JASA ARTICLE

Assessing room acoustic listening expertise

Markus von Berg,[1,a)] Jochen Steffens,[1,b)] Stefan Weinzierl,[2] and Daniel Müllensiefen[3]
[1]*Institute of Sound and Vibration Engineering (ISAVE), Hochschule Düsseldorf, Münsterstaße 156, Düsseldorf, 40476, Germany*
[2]*Audio Communication Group, Technische Univeristät Berlin, Einsteinufer 17c, Berlin, 10587, Germany*
[3]*Goldsmiths, University of London, 8 Lewisham Way, New Cross, London, SE14 6NW, United Kingdom*

ABSTRACT:
Musicians and music professionals are often considered to be expert listeners for listening tests on room acoustics. However, these tests often target acoustic parameters other than those typically relevant in music such as pitch, rhythm, amplitude, or timbre. To assess the expertise in perceiving and understanding room acoustical phenomena, a listening test battery was constructed to measure the perceptual sensitivity and cognitive abilities in the identification of rooms with different reverberation times and different spectral envelopes. Performance in these tests was related to data from the Goldsmiths Musical Sophistication Index, self-reported previous experience in music recording and acoustics, and academic knowledge on acoustics. The data from 102 participants show that sensory and cognitive abilities are both correlated significantly with musical training, analytic listening skills, recording experience, and academic knowledge on acoustics, whereas general interest in and engagement with music do not show any significant correlations. The regression models, using only significantly correlated criteria of musicality and professional expertise, explain only small to moderate amounts (11%–28%) of the variance in the "room acoustic listening expertise" across the different tasks of the battery. Thus, the results suggest that the traditional criteria for selecting expert listeners in room acoustics are only weak predictors of their actual performances.
 https://doi.org/10.1121/10.0006574
(Received 14 May 2021; revised 20 August 2021; accepted 17 September 2021; published online 8 October 2021)
[Editor: Francesco Martellotta] Pages: 2539–2548

I. INTRODUCTION

Studies on room acoustic perception often describe test participants as "expert listeners." In some cases, a person's "listening expertise" is determined by musical training (Soulodre and Bradley, 1995) and/or musical listening experience (Barron, 1971), although the experiments often focus on acoustical features that are different from typical musical parameters such as pitch, rhythm, and amplitude of timbre. Other researchers refer to room acoustic knowledge as a pertinent degree of educational training (Wilkens, 1977) or professional activity (Weinzierl *et al.*, 2018). Sometimes the expertise is also empirically evaluated based on the consistency of the participants' responses (Lokki *et al.*, 2012). In many instances, the participants' perceptual expertise in room acoustics is just assumed without any further justification or empirical evidence (Barron, 1988).

However, it is largely unclear as to what constitutes the "acoustic expertise" of listeners in general, whether it is a *single* skill or a bundle of separate skills. Furthermore, it is yet unclear which indicators could be used to measure perceptual expertise in room acoustics empirically in a valid and reliable way.

The present study is an attempt to close this gap, gather empirical evidence on perceptual expertise in room acoustics, and investigate its relationship with musical expertise and sophistication. That is, we compare different expertise criteria as predictors for the participants' performance on tasks around room acoustics, aiming to distinguish between the potentially differing effects of individual expertise criteria with regard to the performance in listening tasks that focus more on sensory vs cognitive processing of auditory information.

A. Knowledge, experience, and perception

It is easy for most people to attribute certain aspects of a sound to the influence of the room, that is, to recognize the extent to which the loudness, timbre, and resonance of a speaker or musical instrument are shaped by the room rather than by the source itself. From a audio processing perspective, however, this can be regarded as a complex cognitive task because "hearing a room" requires a separation of the perceived sound into an original source signal and a spatial response of the room.

Because the perceived contribution of the room is part of the overall sensory impression, the ability to identify the properties of the room will certainly depend on the performance of the auditory system such as the ability to reliably discriminate small differences in spectral and temporal aspects of sound. At the same time, it appears reasonable to

[a)]Also at: Audio Communication Group, Technische Univeristät Berlin, Einsteinufer 17c, Berlin, 10587, Germany. Electronic mail: markusmartin.vonberg@hs-duesseldorf.de
[b)]Also at: Audio Communication Group, Technische Univeristät Berlin, Einsteinufer 17c, Berlin, 10587, Germany.

assume that assigning these differences to the room acoustical properties is easier for listeners who are familiar with the principles of sound propagation and have sufficient experience regarding the influence of different spaces on the sound of music. Finally, the accuracy and consistency of the assessments of the specific room acoustic attributes is likely to depend also on the understanding of the terminology used in the listening experiments. The terminology and corresponding concepts would often be acquired through both professional experience and/or specific training in room acoustics that participants have received prior to taking the listening tests.

Previous research has demonstrated some effects of professional expertise on the results of the listening experiments. For instance, in a study by Kreimann *et al.* (1990), clinical experts used different, individual strategies to identify dysphonia (a larynx malfunction that causes hoarse voice) in the vocal recordings compared to a sample of naive listeners who rather consistently relied on the voices' fundamental frequency. Also, a study by Kuusinen and Lokki (2020) observed that listening test participants benefited from acoustic expertise and experience in a discrimination task in which different room impulse responses were presented with the same source signal but not if the source signal also changed between the different stimuli.

Considering musical expertise, Wong *et al.* (2007) and Di Liberto *et al.* (2020) measured an improved tracking of linguistic sound features, such as pitch or amplitude envelope, in the neural responses of musicians with several years of formal musical training. However, regular concert attendance and the use of high-fidelity audio systems did not predict the results of the sensory tests for selecting assessors to evaluate the stereo width of an audio system (Wickelmaier and Choisel, 2005). Similarly, no advantage of professional expertise could be found for recognizing a concert hall's shape by its sound (Greif *et al.*, 2020).

These findings do not only show that expertise and experience affect certain aspects of auditory perception depending on the specific task. They also indicate the inconsistent understanding of "expertise" and divergence in opinions on how expertise should be measured (Zacharov and Lorho, 2006). For instance, Di Liberto *et al.* (2020) selected test persons by the years of formal training on a musical instrument. Wickelmaier and Choisel (2005), on the other hand, only asked whether participants played an instrument at all and if they possessed high-fidelity audio equipment or regularly visited concerts.

In the present study, the *Goldsmiths Musical Sophistication Index* (Gold-MSI) was used to assess musical expertise. It defines "musical sophistication" as a multidimensional construct consisting of a general sophistication factor and five subscales, which cover formal training as well as everyday listening habits and self-reported perceptual abilities in the form of empirically validated self-report questionnaires (Müllensiefen *et al.*, 2014). The Gold-MSI provides a quick and yet comprehensive measure of the musical expertise taking into account more than just formal training on a musical instrument. The acoustical expertise was measured with technical questions as well as questions about academic education and practical experience in related fields.

B. Room acoustical qualities

In studies of the perception of room acoustic properties, three aspects have consistently been shown to be constitutive in determining the acoustic impact of rooms, such as musical performance spaces—see Kuusinen and Lokki (2017) or Weinzierl and Vorländer (2015) for an overview. These are

- *strength*, i.e., the degree of perceived amplification of a sound source through the sum of all room reflections;
- *reverberance*, i.e., the intensity and duration of perceived reverberation; and
- *timbre*, i.e., the perceived sound color of the reverberant sound field.

These aspects, together with a higher-level factor describing the individual fit of room acoustics and audio content, were also found to be the primary dimensions of the room acoustic quality inventory, which was established in the course of a recent study involving a large number of rooms and subjects (Weinzierl *et al.*, 2018).

Therefore, in the present study, the sensory performance of different listeners was tested by measuring their ability to discriminate small variations of the absolute values and spectral slope of the frequency-dependent reverberation time ($T30$) as indicators of the perceived reverberance and sound color of the room. The abilities in higher cognitive analysis of room acoustic perception were assessed by an identification test in which the room impulse responses excited by different source signals were to be recognized as the same or different. The rationale of these two tests, combined with the assessment of knowledge and experience, follows the classification of individuals and their assessment abilities for the taste of food according to ISO 8586 (2014), which distinguishes between expert knowledge, immediate sensory assessment, and the ability to identify chemical or biological origins of these impressions—a classification Zacharov and Lorho (2006) also proposed for participants in listening tests.

C. Ability estimation and item response theory

The rating of stimuli on a scale, that is, to assign different numerical values to different sensory objects, is a result of the abilities of the rater and difficulty of the items being rated. Although these influences are always confounded in classical test theory (Hambleton and van der Linden, 1982), item response theory (IRT) provides a framework to determine them separately (de Ayala, 2013).

In IRT models, the response of a person to an item is modeled by a parameter representing the participant's ability and up to four parameters related to the properties of the item. These include the *difficulty* (representing the

https://doi.org/10.1121/10.0006574

"location" on the difficulty range), the *discrimination* (representing how strongly the ratings vary with a person's ability), and two additional parameters representing the guessing rate and degree of inattention during the rating process (Harrison and Müllensiefen, 2018).

To obtain an estimate for the individual ability of the participants that is independent of the specific items used and allows for efficient adaptive testing methods (Hambleton and van der Linden, 1982), the listening tests in this study were designed following the IRT framework. Especially in the test on cognitive processing of room acoustical information, the variation in the item difficulty was a key element in the design of the test items. The principle idea was that room responses with larger differences regarding a certain acoustical parameter are easier to distinguish and, thus, reduce the item difficulty. The actual item difficulty was then estimated from the test results, which is usually called item calibration (Baker, 2001).

II. METHODS

A. Participants

102 participants took part in the experiment (71 male, 30 female, 1 diverse). The age ranged from 16 to 59 years old with an average of 25.7 years old [standard deviation (SD = 7.9)]. About half of the participants were students or graduates from degrees related to acoustics or audio technology. The average test duration was 32 min (SD = 8).

B. Test design

1. Perceptual sensitivity test

The first test aimed to assess the perceptual sensitivity for different room acoustical qualities. Participants were presented with 21 numbered buttons ordered on a horizontal scale and triggering sounds with room acoustical features, systematically modified from left to right. Another button triggered a reference sound, and the test persons were asked to identify the sound on the scale that matches this reference. The distance from the scale step that the participants chose to the actual reference position was recorded as a measure of the minimal stimulus difference that each participant was able to perceive. Due to the number of response options, this procedure has a low guessing probability of 4.8%—based on the assumption that each button on the scale is equally likely to be selected. To avert participants anticipating that the reference would always be located in the mid range of the scale, the reference sound and minimum and maximum reverberation times of the stimuli were different in each item.

Two versions of this test were implemented with a variation of the reverberation time and its frequency-dependent distribution, intended to induce changes in the perceived reverberance and sound color (see Fig. 1). These variations were generated by binaural impulse responses simulated with the RAVEN software (Schröder and Vorländer, 2011). For both test parameters, a room with a volume of 1200 m^3 was modeled, and only the amount of frequency-dependent sound absorption of the walls, floor, and ceiling was changed to modify the reverberation characteristics.

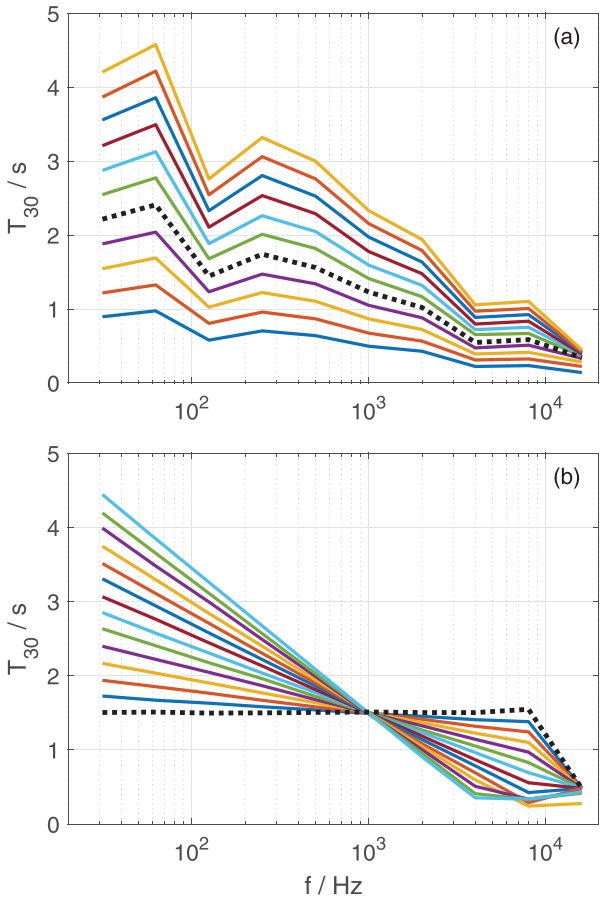

FIG. 1. (Color online) The variation of the reverberation time (a) and spectral envelope (b). The dotted line represents the initial room's reverberation time. For a better overview, only every third generated reverberation curve is plotted. (a) and (b) show the actual simulation results and, thus, a few irregularities can be seen in (b).

For reverberation time, an average of 1.5 s was chosen as recommended by the German standard DIN 18041 (2004) for rooms of this size used for musical performances. To define the spectral envelope for this room's response, the reverberation time in octave band resolution was randomly scattered around this mean within frequency-dependent tolerances (70%–20%), which are also defined in the standard. To create sets of gradually changing room responses, the reverberation time in each octave band was altered in 5% steps of the initial reverberation time from 40% of this reverberation time to 250%. For better visibility, every third generated reverberation curve is displayed in Fig. 1(a). The 5% step size matches the just noticeable difference of reverberance according to ISO 3382-1 (2009).

For the second version of the perceptual sensitivity test, a linear variation of the reverberation's frequency shape was realized by "tilting" the reverberation time, that is, by increasing it toward lower frequencies and decreasing it toward higher frequencies, as depicted in Fig. 1(b), again, for every third room response. In the 1 kHz band, the

reverberation time was always kept at 1.5 s; at 16 kHz, the value was always at approximately 0.5 s, assuming that the reverberation time in this octave band is mainly determined by the air absorption and attempting to avoid generating ecologically nonvalid auralisations.

All room responses were convolved with the same source signal (i.e., an anechoic cello recording), and five items were created from these two sets of room responses. At the beginning of the perceptual sensitivity test, an example scale with instructions and no reference was presented to the subjects to avoid misunderstandings about the test procedure. As an orientation as to which acoustical features to consider, the scales were labeled on the left and right sides with "less/more reverberation" in the reverberation time test and "bright/dark" in the spectral envelope test.

The ranges of reverberation times and spectral envelopes covered by the scales used in the perceptual sensitivity test were configured such that the difference between the reference and reverberation time or spectral envelope on the left and right end of the scale was never less than 20%. This way, the location of the reference on the scale could always be narrowed down from its left and right side. To randomize the location of the references across subjects, two scales covering different ranges of reverberation times and spectral envelopes were defined for each item, and one was randomly selected.

2. Room recognition test

The second test focused on the cognitive processing of the room acoustical stimuli. It was designed to show whether a listener is able to conceptualize and recognize a room as an auditory object with a distinct acoustic influence. Three sounds were presented in a two-alternative forced-choice procedure (ABX test). All three were created with different excerpts of an anechoic piano recording as source signals, and two were convolved with different room responses. One of the two room responses was also used for the third stimulus, and subjects were asked to identify which of the first two stimuli had the same room response as the third.

For an immediate comparison of the cognitive abilities required for this task with the perceptual sensitivity assessed in the previous test, there were again two test versions in which either the reverberation time or its spectral envelope was modified, and the stimuli were created with the same room impulse responses as in the test for perceptual sensitivity. Both test versions comprised 20 items, i.e., test repetitions with different stimulus sets. One room response was used across all items and had to be compared to a room response with a different reverberation time or spectral envelope in every item. Thus, quantitative differences of the reverberation time and spectral envelope of the two room responses presented in each item continuously changed, which was intended to affect the cognitive and perceptual item difficulty. The smallest difference in the reverberation time between the two rooms was only 10%. In the iteration with the largest differences, one reverberation time was three times longer than the other. Because Kuusinen and Lokki (2020) already showed that comparing room responses with different source signals is considerably more difficult than with the same source signal, a minimal difference between the room responses, which was double as the threshold used in the perceptual sensitivity test, was considered sufficient.

To limit the test duration and ensure equal conditions for all participants, stimuli could only be played back once. Participants were introduced to the test procedure with an example using large and clearly audible reverberation time differences between the two rooms. In this introductory example, participants were allowed to listen to the stimuli as often as they liked.

In both tests (i.e., the perceptual sensitivity and room recognition test), the item order was randomized across the participants. In addition, in the room recognition test, it was randomized whether the room response with the larger or lesser reverberation time or more or less reverberation at low frequencies, respectively, was used for two of the three stimuli in each item.

3. Assessment of knowledge, experience, and musicality

A multiple-choice test on prior knowledge of room acoustics was included, comprising eight questions with four response options and only one correct answer. The questions ranged from simple topics ("How does the opening of a window affect reverberation?," "Which physical parameters influence the speed of sound?") to more difficult topics ("Where are the sound pressure maxima of standing waves in a rectangular room with perfectly reflecting surfaces?").

Furthermore, in a self-report inventory on relevant practical experience,[1] participants were asked to rate their previous experience with room acoustics in their academic education or in the context of musical and recording activities on a five-point scale from "never" to "very often." To capture the participant's formal musical training, as well as a more general interest in music, musical sophistication was assessed through three subscales of the Gold-MSI by Müllensiefen *et al.* (2014)—"perceptual abilities," "active engagement," and "musical training." The factor, perceptual abilities, comprises nine questions about analytic listening, active engagement covers the role of music in everyday life with nine items, and musical training assesses formal musical training with seven items. The remaining factors, "singing abilities" and "emotion," were considered irrelevant in the context of the present study.

C. Stimulus generation

Binaural room impulse responses were simulated with the RAVEN software (Schröder and Vorländer, 2011) with the room models created in Trimble SketchUp and the simulation series configured in MATLAB (The MathWorks, Natick, MA). The MATLAB plugin offers an iterative modification of

https://doi.org/10.1121/10.0006574

absorption and scattering properties of a modeled room to approximate specified reverberation times in each octave band between 32 Hz and 16 kHz. This way, the reverberation time curves in Fig. 1 could be realized. Simulations were calculated with 200 000 particles and image sources up to second order.

The room used in both the perceptual sensitivity and room recognition tests was designed to resemble a concert venue and had a rectangular shape (20.75 m length, 10.00 m width, and 6.00 m height) with a stage plateau at one side (4.50 m in depth and 1.00 m in height). The source was positioned at the center of the stage at a height of 1 m above the stage floor. Directed toward the source, the receiver was positioned at a distance of 10 m from the source with an equal distance to both side walls of the room (5 m) and at a height of 1.50 m. Anechoic recordings of musical instruments were chosen as source signals to present the room responses combined with sources whose sound properties were expected to be generally familiar to the participants. In the perceptual sensitivity test, a short excerpt of an anechoic cello recording by Hansen and Munch (1991) was used. For the room recognition test, three excerpts of an anechoic recording of a piano version of the overture of George Gershwin's "Girl Crazy" were chosen, recorded at Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University as part of the sound database included in the RAVEN software. The same instrument was selected for all of the stimuli of the room recognition test because the task of recognizing the timbre of a room's reverberation time was considered to be too difficult when used with different sources, which themselves might strongly differ in their spectra.

D. Experimental setup

The experiment was administered both in German and English and conducted at two universities in Germany (Technische Universität Berlin and Hochschule Düsseldorf–University of Applied Sciences), one in the United Kingdom (Goldsmiths, University of London), and a high school in Germany (Albert-Einstein-Gymnasium Kaarst). At each university, participants were tested individually in a soundproof listening environment. At the high school, a computer laboratory was prepared so that nine participants could take the test simultaneously and the experiment would not exceed regular class periods.

The binaural stimuli were presented via headphones for which the headphone transfer functions were compensated using the equalization filters of the FABIAN database (Brinkmann *et al.*, 2017a; Brinkmann *et al.*, 2017b). For better control of the playback volume, external audio interfaces were used. Because of the amount of required playback systems, several models of headphones (Sennheiser HD650, Sennheiser Electronic, Wedemark, Germany; STAX SR-303, STAX, Fujimi, Japan with HEAD Acoustics HPS IV preamplifier, HEAD Acoustics, Herzogenrath, Germany; AKG K702, Harmann International Industries, Stamford, CT; Beyerdynamic DT770 Pro, Beyerdynamic, Heilbronn, Germany) and audio interfaces (Focusrite Scarlett 6i6, Focusrite, High Wycombe, UK; HEAD Acoustics PEQ V, HEAD Acoustics, Herzogenrath, Germany; Behringer U-Control UCA-222, MUSIC Group IP, Road Town, British Virgin Islands; Steinberg UR-22, Steinberg Media Technologies, Hamburg, Germany) were used at the different institutions. In addition to having a bespoke equalization filter available in the FABIAN database, one important criterion for selecting the type of headphones was a circum-aural design, making the frequency responses less prone to irregularities as a result of variable positioning (Riederer, 1998).

The playback volume was previously adjusted by the experimenters, aiming at a volume that would allow careful examination of the presented content without becoming annoying or stressful across the duration of the experiment. The loudness changed between the stimuli and ranged from approximately 7 to 25 sone. Participants were not allowed to change the volume. The entire test was implemented using the survey web application LimeSurvey (LimeSurvey GmbH, 2003); therefore, no additional software was needed on the computers used in the experiment.

E. Test procedure

We applied a between-subjects design by providing two versions of the test battery. This was done because repeating both listening tests with each of the two different acoustical parameters was considered to be overly time-consuming and exhausting for participants. In the first test version, both the perceptual sensitivity and room recognition test were presented with a changing reverberation time, whereas in the second version, both tests were presented with a changing spectral envelope. To ensure that an equal number of school children and graduates or students of acoustic-related programs took each test, at each institution, the participants were alternately assigned to the first and second test versions. However, because of technical issues affecting the server hosting the online survey, eventually, 54 participants completed the test with varied reverberation times, whereas 48 people were presented with the version with a changing spectral envelope.

The listening tests and questions on knowledge, experience, and musicality were presented in an order[1] that aimed at preventing fatigue and concentration problems by alternating cognitively demanding tasks, such as the listening tests or questions on academic knowledge, with the less challenging self-report questionnaires, such as the Gold-MSI.

F. Statistical analysis

The statistical analyses of the experimental data[1] were implemented in *R* (The *R* Foundation, 1993) using the package psych (Revelle, 2019) for correlation and factor analysis. The calculation of explanatory logistic IRT models was conducted using the packages lme4 (Bates *et al.*, 2015) and 'psyphy' (Knoblauch, 2014). The package catR (Magis and Raîche, 2011) was used for the estimation of the participant abilities and the selection of the prediction models employed functions implemented in the package StepReg (Li *et al.*, 2020). The significance level was set to 0.05 in all of the analyses.

G. Ethics

The study received ethical approval from the ethics committee at Goldsmiths, University of London. As this approval only covered the adult test takers, the parents of the school children were informed about the experiment a few days in advance and given the opportunity to decline participation in the study. In class, the students were, again, explicitly asked whether they wanted to participate or not.

III. RESULTS

A. Gold-MSI and previous expertise

The raw, unweighted scores of each Gold-MSI factor were transformed into percentiles derived from a large sample of 147 633 participants by Müllensiefen *et al.* (2014). A histogram of the percentiles is shown in Fig. 2(a). The mean values of these percentiles were 40.9% for active engagement (SD = 27.2%), 40.4% for perceptual abilities (SD = 28.3%) and 45.3% for musical training (SD = 25.6%).

Figure 2(b) displays the answers to the self-report on previous experience. For each category (academic education, musical performance, and recording activities), about 40%–50% of the subjects reported that they never had contact with room acoustics (52.0% for academic, 39.2% for musical, and 44.1% for recording-related situations). Only 2.0% of the participants stated that they dealt very often with room acoustics as part of a musical activity (2.9% for academic and 4.9% for recording contexts).

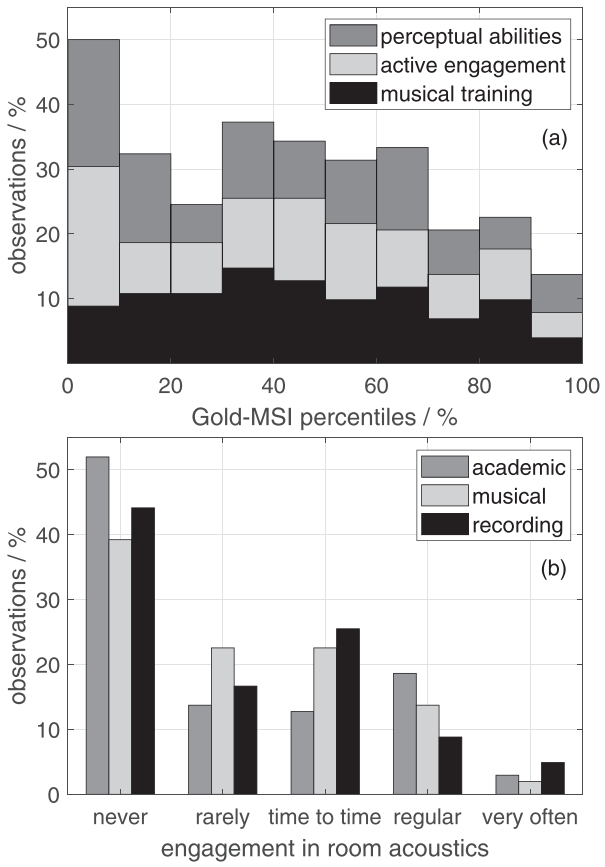

FIG. 2. The distribution of the musical and acoustic experiences within the sample of 102 participants. (a) The frequencies of the percentiles for each factor of the Gold-MSI and (b) the distribution of the degree of previous experience with room acoustics based on self-reports are shown.

B. Ability estimates

1. Perceptual sensitivity test

The data from the perceptual sensitivity task were subjected to a factor analysis to extract a single factor representing participant ability. Therefore, the responses were transformed into absolute distances from the reference position to ensure that large deviations always corresponded to large response values while the minimal value of zero represented the selection of the correct scale position. The mean distance from the reference position across all items and participants was 1.65 (SD = 1.88) in the reverberation time group and 2.93 (SD = 2.95) in the spectral envelope group. Internal consistency, in terms of inter-item correlations, was assessed by means of Cronbach's α (=0.56 for the reverberation time and 0.70 for the spectral envelope group) and McDonald's ω, which was at 0.59 for reverberation time and 0.71 for the spectral envelope. McDonald's ω is less widely used compared to Cronbach's α but relies on more realistic assumptions and has repeatedly been recommended as a more robust reliability estimate of the internal consistency (Dunn *et al.*, 2014). As input data for the factor analysis, polychoric correlation matrices (Olsson, 1979) of the responses were calculated. Originally designed for comparison of the ordinal data, polychoric correlation assumes that a latent trait's metric is not properly captured if the measured ordinal ranks are treated as equidistant numerical values. Although the metric of this test is based on a perceptual threshold for reverberance suggested by ISO 3382–1 (2009), the stimulus differences were always 5% of the reference reverberation time at 1.5 s, as described in Sec. II B 1. Thus, the relative increase in reverberation time between two adjacent stimuli changed along the scales, which probably also affected the perceived change of reverberance.

The Kaiser-Meyer-Olkin (KMO) criterion was examined to test the correlation matrices' suitability for factor analysis. Both matrices indicated a good fit with a KMO of 0.68 and a minimum item-specific mean sampling accuracy (MSA) of 0.64 in the reverberation time group and a KMO of 0.71 and a minimum MSA of 0.68 in the spectral envelope group. Factor analyses were conducted for both test groups. In both the reverberation time and spectral envelope group, one item exhibited a factor loading significantly below the frequently applied threshold of 0.40 (Howard, 2016) and was, thus, excluded and the factor analysis repeated. The extraction of a single factor was confirmed by the Kaiser criterion and scree test. Factor scores for all participants were calculated and, subsequently, used as person ability estimates on the perceptual sensitivity task.

2. Room recognition tests

In the room recognition test, the number of correct answers for both test parameters showed a rather consistent

https://doi.org/10.1121/10.0006574

decrease toward smaller stimulus differences, ranging from maximum success rates of about 90% to values around the guessing probability of 50% [Fig. 3(a)].

The logistic link functions for the IRT models were estimated by fitting a generalized linear (GLM) model. The lower asymptote of this logistic function accounts for the guessing probability, and the upper asymptote accounts for the test-specific possibility of momentary inattention. For the results of the reverberation time test, a lower asymptote of 0.49 confirmed that the theoretical guessing probability of a two-alternative forced-choice test of 0.50 and an upper asymptote of 0.88 were estimated. This value implies that for 2 of the 20 items, subjects would give wrong answers resulting from inattention, which seems to be a reasonable assumption.

For the data of the spectral envelope test, the GLM model fit did not converge. This might be because (unlike the results of the reverberation time test) the rate of correct answers is not skewed for items with larger stimulus differences (see Fig. 3), indicating the absence of a sufficient number of easy items. Thus, the asymptotes estimated for the reverberation time group were also used for the IRT model of the spectral envelope test because the test-paradigm-specific properties modeled by these asymptotes were expected to be independent of the acoustic parameters changed in the stimuli.

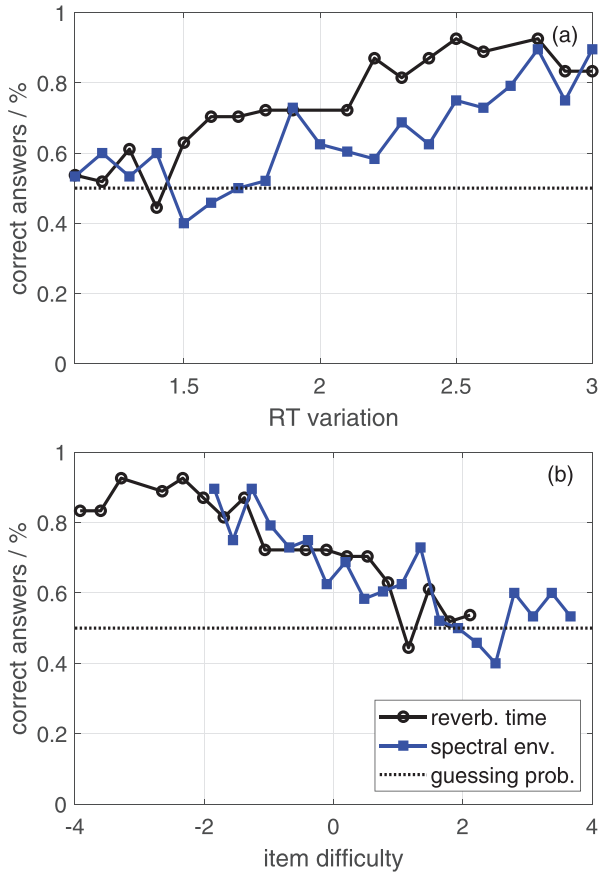

FIG. 3. (Color online) (a) The success rates for each item. For the reverberation time results, the x axis refers to the ratio of the frequency-independent reverberation time of the two rooms compared on each item, and for the spectral envelope test, it refers to the ratio of the reverberation time at 32, representing the change in the spectral envelope as described in Sec. II C. (b) The success rates plotted over the item difficulty, estimated from the GLM models following the IRT.

The mixed-effects models were estimated with the ratio of the reverberation time or spectral envelope between the two rooms compared on each item as a fixed effect, the subjects' individual contribution as a random effect, and the previously customized link function. The variance explained by the models was assessed by means of the marginal and conditional coefficients of determination, R_m^2 and R_c^2, for the generalized mixed-effects models (Nakagawa and Schielzeth, 2013). In the reverberation time group, the fixed effects accounted for 53% of the observed variance ($R_m^2 = 0.53$) and both fixed and random effects for 68% ($R_c^2 = 0.68$). In the spectral envelope group, R_m^2 and R_c^2 reached values of 0.46 and 0.64, respectively.

Following the work of de Boeck *et al.* (2011), the IRT model parameters were derived from the mixed-effects models. A constant item discrimination for all items was extracted as the SD of the participant random intercept, and the item difficulty was determined by dividing the sum of the fixed effects by the negative item discrimination. Finally, the individual participant ability could be estimated from the four item parameters and responses using the weighted likelihood estimation. In the reverberation time test, item discrimination was at 1.25, and in the spectral envelope test, an item discrimination of 1.29 was calculated. According to Baker (2001), item discrimination values above one indicate an acceptable level of test information (i.e., precision of ability estimation). Figure 3(b) shows the success rates plotted over the item difficulties for both tests.

C. Academic knowledge on room acoustics

The multiple-choice questions obtaining academic knowledge on room acoustics could be ordered by an almost linear decrease in correct answers from 83% to 33%, covering a wide range above the guessing probability of 25%. Based on the number of correct answers, the questions were assigned a pseudo-rank and a GLM model was fitted similarly to those calculated for the room recognition tests ($R_m^2 = 0.17$, $R_c^2 = 0.60$). This allowed us to account for the guessing probability and apparent differences in the question difficulty as item parameters. A latent ability was estimated, representing an aggregate measure of an individual's degree of knowledge about room acoustics. A correlation of $r = 0.67$ ($p < 0.001$) with self-reported academic involvement also supports the validity of this measure.

D. Prediction effects of musicality and expertise

In the next step, the Gold-MSI factors, previous experience, and degree of knowledge on room acousitcs were employed as predictor variables for the performance. This was performed separately for each listening test. The correlations between these variables are shown in Table I.

TABLE I. The correlations (Pearson's r) between the predictors and ability estimates. The significant correlations are marked with "*" for $p < 0.05$ and "**" for $p < 0.01$. No correction was applied to the significance values. No correlations were calculated between the ability estimates for the different room acoustical parameters because these were measured with different participants.

	Reverberation time		Spectral envelope								
	Sensitivity	Recognition	Sensitivity	Recognition	KN	PA	AE	MT	EA	EM	ER
Sensitivity (rev. time)	1.00		—	—							
Recognition (reverberation time)	0.19	1.00	—	—							
Sensitivity (spectral envelope)	—	—	1.00								
Recognition (spectral envelope)	—	—	0.26	1.00							
Knowledge (KN)	0.45**	0.41**	0.38**	−0.05	1.00						
Perceptual abilities (PA)	0.33*	0.34*	0.36**	0.35*	0.29**	1.00					
Active engagement (AE)	0.12	0.01	0.03	0.29*	0.15	0.45**	1.00				
Musical training (MT)	0.35**	0.46**	0.07	0.24	0.34**	0.59**	0.38**	1.00			
Experience: Academic (EA)	0.41**	0.40**	0.35*	0.24	0.67**	0.44**	0.27*	0.44**	1.00		
Experience: Musical (EM)	0.32*	0.38**	0.12	0.26	0.41**	0.47**	0.52**	0.58**	0.52**	1.00	
Experience: Recording (ER)	0.35**	0.46**	0.01	0.29*	0.34**	0.45**	0.52**	0.47*	0.32**	0.68**	1.00

The prediction models for each estimated ability were selected using a forward stepwise selection procedure. The algorithm starts with fitting a regression model with only one predictor variable and, subsequently, adds other predictors to the model to minimize the *Akaike information criterion* (AIC; James *et al.*, 2017). In this case, the sample-size corrected Akaike information criterion (AICc) was used, and each variable was tested as the first, single predictor. Only predictor variables that correlated with the given ability estimate at a significance level below 0.10 were tested as possible predictors.

The ability that was estimated from the perceptual sensitivity test with the changing reverberation time significantly correlated with all predictor variables except for active engagement. The stepwise selection procedure returned only the degree of knowledge on room acoustics and musical training as predictors. Only the prediction effect of knowledge on room acoustics was significant ($p < 0.05$; the prediction effect of musical training was at $p < 0.10$) at an adjusted R^2 of 0.19. In the perceptual sensitivity test with a changing spectral envelope, the correlations of the ability estimates with the degree of knowledge on room acoustics and perceptual abilities were observed. Both also turned out to be significant predictors ($p < 0.05$ for both effects), reaching an adjusted R^2 of 0.21.

The ability estimated from the room recognition test with a changing reverberation time also correlated with each predictor variable except for active engagement. The forward stepwise selection procedure returned the recording-related experience, musical training, and academic experience as predictors. However, only recording-related experience turned out to be statistically significant ($p < 0.05$ for the effect of recording-related experience, $p < 0.10$ for the effect of musical training, and $p > 0.10$ for the effect of academic experience). The adjusted R^2 was at 0.28.

In the room recognition test of the spectral envelope group, the only significant correlations were observed for perceptual abilities, active engagement, and the recording-related experience. The perceptual abilities were also selected as a predictor variable and their effect was statistically significant ($p < 0.05$) at a rather small effect size of an (adjusted) R^2 of 0.11.

IV. DISCUSSION AND CONCLUSION

A listening test battery was constructed to assess the listening expertise in the perception of room acoustical phenomena. Two test paradigms were aimed at measuring different abilities: the perceptual sensitivity for variations in the reverberation time and its spectral envelope, as well as the cognitive ability to identify rooms excited by different source signals, with the latter requiring a mental representation of a room's distinct acoustical properties beyond the immediate auditory impression.

The performance on these tests was statistically related to three factors of the Gold-MSI, covering abilities in analytic music listening, musical training, and the importance of music in everyday life, as well as academic knowledge in room acoustics and self-reported experience in room acoustics, music, and music recording of the 102 participants. The distribution of the Gold-MSI factors and self-reports indicates that a diverse sample of laymen, amateurs, semi-professionals, and professionals in music and room acoustics participated in the study. However, neither of these distributions showed clear bimodal characteristics that would support a valid *a priori* categorization of the participants into musical or acoustical experts and nonexperts (see Fig. 2).

Interestingly, perceptual and cognitive ability estimates derived from the two listening tests show very low and insignificant correlations among each other. This finding indicates that the two test paradigms indeed require fundamentally different abilities, confirming the previous findings suggesting that acoustical listening expertise is not a single skill but consists of several abilities (Kuusinen and Lokki, 2020).

In the tests dealing with the detection different reverberation times, performance in both the sensitivity and recognition tasks was correlated with musical sophistication. However, the correlations only were significant for formal training in playing a musical instrument and analytic listening but not for the general interest in and engagement with music. This is in line with findings reported by Wickelmaier and Choisel (2005). When playing an instrument, musicians actively explore how their own sound is shaped by the characteristics of different spaces. This seems to increase both the sensitivity to such changes and the understanding of how room acoustics affect sound. A similar repertoire of experience is obviously built up in music recording, from which participants seemed to profit in the room recognition task and music or acoustic practice, which correlated with both the sensitivity and recognition performances.

The results further suggest that participants with a theoretical knowledge of room acoustics had an advantage both in the perceptual sensitivity and room recognition tests. At first glance, one would not expect perceptual sensitivity to be influenced by knowledge. Our findings, however, indicate that "room reverberation" as a percept is already strongly tied to concepts, that is, to abstract classes of objects formed by both our prior experience and prior knowledge (Murphy, 2004), and our sensory instruments use this prior knowledge to focus on specific cues expected to be relevant to the invoked concept.

In the tests dealing with the detection of different spectral envelopes of the reverberation time, there were generally lower correlations between the ability estimates and descriptors of musicality and professional and academic expertise. This difference is reflected both in the bivariate correlations and the explained variance of the stepwise regression models. Although participants' performance in detecting spectral differences with the same source signal could be predicted by a combination of knowledge on room acoustics and perceptual abilities, the recognition of rooms based on their spectral envelope for different source signals could only be very weakly predicted by the perceptual abilities. Interestingly, experiential knowledge, as provided by musical practice, does not seem to be helpful in discriminating the timbral effects of the source and room.

Selecting the best predictors of all of the measures on music sophistication, music and acoustic experience, and academic room acoustic knowledge in the form of multiple regression models, 11%–28% of the variance in the performance of the four listening tests can be statistically explained.

Thus, to obtain a reliable prediction of "room acoustic listening expertise," a more comprehensive model, including a specific listening test battery that incorporates multiple tasks, must be developed through further research. As these tasks will also contain different aspects that have so far only been collected by self-reports, an increased validity of the predictors can be expected. This might also allow a uniform, evidence-based definition of categories, such as expert or naive listeners, to quickly estimate the listening expertise of a test sample. As long as such a pretest does not exist, however, expertise in the perceptual assessment of room acoustic conditions can only be achieved by familiarization and specific training of all participants prior to the listening test, whereas self-reports and socio-demographic information represent rather unreliable indicators.

[1]See supplementary material at https://www.scitation.org/doi/suppl/10.1121/10.0006574 for the multiple-choice test on prior knowledge and questions for previous experience, a detailed overview of the experiment's structure, and the statistical analysis of the listening tests and regression models in *R*.

Baker, F. B. (**2001**). *The Basics of Item Response Theory*, 2nd ed. (ERIC Clearinghouse on Assessment and Evaluation, College Park, MD).
Barron, M. (**1971**). "The subjective effects of first reflections in concert halls—The need for lateral reflections," J. Sound Vib. **15**(4), 475–494.
Barron, M. (**1988**). "Subjective study of British symphony concert halls," Acta Acust. Acust. **66**(1), 1–14.
Bates, D., Mälcher, M., Bolker, B., and Walker, S. (**2015**). "Fitting linear mixed-effects models using lme4," J. Stat. Softw. **67**(1), 1–48.
Brinkmann, F., Lindau, A., Weinzierl, S., Geissler, G., van de Par, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (**2017a**). "The FABIAN head-related transfer-function data base," available at http://dx.doi.org/10.14279/depositonce-5718.5 (Last viewed September 28, 2021).
Brinkmann, F., Lindau, A., Weinzierl, S., van de Par, S., Müller-Trapet, M., Opdam, R., and Vorländer, M. (**2017b**). "A high resolution and full-spherical head-related transfer function database for different head-above-torso orientations," J. Audio Eng. Soc. **65**(10), 841–848.
de Ayala, R. J. (**2013**). *The Theory and Practice of Item Response Theory* (Guilford Publications, London).
de Boeck, P., Bakker, M., Zwitser, R., Nivard, M., Hofman, A., Tuerlinckx, F., and Partchev, I. (**2011**). "The estimation of item response models with the lmer function from the lme4 package in *R*," J. Stat. Softw. **39**(12), 1–28.
Di Liberto, G. M., Pelofi, C., Shamma, S., and de Cheveigné, A. (**2020**). "Musical expertise enhances the cortical tracking of the acoustic envelope during naturalistic music listening," Acoust. Sci. Technol. **41**(1), 361–364.
DIN 18041. (**2004**). *Hörsamkeit in Räumen—Anforderungen, Empfehlungen Und Hinweise für die Planung (Acoustic Quality in Rooms—Specifications and Instructions for the Room Acoustic Design)* (Deutsches Institut für Normung e.V., Berlin, Germany).
Dunn, T. J., Baguley, T., and Brunsden, V. (**2014**). "From alpha to omega: A practical solution to the pervasive problem of internal consistency estimation," Br. J. Psychol. **105**(3), 399–412.
Greif, J., Ackermann, D., Kokabi, O., and Weinzierl, S. (**2020**). "Kann man die Form eines Konzertsaales hören?" ("Can the geometrical shape of a concert hall be heard?"), in *Fortschritte der Akustik—DAGA 2020* (Deutsche Gesellschaft für Akustik, Berlin), pp. 1153–1156.
Hambleton, R. K., and van der Linden, W. J. (**1982**). "Advances in item response theory and applications: An introduction," Appl. Psychol. Meas. **6**(4), 373–3378.
Hansen, V., and Munch, G. (**1991**). "Making recordings for simulation tests in the archimedes project," J. Audio Eng. Soc. **39**(10), 768–774.
Harrison, P. M. C., and Müllensiefen, D. (**2018**). "Development and validation of the Computerised Adaptive Beat Alignment Test (CA-BAT)," Sci. Rep. **8**, 12395.
Howard, M. C. (**2016**). "A review of exploratory factor analysis decisions and overview of current praxis: What we ware doing and how can we improve?," Int. J. Human-Comput. Interact. **32**(1), 51–62.
ISO (**2009**). 3382-1, *Acoustics—Measurement of Room Acoustic Parameters—Part 1: Performance Spaces* (International Organization for Standardization, Geneva, Switzerland).
ISO (**2014**). 8586, *Sensory Analysis – General Guidelines for the Selection, Training and Monitoring of Selected Assessors and Expert Sensory Assessors* (International Organization for Standardization, Geneva, Switzerland).

James, G., Witten, D., Hastie, T., and Tibshirani, R. (**2017**). *An Introduction to Statistical Learning with Applications in R*, 8th ed. (Springer Science and Business Media, New York).
Knoblauch, K. (**2014**). *Package "psyphy,"* available at https://cran.csiro.au/web/packages/psyphy/psyphy.pdf (Last viewed September 28, 2021).
Kreimann, J., Gerratt, B. R., and Precoda, K. (**1990**). "Listener experience and perception of voice quality," J. Speech Hear. Res. **33**, 103–115.
Kuusinen, A., and Lokki, T. (**2017**). "Wheel of concert hall acoustics," Acta Acust. Acust. **103**(2), 185–188.
Kuusinen, A., and Lokki, T. (**2020**). "Recognizing individual concert halls is difficult when listening to the acoustics with different musical passages," J. Acoust. Soc. Am. **148**(3), 1380–1390.
Li, J., Lu, X., Cheng, K., Liu, W., and Li, M. J. (**2020**). *Package "StepReg,"* available at http://cran.irsn.fr/web/packages/StepReg/StepReg.pdf (Last viewed September 28, 2021).
LimeSurvey GmbH (**2003**). "Limesurvey: The online survey tool," available at https://www.limesurvey.org (Last viewed September 28, 2021).
Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S. (**2012**). "Disentangling preference ratings of concert hall acoustics using subjective sensory profiles," J. Acoust. Soc. Am. **132**(5), 3148–3161.
Magis, D., and Raîche, G. (**2011**). "catR: An *R* package for computerized adaptive testing," Appl. Psychol. Meas. **35**(7), 576–577.
Müllensiefen, D., Gingras, B., Musil, J., and Stewart, L. (**2014**). "The musicality of non-musicians: An index for assessing musical sophistication in the general population," PloS One **9**(2), e89642.
Murphy, G. (**2004**). *The Big Book of Concepts* (MIT Press, Cambridge, MA).
Nakagawa, S., and Schielzeth, H. (**2013**). "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," Methods Ecol. Evol. **4**(1), 133–142.
Olsson, U. (**1979**). "Maximum likelihood estimation of the polychoric correlation coefficient," Psychometrika **44**(4), 443–460.
Revelle, W. (**2019**). *Package "psych,"* available at https://personality-project.org/r/psych-manual.pdf (Last viewed September 28, 2021).
Riederer, K. A. J. (**1998**). "Repeatability analysis of head-related transfer function measurements," in *105th AES Convention*, San Francisco, Paper No. 4846.
Schröder, D., and Vorländer, M. (**2011**). "RAVEN: A real-time framework for the auralization of interactive virtual environments," in *Forum Acusticum*, pp. 1541–1546.
Soulodre, G. A., and Bradley, J. S. (**1995**). "Subjective evaluation of new room acoustic measures," J. Acoust. Soc. Am. **98**(1), 294–301.
The *R* Foundation (**1993**). "*R*," available at https://www.r-project.org/ (Last viewed September 28, 2021).
Weinzierl, S., Lepa, S., and Ackermann, D. (**2018**). "A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI)," J. Acoust. Soc. Am. **144**(3), 1245–1257.
Weinzierl, S., and Vorländer, M. (**2015**). "Room acoustical parameters as predictors of room acoustical impression: What do we know and what would we like to know?," Acoust. Aust. **43**(1), 41–48.
Wickelmaier, F., and Choisel, S. (**2005**). "Selecting participants for listening tests of multichannel reproduced sound," in *118th AES Convention*, Barcelona, Paper No. 6483.
Wilkens, H. (**1977**). "Mehrdimensionale Beschreibung subjektiver Beurteilungen der Akustik von Konzertsälen" ("Multidimensional description of subjective evaluations of the acoustics of concert halls"), Acustica **38**, 10–23.
Wong, P. C. M., Skoe, E., Russo, N. M., Dees, T., and Kraus, N. (**2007**). "Musical experience shapes human brainstem encoding of linguistic pitch patterns," Nat. Neurosci. **10**(4), 420–422.
Zacharov, N., and Lorho, G. (**2006**). "What are the requirements of a listening panel for evaluating spatial audio quality?," in *Proceedings of the International Workshop on Spatial Audio and Sensory Evaluation Techniques*, Guilford, United Kingdom.