifying this KPA example (Stimulus 12) to the right condition more often than for any other stimulus. Second, we let participants classify one audio example at a time to the two conditions (A–Not A design), which more accurately reflected the way we listen to music in real life: We (almost) never compare and contrast two versions of the same song that were recorded using different equipment.

Considering the complex and non-linear underlying interactions between the various components relevant for the sound production by electric guitars, it remains open whether the "true" sound from the original hardware amplifier will ever be undistinguishable from the sound of simulations. As explained by Zollner (2008–2014, 2011–2013) in the case of modeling amps, although the development of algorithms for amplifier simulations has led to impressive results in the last decade, the non-linearity (and unpredictable electric behaviors) of the tube's grid impedance, its dynamic changes as a function of playing style, the dampening of the pickup resonance by the direct galvanic coupling between guitar and amplifier, the interaction between pickups and the speaker cabinet, and the varying line voltage current will remain a challenge for algorithmic simulations.

Considering this study and former studies as well as other background information, it seems safe to say that the KPA is well able to produce a very good and totally sufficient sound for popular music. Our evidence shows that average listeners would never notice, much less complain about, a bad guitar sound if the guitarist used the KPA in the right way. However, a small group of listeners

sometimes might be able to classify audio examples to their recording conditions, but this happens only in a very controlled environment or when listeners have extensive experience with amps or a high degree of musical sophistication. Nevertheless, in our sample, three participants reached a correct response rate of more than 90%. Such high correct classification rates among expert listeners were also observed in a previous study on the classification of live orchestra sounds versus sample based orchestra realizations (Kopiez et al., 2016): A small group of composers and arrangers reached a correct response rate of more than 80% for the correct classification of sounds. All in all, we conclude that professional guitarists and sound engineers can rely on the KPA for live and recording situations without a noticeable loss of sound quality.

Limitations

The design of our study probably even overestimated the discrimination ability of listeners, as we employed a controlled setting and did not use a cover story: Our stimuli contained just the solo guitar recording and not a whole band playing, which is unusual for popular music. In an embedded sound, the other instruments would mask some of the timbre qualities of the guitar sound. Furthermore, listeners normally do not focus so much on listening to music when consuming it, let alone on a special aspect of the sound. The outcome of this directed listening could be expected to be more accurate than undirected, non-technical, and casual listening.

Although the listening conditions of the participants in our online experiment were controlled for by the HALT procedure, they could not meet the high standards of controlled laboratory settings for psychoacoustic studies based on calibrated equipment. However, as far as the control variables from the HALT showed, the various listening conditions had no influence on the participants' responses (see Table S9). The comparison of our results with those obtained from laboratory settings remain a task for future research. In such a highly controlled setting, researchers would be able to investigate the question of distinguishability between sound sources under optimal listening conditions. In contrast, the design of the present study was a compromise between the demand for control of the listening situation and the advantages of an online procedure (high sample size, high statistical test power, low costs). Therefore, the question of distinguishability under the best conditions remains unresolved, and we conclude that under varying listening conditions, participants can hardly discriminate the sound sources.

Future Perspectives

Further studies should vary the research design to aim either for (a) even higher control concerning the listening conditions (and to answer the question of distinguishability under optimal conditions) or (b) higher ecological validity by including other instruments or by embedding profiled sounds into a band arrangement for the production of musical examples. These studies should consider an even greater number and stylistic variety of musical stimuli. Stimuli could be constructed by expert guitarists, especially to highlight the weaknesses of the profiling procedure, thus revealing limitations of this technology. As an alternative to our holistic approach of listening to entire phrases, another approach would be to use musical "micro units" from full-length sound examples for comparison, for example, only one note, one chord, or a string bending sound. A comparison of those micro sounds could reveal general strengths and weaknesses among the categories. This approach can be illustrated by listening to the beginnings of Stimuli 4 and 12 (both from the KPA) which show a significant discoloring of timbre when compared to the OA versions of the phrases (items 3 and 11). Moreover, different hardware (e.g., other amps) should be used to create more profiles with the KPA for a larger variety of sounds. Another question would be whether participants (of a certain degree of general musical sophistication) could learn to discriminate the two sound sources within a limited amount of time. A learning experiment with test sessions before and after a learning procedure could provide insight.

Nevertheless, there is one important feature lacking in the KPA: It cannot provide a big hardware rack of amps and cabinets on stage, which might be vital for some musical styles of rock music. Live music is always an audio-visual event, and for some musical genres, the visual impression of a KPA might not meet the expectations of the audience. A striking example of the importance of speakers and other hardware on stage might be the "Wall of Sound" invented by Grateful Dead's sound engineer Owsley Stanley (see, for example, Osborne, 2018).

To summarize, the present study is the first one to examine participants' ability to perceptually classify audio examples created by normal guitar amps or the Kemper Profiling Amp to the respective conditions. The average discrimination ability turned out to be slightly above chance level for the employed stimuli, but most of the participants were far from reliably classifying the audio examples to the right condition. Considering the design of this study, it should be safe to say that musicians can rely on the sound of the KPA.

All in all, this research is in line with previous research comparing new digital technology with analog sounds (Kopiez et al., 2016) and with research showing that neither listeners nor expert violinists could reliably distinguish old Italian violins (e.g., Stradivarius) from new violins (Fritz et al., 2017; Levitin, 2014). Neither study supports a general skeptical view that critics sometimes have of new music technology.

Author contribution

ND, RK, and AW planned and conducted the study and analyzed the data. PW constructed and recorded the stimuli. ND wrote a first draft of the manuscript. All authors edited and reviewed the manuscript and approved of the final version.

Declaration of conflicting interests

The authors declared no potential conflicts of interest with respect to the research, authorship, and/or publication of this article.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

ORCID iD

Nina Düvel https://orcid.org/0000-0003-0877-0483
Anna Wolf https://orcid.org/0000-0002-6196-6488

Action Editor

Yoshitaka Nakajima, Kyushu University, Department of Human Science.

Peer review

Daniel Levitin, McGill University, Department of Psychology. One anonymous reviewer.

Supplemental material

Supplemental material for this article is available online.

Notes

1. A short introductory video to the basic principles of the Kemper Profiling Amp and the steps of the profiling procedure can be obtained in English from https://www.youtube.com/watch?v=ZjdaZKpDwdU.
2. Some informal listening and playing tests of guitarists are available on the internet (e.g., https://www.youtube.com/watch?v=INJ_H5PiuTE).
3. The following statements were included: "The KPA has a characteristic sound"; "The KPA can simulate the sound of other amps"; "In the following, I have to decide whether an audio example was produced with an OA or whether the sound was simulated by the KPA"; and "The KPA only works if the OA is still connected to the KPA." Statements 2 and 3 had to be ticked as correct.

References

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

Baniya, B. K., Lee, J., & Li, Z.-N. (2014). Audio feature reduction and analysis for automatic music genre classification. In *Proceedings: 2014 IEEE International Conference on Systems, Man and Cybernetics* (pp. 457–462).

Bennett, B. M., & Underwood, R. E. (1970). On McNemar's test for the 2 x 2 table and its power function. *Biometrics*, *26*, 339–343.

Bi, J. (2015). *Sensory discrimination tests and measurements: Sensometrics in sensory evaluation* (2nd ed.). Chichester, UK: John Wiley & Sons.

Bi, J., & Ennis, D. M. (2001). Statistical models for the A-Not A method. *Journal of Sensory Sturlies*, *16*, 215–237.

Caetano, M., Saitis, C., & Siedenburg, K. (2019). Audio content descriptors of timbre. In K. Siedenburg, C. Saitis, S. McAdams, A. N. Popper, & R. R. Fay (Eds.), *Timbre: Acoustics, perception, and cognition* (pp. 297–333). Springer.

Cumming, G., & Calin-Jageman, R. (2017). *Introduction to the new statistics: Estimation, open science, & beyond*. New York: Routledge.

Daniel, P. (2008). Psychoacoustical roughness. In D. Havelock, S. Kuwano, & M. Vorländer (Eds.), *Handbook of signal processing in acoustics* (pp. 263–274). New York: Springer.

Davis, S. B., & Mermelstein, P. (1980). Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, *28*, 357–366. http://doi.org/10.1109/tassp.1980.1163420

dBSONIC [computer software] (Version 4.501). (2012). 01dB-Metravib. Retrieved from http://www.01db-metravib.com

Deutsche Gesellschaft für Psychologie e.V. (2016). *Berufsethische Richtlinien* [Guidelines for professional ethics]. Retrieved from https://www.dgps.de/fileadmin/documents/Empfehlungen/berufsethische_richtlinien_dgps.pdf.

Eichas, F., & Zölzer, U. (2018). Gray-box modeling of guitar amplifiers. *Journal of the Audio Engineering Society*, *66*, 1006–1115. http://doi.org/10.17743/jaes.2018.0052

Fastl, H., & Zwicker, E. (2010). *Psychoacoustics: Facts and models* (3rd ed.). Berlin; Heidelberg [u.a.]: Springer.

Fritz, C., Curtin, J., Poitevineau, J., & Tao, F.-C. (2017). Listener evaluations of new and old Italian violins. *Proceedings of the National Academy of Sciences*, *114*, 5395–5400. https://doi.org/10.1073/pnas.1619443114

Golubock, J. L., & Janata, P. (2013). Keeping timbre in mind: Working memory for complex sounds that can't be verbalized. *Journal of Experimental Psychology: Human Perception and Performance*, *32*, 399–412. http://doi.org/10.1037/a0029720

Greeves, D. (2012). *Kemper profiling amplifier*. Retrieved from https://www.soundonsound.com/reviews/kemper-profiling-amplifier

Hanover University of Music Drama and Media. (2017). *Leitlinien Guter Wissenschaftlicher Praxis* [Guidelines for good scientific practice]. Retrieved 14 February, 2019, from https://www.musikwissenschaft.hmtm-hannover.de/fileadmin/www.musikwissenschaft/Downloads/HMTMH_Regeln_guter_wissenschaftlicher_Praxis.pdf

Herbst, J.-P. (2019). Old sounds with new technologies? Examining the creative potential of guitar 'profiling' technology and the future of metal music from producers' perspectives. *Metal Music Studies*, *5*, 53–69. http://doi.org/10.1386/mms.5.1.53_1

Herbst, J.-P., Reuter, C., & Czedik-Eysenberg, I. (2018). Guitar profiling technology in metal music production: Public reception, capability, consequences and perspectives. *Metal Music Studies*, *4*, 481–506. http://doi.org/10.1386/mms.4.3.481_1

Huber, D. M., & Runstein, R. E. (2010). *Modern recording techniques* (7th ed.). Amsterdam: Elsevier.

Kemper, C. (2009). Germany Patent No. DE 10 2006 *0035* 188 B4. D. P.-u. Markenamt.

Kemper, C. (2015a). United States Patent No. US 8,796,530 B2. U. S. P. a. T. Office.

Kemper, C. (2015b). Europe Patent No. EP 1 883 064 B1. E. P. Office.

Kemper, C. (2015c). United States Design Patent No. US D738,349 S. U. S. P. a. T. Office.

Kemper, C. (2017). Germany Patent No. DE 10 2015 *110* 938 B4. D. P.-u. Markenamt.

Kemper, GmbH. (2019). *The kemper profiler*. Retrieved from https://www.kemper-amps.com/profiler/overview.

Kemper Profilier: Profiling guide. (2015). Retrieved from http://www.audioline.it/catalogo/allegati/2452_Kemper_Profiler_Rack.pdf.

Kopiez, R., Wolf, A., Platz, F., & Mons, J. (2016). Replacing the orchestra? – The discernibility of sample library and live orchestra sounds. *PloS One*, *11*, e0161911. http://doi.org/10.1371/journal.pone.0158324

Kruschke, J. K. (2015). *Doing Bayesian data analysis: A tutorial with R, JAGS, and Stan* (2nd ed.). Amsterdam: Academic Press.

Lartillot, O. (2019). *MIRtoolbox 1.7.2 user's manual*. Aalborg University.

Lartillot, O., Toiviainen, P., & Eerola, T. (2008). A Matlab toolbox for music information retrieval. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications, studies in classification, data analysis, and knowledge organization* (pp. 261–268). Berlin, Heidelberg: Springer-Verlag.

Levitin, D. J. (2014). Expert violinists can't tell old from new. *Proceedings of the National Academy of Sciences*, *111*, 7168–7169. https://doi.org/10.1073/pnas.1405851111

Logan, B. (2000). Mel frequency cepstral coefficients for music modeling. In *Proceedings of the international symposium on music information retrieval (ISMIR)*. Plymouth, USA, 23–25 October 2000. Retrieved from http://ismir2000.ismir.net/papers/logan_paper.pdf

Loughran, R., Walker, J., O'Neill, M., & O'Farrell, M. (2008). The use of mel-frequency cepstral coefficients in musical instrument identification. In The International Computer Music Association (Ed.), *International Computer Music Conference Proceedings*. Retrieved from https://quod.lib.umich.edu/i/icmc/bbp2372.2008.083/1

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide* (2nd ed.). Mahwah, N.J.: Lawrence Erlbaum.

Majewski, M., & Malecki, P. (2015). *Differences between recorded and emulated guitar sounds*. Paper presented at the Audio Engineering Society Convention 138. Retrieved from http://www.aes.org/e-lib/browse.cfm?elib=17621

McAdams, S., Winsberg, S., Donnadieu, S., De Soete, G., & Krimphoff, J. (1995). Perceptual scaling of synthesized musical timbres: Common dimensions, specificities, and latent subject classes. *Psychological Research*, *58*, 177–192. http://doi.org/10.1007/BF00419633

McNemar, Q. (1947). Note on the sampling error of the difference between correlated proportions or percentages. *Psychometrika*, *12*, 153–157. https://doi.org/10.1007/BF02295996

Miettinen, O. S. (1968). The matched pairs design in the case if all-or-none responses. *Biometrics*, *24*, 339–352.

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*, e89642. http://doi.org/10.1371/journal.pone.0089642

Mynett, M. (2017). *Metal music manual: Producing, engineering, mixing, and mastering contemporary heavy music*. New York: Routledge.

Osborne, L. (2018). *Remembering the Grateful Dead's 'Wall of Sound': An absurd feat of technological engineering*. Retrieved from https://enmoreaudio.com/remembering-the-grateful-deads-wall-of-sound-an-absurd-feat-of-technological-engineering/

Pakarinen, J., & Yeh, D. T. (2009). A review of digital techniques for modeling vacuum-tube guitar amplifiers. *Computer Music Journal*, *33*, 85–100.

Revelle, W., & Condon, D. M. (2018). Reliability. In P. Irwing, T. Booth, & D. J. Hughes (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (Vol. *2*, pp. 709–749). London: John Wiley & Sons.

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. *Musicae Scientiae*, *18*, 423–447. http://doi.org/10.1177/1029864914541851

Sharples, S. (2017). *Can people tell the difference between live and recorded music?* Retrieved from https://www.psneurope.com/live/can-people-tell-the-difference-between-live-and-recorded-music

Siedenburg, K., Fujinaga, I., & McAdams, S. (2016). A comparison of approaches to timbre descriptors in music information retrieval and music psychology. *Journal of New Music Research*, *45*, 27–41. http://doi.org/10.1080/09298215.2015.1132737

Sneap, A. (2012). *Kemper Profiling Amp – Artist talk with Andy Sneap*. Retrieved from https://www.youtube.com/watch?v=Nb1zI6pEu0A

Spicer, N. (2015). *Kemper Profiler RigCheck - Nathan Spicer on tour with Katy Perry*. Retrieved from https://www.youtube.com/watch?v=8rEBzzg5EfE&t=2 s

Théberge, P. (2001). 'Plugged in': Technology and popular music. In S. Frith, W. Straw, & J. Street (Eds.), *The Cambridge companion to pop and rock* (pp. 3–25). Cambridge University Press.

Thiesen, F. C., Kopiez, R., Reuter, C., & Czedik-Eysenberg, I. (2019). A snippet in a snippet: Development of the Matryoshka principle for the construction of very short musical stimuli (plinks). *Musicae Scientiae*. Advance online publication. doi:10.1177/1029864918820212

Vinnicombe, C. (2012). Kemper Profiling Amp review. Retrieved from https://www.musicradar.com/reviews/guitars/kemper-profiling-amp-538742

Wagenmakers, E.-J., Love, J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., . . . Morey, R. D. (2018a). Bayesian inference for psychology. Part II: Example applications with JASP. Psychonomic Bulletin & Review, *25*, 58–76. http://doi.org/10.3758/s13423-017-1323-7

Wagenmakers, E.-J., Marsman, M., Jamil, T., Ly, A., Verhagen, J., Love, J., . . . Morey, R. D. (2018b). Bayesian inference for psychology. Part I: Theoretical advantages and practical ramifications. Psychonomic Bulletin & Review, *25*, 35–57. http://doi.org/10.3758/s13423-017-1343-3

Weihe, P. (2012). Peter Weihe zur Praxis mit dem Kemper Profiling Amp [Peter Weihe about the practice with the Kemper Profiling Amp]. Retrieved from http://www.xound.com/specials/peter-weihe-zur-praxis-mit-dem-kemper-profiling-amp.html

Wickens, T. D. (2002). *Elementary signal detection theory*. New York: Oxford University Press.

Wycisk, Y., Kopiez, R., & Wolf, A. (2018). *Control of headphone and loudspeaker characteristics in online experiments*. Paper presented at the 15th International Conference on Music Perception and Cognition, Graz. Retrieved from https://static.uni-graz.at/fileadmin/veranstaltungen/music-psychology-conference2018/documents/ICMPC15ESCOM10abstractbook.pdf

You, J., & Jeon, J. Y. (2008). Just noticeable differences in sound quality metrics for refrigerator noise. *Noise Control Engineering Journal*, *56*, 414–424. https://doi.org/10.3397/1.2987734

Zollner, M. (2007). *Physics of the electric guitar: 10.1.2 Tube input-impedance* (T. Zwicker, Trans.). Retrieved from https://www.gitec-forum-eng.de/wp-content/uploads/2019/02/poteg-10-01-02-tube-input-impedance.pdf

Zollner, M. (2008-2014). *Physics of the electric guitar: 11.8 Loudspeaker enclosures* (T. Zwicker, Trans.). Retrieved from https://www.gitec-forum-eng.de/wp-content/uploads/2019/03/poteg-11-8-1-enclosures-and-boxes-basics.pdf

Zollner, M. (2011-2013). *Physics of the electric guitar: 10.10.8 Modeling Amps* (T. Zwicker, Trans.). Retrieved from https://www.gitec-forum-eng.de/wp-content/uploads/2019/02/poteg-10-10-8-modeling-amps.pdf