A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI)

Stefan Weinzierl, Steffen Lepa, and David Ackermann

Citation: The Journal of the Acoustical Society of America **144**, 1245 (2018); doi: 10.1121/1.5051453
View online: https://doi.org/10.1121/1.5051453
View Table of Contents: https://asa.scitation.org/toc/jas/144/3
Published by the Acoustical Society of America

A measuring instrument for the auditory perception of rooms: The Room Acoustical Quality Inventory (RAQI)

Stefan Weinzierl,[a] Steffen Lepa, and David Ackermann
Audio Communication Group, Technische Universität Berlin, Einsteinufer 17c, Berlin, D-10587, Germany

(Received 4 June 2018; revised 17 July 2018; accepted 8 August 2018; published online 10 September 2018)

With the Room Acoustical Quality Inventory (RAQI), a measuring instrument for the perceptual space of performance venues for music and speech has been developed. First, a focus group with room acoustical experts determined relevant aspects of room acoustical impression in the form of a comprehensive list of 50 uni- and bipolar items in different categories. Then, $n = 190$ subjects rated their acoustical impression of 35 binaurally simulated rooms from 2 listening positions, with symphonic orchestra, solo trumpet, and dramatic speech as audio content. Subsequent explorative and confirmative factor analyses of the questionnaire data resulted in three possible solutions with four, six, and nine factors of room acoustical impression. The factor solutions, as well as the related RAQI items, were tested in terms of reliability, validity, and several types of measurement invariance, and were cross-validated by a follow-up experiment with a subsample of 46% of the original participants, which provided re-test reliabilities and stability coefficients for all RAQI constructs. The resulting psychometrically evaluated measurement instrument can be used for room quality assessment, acoustical planning, and the further development of room acoustical parameters in order to predict primary acoustical qualities of venues for music and speech. https://doi.org/10.1121/1.5051453

[FM] Pages: 1245–1257

I. INTRODUCTION

Throughout the first half of the 20th century, the investigation of perceptual properties of room acoustical environments was mainly focused on the identification of preferred values for the reverberation time and its frequency dependence. The first perceptual experiments in rooms with modified absorption were already conducted in 1902 by Sabine (1906) and later by Bagenal (1925), while other studies tried to interpolate the reverberation times of existing concert halls recognized for their superior acoustics in order to find target values for acoustical planning (Watson, 1923; Sabine 1928; Knudsen, 1931).

After 1950, the focus gradually widened to other perceptual qualities of performance venues. Somerville and Gilford (1957) defined a glossary of 14 acoustic terms, which were "commonly used to describe the subjective qualities of a concert hall or studio." A similar list of 18 attributes was proposed by Beranek (1962) in his landmark book on *Music, Acoustics and Architecture*, along with providing assumed relations between these perceptual qualities and physical properties of the hall, based on his own intuition and experience. With 16 attributes selected from Beranek's list, Hawkes and Douglas (1971) conducted listening experiments in order to find underlying perceptual concepts on which the ratings of the 16 attributes could be based, many of which turned out to be highly correlated among each other. By using principal component analysis and multidimensional scaling (MDS), they arrived at different solutions with 4–6 independent factors. While Hawkes and Douglas (1971) administered their questionnaire in different British concert halls, Lehmann and Wilkens (1980) used dummy head recordings of the Berlin Philharmonic Orchestra in six different halls in order to vary acoustical stimuli experimentally and captured the assessment of subjects on a semantic differential with 19 different (German) attributes. Their analysis of room acoustic impression ratings delivered three principal components, explaining 89% of the total variance. These were interpreted as *Strength and extension of the sound source*, *Definition*, and *Timbre* of the overall sound (Wilkens, 1977).

The attributes used to describe perceptual qualities of the room acoustical impression in all studies mentioned above were always defined by the investigators themselves, based on their theoretical or practical experience with room acoustic design. In contrast, several studies appeared after 1990, aiming at an empirical approach to identify comprehensive vocabularies valid for listeners with different experiential backgrounds. Such studies, including a qualitative part for the verbal elicitation of the terminology and a quantitative part for the statistical analysis of the generated terms, were focused both on the evaluation of spatial audio reproduction systems (Berg and Rumsey, 2006) and the perception of natural acoustical environments (Lokki *et al.*, 2012). Using stimuli provided by impulse response measurements in 8 different concert halls, encoded in Ambisonics B-format and reproduced by a 14-channel loudspeaker system, Lokki *et al.* (2012) generated a vocabulary of 60 attributes, which were elicited from 17 subjects, using a method called vocabulary profiling. Based on the individual ratings of these

[a]Electronic mail: stefan.weinzierl@tu-berlin.de

attributes, the authors identified three principal components explaining 67% of the total variance. For interpretation, the attributes were clustered based on their loadings on the principal components. A group of attributes interpreted as *proximity* descriptors was identified as crucial for the preference of the room acoustical environments presented. It was not the focus of this study, however, to develop a generic and psychometrically validated measuring instrument for room acoustical perception.

This, in turn, is the aim of the present paper. Previous research already provided an idea of the perceptual dimensions human listeners may employ when describing the acoustical impression of a room. Due to shortcomings in terms of experimental stimuli, participating subjects, and statistical analysis techniques employed in these works, room acoustics, however, still lacks a psychometrically valid measurement instrument that would be suited as a standardized scale for room acoustical impression. In the following, we will depict the most important shortcomings and how these were addressed in the present study.

A problem that pertains to most of the *early* studies in acoustical room impression measurement (Hawkes and Douglas, 1971; Wilkens, 1977; Lehmann and Wilkens, 1980) is the lack of experimental control concerning the stimuli presented. Any studies that employed real existing rooms, during a concert or as a dummy head recording in the lab, could only work with a lower number of room stimuli, and always risked the influence of "hidden confounders" such as the audio content, the visual impression, or the musical performance. Presenting the whole breadth of possible room acoustical conditions as purely acoustic stimuli to a large number of participants with full control of all confounding variables seems only possible with state-of-the-art technologies for auralization.

A second challenge lies in the sample of rooms presented. The general identification of latent variables of room acoustical perception, i.e., a stable factor analytic solution of the measured data, which is valid beyond the specific sample of rooms used in the test, cannot be expected for a too small set of stimuli. In order to identify five largely independent perceptual dimensions, a set of at least $2^5 = 32$ stimuli would be required so that all perceptual qualities are able to vary independently across stimuli and hence can be properly identified by a factor analysis. Furthermore, only with a sufficiently large sample of rooms, the results can be considered representative for the targeted population of room acoustical conditions. Comparing these requirements with the sample sizes used in the above mentioned studies, with typically six (Lehmann and Wilkens, 1980) or eight rooms (Lokki *et al.*, 2012), it becomes obvious that neither the dimension of the perceptual space, i.e., the number of latent variables, nor the structure and interpretation of the adopted factor solution can be reliably determined. The same deficit has been identified by Gade (2010) for the problem of room acoustical perception by musicians on stage.

A related problem is the general low number of subjects and a convenience selection bias of the participant samples, which pertains to all previous works in the field (Hawkes and Douglas, 1971; Wilkens, 1977; Lehmann and Wilkens, 1980; Berg and Rumsey, 2006; Lokki *et al.*, 2012). Psychological measurement scales are typically constructed to work with the normal population, or at least a certain social group from which a quasi-random sample should be drawn. Instruments developed only with students from an acoustical institute, for example, might over- or underemphasize certain perceptual dimensions. Furthermore, in order to perform factor analyses that are necessary to assess the dimensionality of perceptual constructs represented by questionnaire item batteries, it is recommend to use sample sizes of $n = 100$–200 subjects or at least a sample of three times the number of employed items (MacCallum *et al.*, 1999), which was never the case in previous experiments. Hence, an improved systematic approach would have to draw on a substantially higher number of individuals, ideally recruited from the general population—as long as the scale is not explicitly constructed solely for room acoustical experts.

Finally, prior studies generally omitted to analyze the psychometric qualities of perceptual constructs and questionnaires based on them in terms of validity, reliability, and measurement invariance. Concepts, analysis techniques, and quality criteria for this purpose have been developed extensively in the social sciences during the 20th century (Vooris and Clavio, 2017). Typical requirements for the quality of up-to-date psychological questionnaire instruments comprise the use of true *latent measurement models* [confirmatory factor analysis (CFA) instead of principal component analysis (PCA); see Fabrigar *et al.*, 1999], the demonstration of convergent and discriminant validity (do the scale's subdimensions actually measure what they are supposed to measure and are sub-dimensions sufficiently different from each other?), as well as demonstrations of sufficient construct reliability (how precise does the scale measure?) and measurement invariance (Millsap, 2011) across time, stimuli, and populations of interest (are the scale's measurements independent of the experimental factors employed?).

In order to achieve and demonstrate an acceptable degree of validity, reliability, and measurement invariance, and practically deal with different sources of measurement error in applied studies (Schmidt and Hunter, 1999), psychological scale development nowadays typically relies on the technique of *latent variable models* (Loehlin, 2004). Since past studies on room acoustics predominantly drew on PCA/clustering techniques, only a minority of these questions could be systematically addressed. Furthermore, none of the prior studies calculated adequate scale reliability coefficients (Cho, 2016) and most of them also neglected item *re-test reliability* completely (Lokki *et al.*, 2012, being an exception, however, without documenting latency between measurements). The latter, however, is typically deemed the most important coefficient in development of reliable scales since Cronbach (1947).

The present approach is an attempt to systematically develop a standardized measurement instrument for room acoustical quality assessment. This was achieved by first acquiring expert knowledge from different domains of room acoustics to provide a comprehensive terminology for such an instrument. In a second step, listening experiments with acoustical experts and laymen using a large number of rooms

of different types was conducted with different audio content (single instruments, orchestral music, and dramatic speech in order to address the most important acoustic performance types and their specifics). The goals of the subsequent analysis were to

- find an exhaustive list of verbal attributes that describes all relevant room acoustical properties,
- identify the best suited items of this list to form a standardized measurement instrument,
- analyze the underlying dimensions of room acoustical impressions,
- construct a measurement instrument based on these dimensions and corresponding items,
- demonstrate reliability of the new instrument across and within raters,
- demonstrate measurement invariance of the new instrument across experimental conditions such as audio content type and subject samples, and
- demonstrate sufficient discriminant validity of its subdimensions.

In order to realize this in an experimental setting that permitted controlling for any possible confounders, we drew on room acoustical simulation, anechoic audio content, and auralization by dynamic binaural synthesis. Based on the comprehensive psychometric data collected, we will suggest three factor solutions of different size, along with practical recommendations for their use.

II. METHODS

A. General considerations

Aiming at a generic measuring instrument for the room acoustical impression that can be used for a variety of room acoustical conditions and for listeners with different levels of expertise, we combined a qualitative step for the generation of the perceptual attributes with a stimulus-based evaluation of the questionnaire with respect to different test-theoretical criteria. Whereas individual elicitation methods are always confronted with the problem of merging individual vocabularies into a valid group vocabulary, group methods directly aim at deriving a consensual language. Since we had access to a large reservoir of room acoustical experts, who could be invited for face-to-face moderated roundtable discussions, we decided for the focus group method (Stewart *et al.*, 1990), which has been successfully applied for similar tasks (Lindau *et al.*, 2014). For the stimulus-based evaluation and refinement, we used the preliminary vocabulary generated by the focus group as a rating instrument to have the room acoustical impression of 35×2 (rooms $\times$ listener positions) different room acoustical conditions evaluated by 190 subjects. Based on the statistical analysis of these data, the vocabulary was supposed to be reduced to a questionnaire serving as a validated measuring instrument.

B. Expert focus group

Methodologically, a focus group may be regarded as a combination of a guided interview and a group discussion. This combination is particularly well-suited for the elicitation of expert knowledge, as experts are routinely used to discourse-based revelation of consensual knowledge (Krueger, 2014). As a moderator, the first author had to control for unwanted group effects, e.g., by restraining "leading" and motivating "hiding" discussants, and being sensitive to nonverbal communication. As participants, a group of German speaking experts was invited, representing a wide professional experience in room acoustical consulting as well as room acoustical and psycho-acoustical, academic expertise. With some changes over the different meetings regarding group size and composition, a total of 12 experts (aged 35–77 yr) participated (see the Acknowledgments for the composition of the group). Discussions were held at three meetings in Berlin between November 2014 and March 2015.

The vocabulary to be developed was defined as a list of auditory qualities and respective rating scales for room acoustical environments for speech and music, from rehearsal spaces, lecture halls, to chamber music and symphonic concert halls to large cathedrals. Typical intended applications of the future vocabulary were listening tests of existing halls, perceptual qualities for the description, and the targeted planning of new rooms, as well as for the further development of room acoustical parameters as predictors of these qualities. These objectives were aggregated into a mission statement, which was *creating a vocabulary for the perceptual assessment of room acoustical environments for speech and music from the audience's perspective*. With the final addition, we acknowledged that speakers, singers, or musical performers have a different experience of their spatial environments and might use a different vocabulary to describe it.

The experts were instructed that terms should address the perceptual domain, rather than physical quantities or measures. Furthermore, they were asked to aim at completeness of the overall vocabulary, while, at the same time, consider its practical relevance. Descriptors should be formulated as semantically unidimensional as possible, and complemented by a short explanatory description and labels for scale poles. All major decisions were to be agreed upon by simple majority. In order to consider the state of the art, all attributes used in important, earlier publications (Hawkes and Douglas, 1971; Lehmann and Wilkens, 1980; Barron, 1988; Jullien *et al.*, 1992; Beranek, 1996; Lokki, 2005; Lokki *et al.*, 2012; Lindau *et al.*, 2014) were presented to the group as terminological options. Each session lasted about 4 h, summing up to 12 h of discussions. The result of the focus group, a preliminary German room acoustical quality inventory (RAQI) consisting of 50 items, was the basis for the subsequent listening test aiming at a stimulus-based evaluation and reduction of the vocabulary.

Parallel to the listening test, we invited four room acoustical experts to translate the vocabulary to English. The target language was assumed to be a "technical community language" in the field of acoustics, which is neither a real United States (U.S.) nor United Kingdom (U.K.) native English, or any other native English. Hence, the criterion for the translator panel was not native language skills but rather longstanding professional expertise in an international, English-speaking room-acoustical context (see the

Acknowledgments for the composition of the group). The translators were provided with the German descriptors, sometimes already including English terms suggested by the German panel, and were asked to produce adequate English terms. The translation took place over two sessions of 4 h between March and April 2017.

C. Listening experiment: Stimuli

Since the generated questionnaire was meant to be valid for music and speech venues of different size and architectural design, we endeavored to create a representative sample of room acoustical environments for this purpose. As a guiding principle we created three-dimensional (3D) models for rooms of different size (166 m^3–43 790 m^3), different mean absorption ($\alpha = 0.04$–0.42), and different architectural types (Fig. 1). Most of these models correspond to existing concert halls, lecture halls, etc., and some were created without a real equivalent in order to fill gaps in the guiding parameter space.

For the auralization of solo music and speech, binaural room impulse response (BRIR) datasets were simulated for one central, front-stage source position and two receiver positions. Both had a distance to the source of at least twice the critical distance to avoid a predominance of the direct sound. One was at a central position in the parquet floor and the other was lateral and further from the stage. According to ISO 3382–1 (2009), the source was positioned at a height of 1.5 m and the receivers at 1.2 m. For the simulation of the orchestra, BRIR datasets were simulated for 66 source positions on stage, corresponding to a typical symphonic stage plan (Fig. 2). The source directivities used for the simulation were taken from measurements with a 32-channel spherical microphone array (Shabtai *et al.*, 2017; Weinzierl *et al.*, 2017). After spherical harmonic decomposition, the pitch-dependent directivities were subject to a weighted average, using the typical pitch distribution of classical orchestra

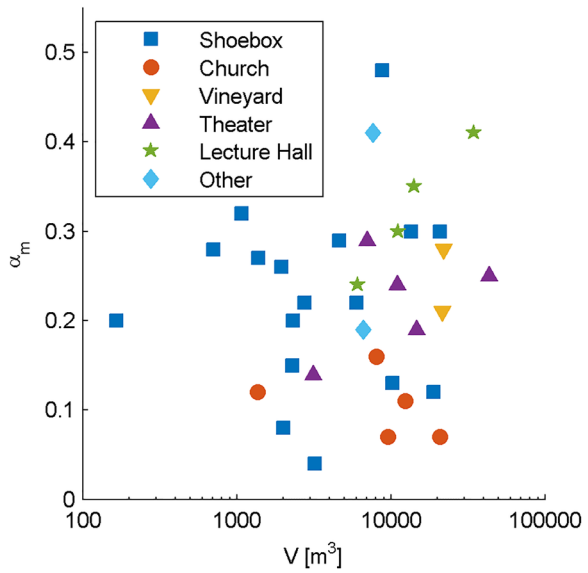

FIG. 1. (Color online) Computer models for 35 venues of different size (V), mean absorption (α_m) and architectural design (see inset) were created as a representative sample of room acoustical environments for music and speech.

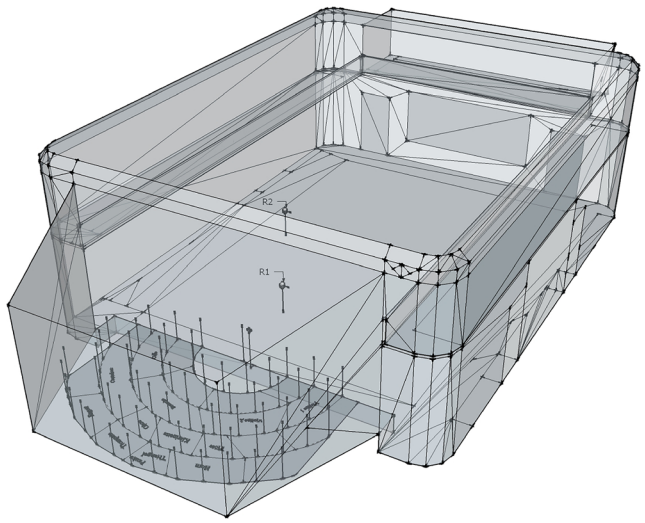

FIG. 2. (Color online) Room acoustical model, including the symphonic stage plan used to define the source positions for the BRIR datasets simulated for the auralization of the orchestra, as well as the two receiver positions used.

instruments as weights (Quiring and Weinzierl, 2016), and imported into the room acoustical simulation in OpenDAFF format (Wefers, 2010).

The BRIRs were simulated with the RAVEN software, using a hybrid mirror image and ray tracing algorithm (Vorländer *et al.*, 2014), and head-related transfer functions (HRTFs) taken from the FABIAN database (Brinkmann *et al.*, 2017b; Brinkmann *et al.*, 2017a). For every source-receiver transfer path, 71 BRIRs were simulated for horizontal head orientations between ±70° in steps of 2° and 0° vertical orientation (Lindau and Weinzierl, 2009). The BRIRs were separated at the assumed perceptual mixing time into a dynamically convolved, early part, and a static, late part (Lindau *et al.*, 2012), and converted to the SOFA format (AES 69, 2015). For dynamic binaural synthesis, we used the SoundScape Renderer (Geier *et al.*, 2008), as well as extra-aural headphones optimized for spectral compensation of the headphone transfer function (Erbes *et al.*, 2012).

Since the perceived acoustic qualities of rooms can be expected to be influenced by the source signal, we selected three different audio contents for excitation:

- Solo music (J. Clarke: *Trumpet Voluntary*) (1:30 min);
- Orchestral music (J. Brahms: *Symphony No. 4*, 3rd movement) (1:30 min);
- Dramatic speech (Cicero: *Catiline Speech*/German translation) (6:00 min).

Speech and solo trumpet were recorded in the fully anechoic chamber of TU Berlin ($V = 1070$ m^3, $f_g = 63$ Hz). For the auralization of an orchestra, an anechoic recording of the 4th symphony of J. Brahms was kindly provided to us as single tracks (Vigeant *et al.*, 2010). These recordings were edited by a professional sound engineer in order to reduce asynchronicities resulting from the sequential recording process. Subsequently, we multiplied the existing tracks with three first and three second violins, one viola, one violoncello, and one double bass by adapting the algorithm of

Pätynen *et al.* (2011), using a single-channel recording of the Berlin Philharmonic Orchestra as a model for the orchestral distribution of onsets and pitch. By this means, the string section of the recording was augmented to 12 first and 11 s violins, 10 violas, 9 violoncelli, and 8 double basses (Grigoriev, 2017). The speech and the trumpet stimulus were recorded with a professional actor and a professional musician in the fully anechoic chamber of TU Berlin.

For solo music and speech, BRIR datasets for 35 rooms at 2 listening positions each were generated. For the orchestral piece, 25 rooms at 2 listening positions were selected, leaving off 10 rooms where the stage area would not be large enough for an orchestra. Thus, in total, 190 room acoustical conditions were simulated for the listening experiment.

D. Listening experiment: Procedure

Participants were recruited from the subject pool of the Audio Communication Group, by advertisements in mailing lists, and via classified ads in various local online portals (see Sec. III B for descriptive statistics on the composition of the subject sample). The subjects, who had no self-reported hearing loss, were offered a compensation of 10 EUR. For the test, they were first presented with a list of 46 items selected from the preliminary RAQI with explanations of the meaning of each item on a sheet of paper. They were asked to thoroughly read the explanation and check back if the meaning of each item and scale poles were unclear. After the start of the experiment, 14 randomly chosen stimuli from the overall set of 190 different possible stimulus combinations were presented aurally via headphones in random order to each subject. The stimulus selection from the pool was pre-calculated in a way that ensured equal frequency for each stimulus of the set across all trials for all participants, thereby forming a balanced incomplete block design. Each trial consisted of a continuously looped aural presentation of the stimulus, with 2 s of silence between the loops. While listening to a stimulus, subjects were presented the 46 items one after another in randomized order on a screen, including an analog visual rating scale reaching from 0 to 100 in the case of asymmetrical response options, or −50 to +50 in the case of symmetrical response options. Four of the items were only suitable for certain stimuli.[1] Subjects were told to either rate the stimulus using the scale or to skip an item that they considered unsuitable for the present stimulus and proceed to the next item. They could use as much time for each rating as they wanted to.

At the end of the experiment, participants filled out another questionnaire on gender, age, and educational background, as well as on their musical expertise in terms of whether they played an instrument, how often they listen to music, and how many concert performances they had visited in the last 12 months. Additionally, they were asked to report on their room-acoustical expertise in terms of the number of years of education related to room acoustics and the number of years they had spent in a job related to room acoustics. Finally, they were asked for their willingness to take part in the follow-up experiment. Altogether, the listening experiment lasted about 110 min [$M = 103.76$; standard deviation (SD) = 19.8 for the rating part alone].

The follow-up experiment used exactly the same stimulus material and employed the same procedures as the initial listening experiment, apart from the questionnaire on sociodemographics and expertise. All participants of the initial experiment who had left their contact information were recontacted and invited to participate again with the same monetary incentive. The follow-up took place approximately $M = 42$ days (SD = 37) after the initial experiment. It was possible to gain $n = 88$ participants of the original sample, resulting in a coverage rate of 46% for the re-test procedure. The re-test itself was an exact replication of the first experiment for each participant. Only the order of the questionnaire items was newly randomized with each stimulus. Most participants were now able to perform the rating quicker, resulting in a mean processing time of $M = 41.5$ min (SD = 46.3).

E. Statistical analysis

For all original questionnaire items of the RAQI, the value range, means, and SDs were calculated, as well as skewness and kurtosis, and the number of missing values per item. Furthermore, histograms for visual inspection were created to check for obvious outliers and possible zero-inflation in single items.

Re-test reliability was estimated for all items by calculating pairwise Pearson correlations between the raw score of any initial and corresponding follow-up item measurements across all data from the $n = 88$ respondents that had taken part in both experiments, including $n = 28$ respondents who had at least one year of professional education or job experience in room acoustics (their score labelled as *expert reliability*). When an item had been skipped in one of the two occasions by a participant, the respective case was not included in the reliability analysis (pairwise missing deletion).

An initial *exploratory factor analysis* (EFA) based on the common factor approach was conducted with the software package MPlus 7 (Muthén and Muthén, 2012). The aim was to estimate the number of independent latent dimensions contained in the full RAQI item data matrix of the experiment (leaving out the four items that had not been presented with all stimuli) by using the scree- and Kaiser-criterion as a starting point for constructing a multidimensional measurement model. For this, eigenvalues for one to nine factors and respective factor loading matrices were calculated using maximum likelihood parameter estimates that are robust to non-normality and non-independence of observations (MLR estimator) and the Crawford-Ferguson oblique rotation method with imputing missing data from the model. This analysis was done on the data matrix from the initial experiment only, leaving the follow-up data for later cross-validation. As it later turned out, the EFA results were ambivalent in terms of the number of independent dimensions (2,4,6,9), which is why some of the following analyses were initially performed for all four variants.

For the four different possible basic measurement models established by EFA, refined simple-structure measurement models where construed in MPlus 7 by using a four-step heuristic: Initially, each item was assigned to one of the

factors on the basis of its highest EFA loading. The resulting model was then optimized by first removing items with substantial cross-loadings (<0.1 difference in EFA loading values) and items with small EFA loadings (<0.4). Then, a series of *confirmatory factor analyses* (CFA) employing the MLR estimator with missing data imputation was conducted to consecutively remove single items on the basis of estimated highest modification indices. In this way, items were removed step-by-step from the measurement models up to a point where an implied removal would have led to less than three items per factor, otherwise, the overall fit of the measurement model was already good. In a final step, modification indices were inspected again to determine if single and theoretically explainable cross-loadings would substantially improve the model fit. Afterwards, root mean square errors of approximation (RMSEA), comparative fit indices (CFIs), and standardized root mean square residual (SRMR) coefficients were calculated as fit indices for the resulting final CFA measurement models for two, four, six, and nine factors, as well as congeneric construct reliability (CR) and average variance extracted (AVE) for each factor. CR denotes the internal consistency of a factor construct (Cho, 2016) and is a measure comparable to Cronbach's alpha, while the AVE measures how well a factor explains the scores of its underlying items, and should conventionally be above 0.5 (Fornell and Larcker, 1981). For an unbiased calculation of CR and AVE, and to achieve further improvement of the model, we also calculated zero-inflated regression models for items with a substantial amount of zero-inflation according to histogram inspection of item scores. After finally deciding for one of the solutions on the basis of theoretical considerations, fit indices, and the Fornell–Larcker criterion of discriminant validity (Fornell and Larcker, 1981), factor scores were estimated for each combination of subject case and stimulus in the dataset. For *cross-validation*, we calculated the fit of the final CFA model again, but this time drawing on the data from the follow-up experiment comprising $n = 1216$ observations from 88 subject clusters, having fixed the means and intercepts to the values of the original experiment.

To test for *measurement invariance across measurement occasions*, we estimated a longitudinal CFA with correlated measurement errors and identical factor-item-configuration as in the final CFA solution. We stepwise increased equality constraints and calculated resulting fit indices in order to check for fit differences due to different assumed types of configural, metric, or scalar measurement invariance (Steenkamp and Baumgartner, 1998). We considered a drop in CFI of greater than or equal to 0.01 and a CFI value lower than 0.95 as an indicator for substantial loss in fit between model versions (Cheung und Rensvold, 2002). After having established scalar invariance in this way, we calculated factor score stabilities (SCs, correlations of factor scores within subjects across time) from the estimated pairwise factor covariances.

To test for *measurement invariance across audio content*, we estimated a multiple group CFA with identical factor-item-configuration as in the final CFA solution, using the type of audio content (speech, solo trumpet, orchestra) as a group factor to divide the experimental data matrix in three subsamples. To test for *measurement invariance across acoustical expertise*, we estimated a multiple group CFA with identical factor-item-configuration as in the final CFA solution, using a binary expertise variable (laymen, experts) as a group factor. For this, we coded all participants with more than one year of professional education in room acoustics as "experts." To test for *measurement invariance across listening positions*, we estimated a multiple group CFA with identical factor-item-configuration as in the final CFA solution, using the listening position (front center, middle right) as a group factor to divide the experimental data matrix in two subsamples. In all cases, we then stepwise increased equality constraints and calculated resulting fit indices to check for fit differences between model versions, again drawing on CFI drop as a decision criterion. When encountering a substantial drop in fit, we freed equality constraints of single items on the basis of modification indices to arrive at a well-fitting alternative model with partial measurement invariance.

Within all performed EFA/CFA analyses, we accommodated for the clustered structure of data by using the *cluster* option of the MPlus 7 software, resulting in standard errors adjusted by a sandwich-estimator to correct for non-independence of the 14 measurements "within" the test subjects.

In order to test to what extent the estimated RAQI factor scores would be helpful in discriminating between the 35 simulated rooms used in the experiment, we performed a *linear discriminant analysis* in SPSS 23 that employed the estimated factor scores from all experimental observations of the original experiment as the training set and the identity of the rooms used in the experiment as the class variable to be recovered. The analysis also comprised of several ANOVAs that tested whether each of the factors significantly contributed to differentiating rooms, as well as of a naive Bayes classification algorithm, which employed estimated decorrelated discriminant functions to recover original room identity.

III. RESULTS

A. Expert focus group

The consensus vocabulary generated by the expert focus group consists of 50 perceptual qualities related to the timbre, geometry, reverberation, temporal behavior, and dynamic behavior of room acoustical environments, as well as overall qualities.[1] While some attributes reflect lower order qualities closely related to temporal or spectral properties of the audio signal ("loudness," "treble/mid/bass range tone color," perceived "size," and "width" of sound sources), other attributes reflect higher-order psychological constructs, supra-modal, affective, aesthetic, or attitudinal aspects ("clarity," "intimacy," "spatial transparency," or "ease of listening"). All descriptors are complemented by bipolar scale labels.

In contrast to previous studies, the vocabulary is generally more specific, for example, by distinguishing between the "duration" and the "strength of reverberation," by specifying clarity as "temporal clarity" (to avoid a more timbre-related interpretation), or by expecting certain attributes

such as the perceived "distance," "size," "width," or "spatial presence" to be assigned to a specific reference object, such as a speaker, a musical instrument, or a group of musical instruments, in order to increase the consistency of the ratings. For the same reason, most of the attributes (except for some where the expert group did not consider it necessary) are complemented by a short circumscription.

B. Listening experiment—Descriptive statistics

The 190 subjects taking part in the experiment had a mean age of 32 yr (SD = 11) and included 113 male persons (59%). A majority of participants had a high school degree or higher educational attainment (86%) and most of them were Germans ($n = 180$). On average, they had visited 5 concerts (median) within the last 12 months and most of them were listening to reproduced music several times a week. In terms of musical and acoustical expertise, 70% of them played an instrument, 9% had studied musicology, 33% studied an acoustic-related subject or had another type of professional education with affinities to room acoustics, and 14% even had a profession directly related to room acoustics.

Four of the 50 items of the original vocabulary were not used in the listening test because they were considered irrelevant for the presented stimuli (Nos. 1, 16, 38, 50). The descriptive statistics for the 46 tested attributes[1] indicate problematic items: Items with a high percentage of missing data, such as those for judgments of the "mid-range characteristics" in timbre, seemed to be either irrelevant or difficult to assess for a large number of participants, which makes them unsuitable for the final scale. Three items having at least 20% of missing values ("mid-range characteristic," "bass range characteristic," "loudness balance between strings and wind instruments") were thus excluded from later factor analysis. Comparing item means with the scale center values indicates items being obviously formulated in a way that lead to an overly easy or too difficult approval, such as for "blend" and "flutter echo." The same two items also exhibit the largest problems in terms of skewness of score distribution. The inspection of histograms also shows zero-inflation for a larger number of items, i.e., items with frequent zero-valued observations, as well as the special case of ease of listening where frequency inflation was obtained at the scale maximum. Finally, we found clear outliers for several reverberation-related items and several of the overall quality items, as well as some problems with excess of kurtosis. While these might violate requirements of certain statistical analyses, neither outliers, nor kurtosis turned out to be a problem in the later factor analyses, while we had to use zero-inflated regression models to compensate for strongly skewed item scores.

Generally, the calculated retest-reliabilities appear rather low for a large number of questionnaire items, given that perceptual impressions should not exhibit true change after a time period of 42 days. Rules of convention (Cicchetti, 1994) would consider correlation coefficients equal or larger than 0.75 as indicators of "excellent" reliability, 0.6–0.74 for "good" reliability of single items. Items with retest-reliabilities below 0.5 contain more measurement error than true scores and appear therefore not suitable for the population addressed in this experiment. This is true for 12 of the items developed in the expert focus group. The retest reliabilities tend to be higher for the 28 room acoustical experts taking part in the test, but 5 of the item scores are still not satisfying even for experts with values below 0.5; these were excluded from the later factor analysis ("mid-range characteristic," "sharpness," "homogeneity," "responsiveness," "loudness balance between strings and wind instruments").

C. Factor solutions

Applying the standard criteria for the optimal number of factors in the EFA does not yield a unique solution. The Kaiser criterion (eigenvalues > 1) points to nine factors, while the scree-criterion (all factors above a visible "knee") could, in this case, allow to decide for two, four, six, or nine factors (Fig. 3).

Calculating incremental increases of the RMSEA across solutions, as suggested by Fabrigar *et al.* (1999), leads to the same result. Accordingly, we performed the CFA analyses with the four-step heuristics (cf. Sec. II E) for all four variants and calculated the fit of the resulting solutions (Table I). For the six- and nine-factor solutions, we also fitted variants that included a double loading of the item "echo" on two factors, since this was indicated by modification indices and also appeared theoretically reasonable—it is a common anecdotal fact that laymen, in particular, have difficulty discriminating between reverberation and echo.

The X^2-test for all of the tested measurement model variants became significant. This, however, is expected for models with large n and many parameters (Hu and Bentler, 1999). Accordingly, following these authors, we evaluated the model fit by inspecting RMSEA, CFI, and SRMR but used the cutoff criteria appropriate for small-n samples, since the true sample size in our repeated measurement experiment

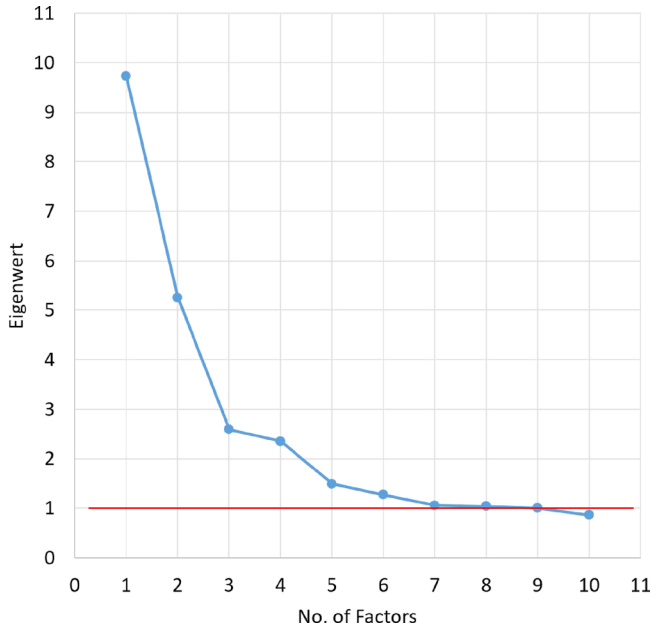

FIG. 3. (Color online) Scree plot for the exploratory factor analysis [EFA, Crawford-Ferguson (CF)-varimax oblique] of the 46 RAQI items.

TABLE I. Fit statistics of tested CFA factor solutions. $n = 2660$ (190 subject clusters); MLR estimation with missing data imputation; SEM R^2 = explained variance when regressing factor 1 (*quality*) on all others.

Model	items	AIC	BIC	X^2	df	p	RMSEA	**CFI**	SRMR	SEM R^2
Two factors	8	184570.352	184717.504	89.772	19	<0.01	0.037	0.986	0.034	0.077
Four factors	14	316018.072	316300.604	340.562	71	<0.01	0.038	0.970	0.041	0.354
Six factors	20	453579.646	454021.102	829.632	155	<0.01	0.040	0.948	0.048	0.476
Six factors_*b* (echo 2*x*)	20	453382.819	453830.162	715.706	154	<0.01	0.037	0.956	0.045	0.485
Nine factors	29	645487.386	646211.374	1914.435	341	<0.01	0.042	0.918	0.056	0.373
Nine factors_*b* (echo 2*x*)	29	645289.297	646019.171	1800.085	340	<0.01	0.040	0.924	0.054	0.374

was only $n = 190$. Along these lines, RMSEA and SRMR turned out to be excellent for all tested solutions (all below 0.05 and 0.06), while CFI was very good (>0.950) only for the two- and four-factor solutions, as well as for the six- and nine-factor model variants with a double loading. Since we did not aim at reaching a most parsimonious solution but to explore room acoustic impression dimensionality as exhaustively as possible, we continued further analyses with the four-, six-, and nine-factor solutions with double loading. The calculated CR scores for each factor dimension turned out to be excellent (>0.7), except for *Coloration* and *Intimacy*, where reliability was only moderate. The AVE, indicating how well the corresponding items can be explained by each factor, misses the usual criterion of AVE > 0.5 only narrowly for some of the factors. The Fornell–Larcker criterion for discriminant validity, stating that the AVE of each of the latent constructs should be higher than the highest squared correlation with any other latent variable and indicating how well the factors describe *different* aspects of the room acoustical impression, is not fulfilled for *Intimacy* and *Liveliness* in the nine-factor solution, where these factors correlate moderately to highly with several other factors. Hence, according to the standard criteria for psychological measurement instruments, the six-factor solution exhibits the best fit with the sample observations at hand.[1]

Depending on the desired degree of differentiation and the allowed size of the questionnaire, however, all three solutions can be recommended as a measurement instrument for room acoustical impression. The considerable correlation of the nine factors among each other, compared to the values for four and six factors, seems to be an acceptable concession in favor of the additional information, which can be gained by the additional sub-dimensions. The possible questionnaires resulting from the four-, six-, and nine-factor solutions, are presented in Table II. They contain three or four items to measure the corresponding latent variables as sub-dimensions of the room acoustic impression. They also contain four room acoustical attributes with good re-test reliabilities, which could not be assigned to any of the dimensions identified by EFA, but might still be interesting to add. These were "metallic tone color," "openness," "attack," und "richness of sound."

D. Cross-validation, measurement invariance, stability, discriminant analysis

We have calculated a series of quality criteria for the six-factor solution as the empirically best substantiated measurement instrument for room acoustical impression. First, a *cross-validation* with the sample of the follow-up experiment, conducted with fixed item loadings and intercepts to the values of the original experiment, resulted in an excellent fit with $X^2 = 405.422$; degrees of freedom = 195; $p < 0.01$; RMSEA = 0.030; CFI = 0.962; SRMR = 0.056.

Different tests of measurement invariance for the six-factor RAQI were performed.[1] Testing for *measurement invariance across time* by help of a longitudinal CFA indicated scalar invariance (Steenkamp and Baumgartner, 1998). Calculating stability coefficients for the six-factor solution (SC, intra subject reliability coefficients[1]) yielded values well above the threshold of 0.7, except for the *quality* factor, which seems to exhibit substantially less within-subject stability (SC) across time. A test of *measurement invariance across different sound sources and audio content* (solo music, orchestra, speech) yielded scalar invariance only after freeing equality constraints of the intercepts for a number of items: The mean values for "brilliance," "boominess," "envelopment by reverberation," "echo," "liking," and "room acoustic suitability" were significantly different for the different audio content. Thus, only metric invariance can be assumed for the complete RAQI across different types of audio content. The analysis of *measurement invariance across subjects with different expertise* signaled only metric invariance, too. Only after freeing the intercepts of "treble range characteristic," "comb filter coloration," and "size," partial scalar invariance could be demonstrated. The test *for measurement invariance across the two listening positions* used in the experiment resulted in scalar invariance without any further problems. We can thus conclude that not only the structure of the model, i.e., the factor structure and the corresponding items, but also their loadings on the different factors are invariant to different times of measurement, different audio content (solo music, orchestra, speech), subjects with different expertise and different listening positions within the same room. The mean values of the item scores (intercepts), however, can vary for different audio content and different listener expertise. Surprisingly, the mean values are *not* significantly different for assessments from different locations in the hall.

The univariate ANOVAs for equality of group means conducted as part of the discriminant analysis resulted in $p < 0.001$ significant differences between rooms on all factors of the final six-dimensional RAQI scale. The classification statistics for recovery of room identity by help of the discriminant functions resulted in 13.5% correct

TABLE II. Three possible solutions of the CFA yielding four, six, and nine factors as sub-dimensions of room acoustical impression. The corresponding questionnaires would contain 14, 20, and 29 items, which are given with corresponding poles. Weights (W) and intercepts (I) should be used to measure factors and for structural equation analysis (see Sec. IV). Four additional items with high re-test reliability, which could not be assigned to one of the factors, are given below.

			Factors	Items	Poles	W	I
9-factor RAQI	6-factor RAQI	4-factor RAQI	Quality	Liking	I like it – I don't like it	1.0	2
				Room acoustic suitability	suitable – not suitable	1.0	56
				Ease of listening	difficult – effortless	0.9	10
				Global balance	balanced – unbalanced	0.8	4
			Strength	Size	small – large	1.0	57
				Loudness	soft – loud	0.7	64
				Width	small – large	0.8	57
			Reverberance	Duration of reverberation	short – long	1.0	47
				Reverberance	dry – reverberant	1.0	54
				Strength of reverberation	weak – strong	1.0	51
				Envelopment by reverberation	weak – strong	0.7	48
			Brilliance	Brilliance	not brilliant – very brilliant	1.0	48
				Tone Color bright/dark	bright – dark	-0.8	-7
				Treble range characteristic	attenuated – emphasized	0.7	3
			Irregular decay	Flutter Echo	none – very strong	1.0	26
				Echo	none – very strong	0.7	40
				Irregularity in sound decay	none – very strong	0.9	32
			Coloration	Boominess	not boomy – very boomy	1.0	37
				Roughness	not rough – very rough	0.7	31
				Comb filter coloration	none – very strong	0.8	34
			Clarity	Temporal clarity	clear – blurred	1.0	10
				Spatial transparency	blurred – transparent	1.0	1
				Precision of localization	precise – diffuse	-0.8	-6
			Liveliness	Liveliness	dead – lively	1.0	11
				Spatial presence	low – high	1.0	63
				Dynamic range	small – large	0.9	50
			Intimacy	Intimacy	remote – intimate	1.0	-3
				Distance	close – distant	-0.8	51
				Warmth	cool – warm	0.5	3
Single items				Metallic tone color	not metallic – very metallic		
				Openness	open – constricted		
				Attack	soft – crisp		
				Richness of sound	low – high		

classifications compared to an *a priori* probability of 3.2% for most rooms (2.1% for the ten rooms not suitable for orchestra play). Hence, the factors are able to reliably discriminate between different rooms, even if the factor scores alone are, of course, not able to identify a specific room.

E. Primary research data

All primary research data of the current study are available as a digital publication (Ackermann *et al.*, 2018). It includes 3D models of the 35 acoustical environments in Sketchup format with source and receiver positions, the acoustic properties of the surfaces, the simulated monaural and binaural impulse responses, as well as the item and factor scores of the listening test for each of the 35 rooms.

IV. DISCUSSION

With the current investigation we have developed a measuring instrument for the perceptual space of performance venues for music and speech. In a first step, a vocabulary of 50 attributes was generated by an expert focus group, including scholars as well as room acoustical consultants. Even if some of the attributes turned out to be unsuitable in the following stimulus-based evaluation, the item battery itself can be considered as an attempt to standardize the mostly inconsistent terminology to assess the qualities of performance venues for music and speech. As the result of extensive discussions about the relevance and the denomination of room acoustical qualities, the list might be an attractive resource for the generation of questionnaires for the assessment of room acoustical environments, ensuring the comparability of the results for subsequent meta-analyses. The assessment of a new performance venue could, for example, then be compared to existing assessments of reference rooms.

In the second part of this study, we employed a listening experiment including a follow-up test 6 weeks later, which used the dynamic binaural synthesis of a diverse sample of

35 room acoustical environments to assess the suitability of 46 of the attributes proposed by the expert focus group. The subject sample largely consisted of music-interested laymen and a smaller number of individuals with professional education in room acoustics. It is particularly instructive to consider the calculated re-test reliabilities: A majority of them turned out to be poor and only three items related to the strength and duration of reverberation exceeded conventional reliability thresholds of $r = 0.7$. Many other items, including popular ones in room acoustics such as "sharpness" or "transparency," turned out to be based on rather unreliable judgements, using the variation over time within subjects as an indicator. In contrast, the variation across subjects (as measured by item standard deviation) was quite well-balanced across items, and only a few items, such as the overall quality judgment, exhibited obvious outliers from normality between subjects. The general weak single item reliability across time indicates that room acoustical impressions of human agents appear to be strongly influenced by time-varying situational factors. These can include random response errors (error related to variations in attention/mental efficiency/distraction across subjects), transient errors (errors caused by occasion-specific variations in mood/feeling/mindset within subjects) and specific errors (personal trait-related error, e.g., in terms of expertise, Schmidt and Hunter, 1999). A common work-around to these problems in prior room acoustical research has been to work with expert samples only and let subjects do only comparative stimulus ratings—which results in the methodological drawback of only allowing a limited number of participants and stimuli critically discussed in the beginning of this paper. The results of the present study, however, demonstrate that such work-arounds are not necessary when employing confirmatory factor analysis and structural equation modeling techniques, because these procedures are able to statistically correct very well for individual measurement errors, as demonstrated by the high stability coefficients of factor constructs[1] in comparison to low item stabilities (retest reliabilities). Furthermore, they also allow to address the large degree of zero inflation for items of usually "unwanted" percepts such as "echo" and "boominess," but also for items characterizing more abstract timbre features such as "roughness" and "comb filter coloration," which seem to be often interpreted in a yes-or-no manner. According to the results of the item analysis about one-third of the attributes proposed by the expert focus group can be considered as unsuitable for a measurement instrument that does *not* employ latent variable modeling techniques. The rest of the attributes still exhibited stability and distribution problems when administered to a sample of laymen with concert experience, but turned out to be more consistent with people with room acoustical education. The assumption that frequent concert experience and recording practice alone should suffice for reliable room acoustical assessments (Lokki *et al.*, 2012) is thus not supported by the present results.

The factor analysis of the remaining items suggests possible solutions with four, six, or nine factors. They can be interpreted as a general room acoustical *Quality* factor, *Strength*, *Reverberation*, *Brilliance* (four-factor solution), *Irregular Decay*, and *Coloration* (six-factor solution), *Clarity*, *Liveliness*, and *Intimacy* (nine-factor solution). The corresponding item batteries consist of 14, 20 and 29 attributes. From a statistical point of view, the 6-factor RAQI scale with 20 items is best suited to account for the full complexity of room acoustical impressions, while at the same time ensuring sufficient statistical independence of the different factors. *Quality* and its underlying items represent a subjective assessment of the perceived "fit" between room acoustical features and audio content taking into account the individual preferences of different listeners. Both effects are related to higher order, top-down aesthetical judgments, which cannot be explained by perceptual attributes alone. It is thus a matter of perspective, whether *Quality* is considered as an independent aspect of room acoustical perception in itself or as a kind of second-order factor that results from the other perceptual qualities. From this perspective, the structural equation model shows that, in our sample of room acoustical environments, subjective overall *Quality* judgements were positively correlated with the perceived *Strength* and *Brilliance* of the room, and negatively correlated with the perceived *Coloration* and *Irregular Decay*. These factors, however, explain only half of the variance in *Quality*.

With *Strength* and *Reverberance*, we identified two subdimensions that are omnipresent in the room acoustical literature. Also, *Clarity* and *Intimacy* as additional factors have been frequently highlighted by previous studies (Hawkes and Douglas, 1971; Lokki *et al.*, 2012). With *Brilliance*, *Coloration*, and *Intimacy* appearing as largely independent factors, it seems that timbre-related qualities play a greater role with more dimensions than previously assumed. The importance of perceived *Irregularities in Decay* and of *Liveliness* as an independent construct has, to our knowledge, hardly been considered so far. Comparing our nine-factor solution with the eight categories compiled by Kuusinen and Lokki based on their own and other studies, however, without further empirical analysis of their interdependence, shows four identical sub-dimensions: *Clarity*, *Reverberance*, *Loudness* (called *Strength* here), and *Intimacy*, while the other factors are different in structure (Kuusinen and Lokki, 2017).

In terms of psychometric quality of the six-factor RAQI, five factors exhibit excellent across-rater consistency (CR) and within-rater stability (SC). Only *Coloration* shows only fair across-rater consistency, while *Quality* shows only fair within-rater stability. With regard to measurement invariance, we were able to demonstrate scalar measurement invariance across measurement occasions with a rather long distance of approximately 42 days. Scores from all RAQI sub-dimensions can thus be directly compared across studies as long as experimental conditions and test subject sample are identical. Similar results pertain to changes in experimental listening position: Scores taken from slightly different listening positions in the same room, as in the present case, did not differ systematically. Although the acoustical transfer functions might be quite different, as was demonstrated even for minor changes of the listening position (de Vries *et al.*, 2001), listeners are obviously able to identify the room and its acoustical properties as a consistent

cognitive object. Results of measurement invariance tests for the influence of room acoustical expertise and the audio content (solo music, orchestra, speech) yielded overall metric but not scalar influence for all items. This implies that the overall factor structure of the RAQI holds very well across these strongly differing experimental conditions, which is a striking finding in itself. Since the observed bias for some items only affected the *size* of the mean scores and not the factor loadings, any form of covariance based analysis (correlations, regression, structural equation modeling) with RAQI factor scores will still be unbiased for different listening positions and different audio content. This includes music vs speech, which both appear to evoke similar perceptual impressions of the room acoustical environment, and listeners with different expertise. A direct comparison of mean score sizes, however, is only possible with identical stimuli and listeners with a comparable degree of expertise.

For the measurement of room acoustical impression, we thus suggest three options. For test-economic reasons one might decide (a) to draw on single-item measures only. In this case, we recommend using only those items of the RAQI scale yielding good retest reliabilities. To allow for the statistical control of measurement errors, however, we advise to draw on latent variable measurements, i.e., the measurement of factors rather than items, in particular when expecting a sample of moderate to low room-acoustical expertise participants. For this purpose, the scale versions with four, six, or nine factors can be used, depending on the desired degree of differentiation. When administering these scales, we advise to take special care explaining items with low retest-reliability (<0.7) during the initial instruction of subjects. For statistical analysis, one can (b) simply form weighted averages of the item scores on each dimension and draw on the weights provided in Table II. A better estimation of the factor scores can, however, be reached with (c) structural equation modeling software (AMOS, Lisrel, MPlus, R.lavaan) with a robust maximum likelihood estimator and weights and intercepts fixed to the values given in Table II. To estimate the scores of the two factors containing items with high zero-inflation probability, zero-inflated regression models should be used, or they could be treated as binary attributes right away. If structural equation modeling is used, the double loading of the "echo" item should be addressed by allowing double loadings. The resulting factor scores can be freely combined in covariance-based statistical analyses (regression, correlation, path analysis) with scores measured in other experiments with different samples, different excitement signals and from different listening positions. The direct comparison of score means (*t*-test, variance analysis), however, is only feasible with identical stimuli and participant samples of comparable expertise. We recommend employing a standardized measure for room-acoustical expertise to see if participant samples are truly comparable. This could, for example, be the number of years in a graduate program or job related to room acoustics, as we did in the present study. What exactly constitutes "room acoustical expertise" from a psychological point of view, however, seems to be an interesting question for future research.

With respect to the exhaustiveness of the presented factor solutions, future studies could consider complementing room acoustical attributes with good re-test reliabilities, which could not be assigned to any of the nine dimensions identified by EFA ("metallic tone color," "openness," "attack," "richness of sound") with additional attributes of similar meaning in order to possibly identify additional latent qualities not identified in the present study. It could also be considered to substitute items that turned out to be strongly expertise-variant ("treble range characteristic," "comb filter coloration," "size") by terms which are more easily accessible to lay persons. Using the RAQI scale in different languages will, in the future, allow to demonstrate the measurement invariance of the RAQI across the German and English language versions empirically.

Comparing the perceptual space identified by the current investigation with the "subjective listener aspects" named in ISO 3382-1 (2009), it is striking that timbral aspects, represented by the factors *Brilliance*, *Coloration*, and, to a certain extent, *Intimacy* in the nine-factor RAQI are currently not represented by standard room acoustical parameters at all. The fact that these play a major role in judgments on the acoustical qualities of performance venues for speech and music has already been pointed out by Lokki *et al.* (2016, p. 561) and coincides with similar findings for the quality of multichannel audio systems (Rumsey *et al.*, 2005). Spatial attributes, such as the "apparent source width" or "listener envelopment" in ISO 3382-1, on the other hand, do not appear as independent dimensions in any of the RAQI configurations. Although the related attributes can be quite reliably assessed, at least, by room acoustical experts, they are highly correlated to other aspects of the room acoustical impression: The perceived "width" is correlated to the perceived "loudness" and "size" of the source, as part of a general *Strength* factor, whereas "envelopment by reverberation" is part of the more general *Reverberance* factor. Even in our quite large sample of room acoustical environments, there were, obviously, no examples where sound sources appeared "wide" without sounding "strong," or where listeners felt "enveloped" by reverberation in rooms with low reverberance.

It might thus be worthwhile to consider the development of new room acoustical parameters that are more closely linked to the perceptual dimensions of room acoustical environments. The published database of all room acoustical environments, room impulse responses, and perceptual ratings used for the current study (Ackermann *et al.*, 2018) could be used as a ground truth to validate these new room acoustical parameters immediately.

ACKNOWLEDGMENTS

This investigation was supported by a grant from the Deutsche Forschungsgemeinschaft (Grant Nos. DFG FOR 1557, DFG WE 4057/3-2). The authors—as members of the focus group themselves—would like to thank for participation in the focus group (in alphabetical order): Jens Ahrens, Jens Blauert, Clemens Büttner, Eckhard Kahle, Markus Noisternig, Alexander Lindau, Sönke Pelzer, Zora Schärer, Ingo Witew, and Franz Zotter. For participation in

the translator group we would like to thank Jens Blauert, Eckhard Kahle, Jens Holger Rindel, and Martin Vercammen. For the recording of the speech and trumpet stimulus we would like to thank Boris Freytag and Tobias Haußig. For supervising many hours of listening tests we would like to thank Dmitry Grigoriev, Hannes Helmholz, and Omid Kokabi.

[1]See supplementary material at https://doi.org/10.1121/1.5051453. [SuppPub1] gives a complete list of the consensus vocabulary of the focus group; all attributes, which are part of the final measuring instrument, are listed in Table II. [SuppPub2] shows descriptive statistics of all rated items. [SuppPub3] shows the CFA measurement models for the four, six and nine factor solutions. [SuppPub4] shows the statistics for all tests on measurement invariance.

Ackermann, D., Ilse, M., Grigoriev, D., Lepa, S., Pelzer, S., Vorländer, M., and Weinzierl, S. (**2018**). "A ground truth on room acoustical analysis and perception (GRAP)," available at http://dx.doi.org/10.14279/depositonce-7003 (Last viewed August 20, 2018).

AES (**2015**). 69, *AES Standard for File Exchange—Spatial Acoustic Data File Format* (Audio Engineering Society, New York).

Bagenal, H. (**1925**). "Designing for musical tone," J. of the Royal Institute of Br. Archit. **32**(20), 625–629.

Barron, M. (**1988**). "Subjective study of British symphony concert halls," Acustica **66**, 1–14.

Beranek, L. L. (**1962**). *Music, Acoustics and Architecture* (Wiley, New York).

Beranek, L. (**1996**). *Concert and Opera Halls—How They Sound* (American Institute of Physics, Melville, NY).

Berg, J., and Rumsey, F. (**2006**). "Identification of quality attributes of spatial audio by repertory grid technique," J. Audio Eng. Soc. **54**(5), 365–379.

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-5718.2 (Last viewed August 20, 2018).

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), 1–8.

Cheung, G. W., and Rensvold, R. B. (**2002**). "Evaluating goodness-of-fit indexes for testing measurement invariance," Struct. Equation Model.: A Multidiscipl. J. **9**(2), 233–255.

Cho, E. (**2016**). "Making reliability reliable: A systematic approach to reliability coefficients," Organ. Res. Methods **19**(4), 651–682.

Cicchetti, D. V. (**1994**). "Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology," Psychol. Assess. **6**(4), 284–290.

Cronbach, L. J. (**1947**). "Test 'reliability': Its meaning and determination," Psychometrika **12**(1), 1–16.

de Vries, D., Hulsebos, E. M., and Baan, J. (**2001**). "Spatial fluctuations in measures for spaciousness," J. Acoust. Soc. Am. **110**(2), 947–954.

Erbes, V., Schultz, F., Lindau, A., and Weinzierl, S. (**2012**). "An extraaural headphone system for optimized binaural reproduction," in *Fortschritte der Akustik—DAGA 2012*, pp. 313–314.

Fabrigar, L. R., Wegener, D. T., MacCallum, R. C., and Strahan, E. J. (**1999**). "Evaluating the use of exploratory factor analysis in psychological research," Psychol. Methods **4**(3), 272–299.

Fornell, C., and Larcker, D. F. (**1981**). "Evaluating structural equation models with unobservable variables and measurement error," J. Market. Res. **18**(1), 39–50.

Gade, A. C. (**2010**). "Acoustics for symphony orchestras; status after three decades of experimental research," in *Proceedings of the International Symposium on Room Acoustics (ISRA 2010)*, Melbourne, Australia, pp. 1–12.

Geier, M., Ahrens, J., and Spors, S. (**2008**). "The SoundScape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *124th AES Convention*, Amsterdam, paper 7330.

Grigoriev, D. (**2017**). "Synthesis of binaural stimuli for a listening test on room acoustic perception," Master thesis, TU Berlin.

Hawkes, R. J., and Douglas, H. (**1971**). "Subjective acoustic experience in concert auditoria," Acustica **24**, 235–250.

Hu, L., and Bentler, P. M. (**1999**). "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," Struct. Equation Model.: A Multidiscip. J. **6**(1), 1–55.

ISO (**2009**). 3382-1, *Measurement of Room Acoustic Parameters Part 1: Performance Rooms* (International Organization for Standardization, Geneva, Switzerland).

Jullien, J.-P., Kahle, E., Winsberg, S., and Warusfel, O. (**1992**). "Some results on the objective and perceptual characterization of room acoustical quality in both laboratory and real environments," Proc. Inst. Acoust. **14**(2), 77–84.

Knudsen, V. O. (**1931**). "Acoustics of music rooms," J. Acoust. Soc. Am. **2**, 434–467.

Krueger, R. A. (**2014**). *Focus Groups: A Practical Guide for Applied Research* (Sage, Thousand Oaks, CA).

Kuusinen, A., and Lokki, T. (**2017**). "Wheel of concert hall scoustics," Acta Acust. Acust. **103**(2), 185–188.

Lehmann, P., and Wilkens, H. (**1980**). "Zusammenhang subjektiver Beurteilungen von Konzertsälen mit raumakustischen Kriterien" ("Relationship between subjective assessments of concert halls and room acoustical parameters"), Acustica **45**, 256–268.

Lindau, A., Erbes, V., Maempel, H.-J., Lepa, S., Brinkmann, F., and Weinzierl, S. (**2014**). "A spatial audio quality inventory for virtual acoustic environments (SAQI)," Acta Acust. Acust. **100**(5), 984–994.

Lindau, A., Kosanke, L., and Weinzierl, S. (**2012**). "Perceptual evaluation of model- and signal-based predictors of the mixing time in binaural room impulse responses," J. Audio Eng. Soc. **60**(11), 887–898.

Lindau, A., and Weinzierl, S. (**2009**). "On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical and lateral direction," in *EAA Symposium on Auralization*, Helsinki, pp. 1–6.

Loehlin, J. C. (**2004**). *Latent Variable Models: An Introduction to Factor, Path, and Structural Equation Analysis*, 4th ed. (Routledge, New York).

Lokki, T. (**2005**). "Subjective comparison of four concert halls based on binaural impulse responses," Acoust. Sci. Tech. **26**(2), 200–203.

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.

Lokki, T., Pätynen, J., Kuusinen, A., and Tervo, S. (**2016**). "Concert hall acoustics: Repertoire, listening position, and individual taste of the listeners influence the qualitative attributes and preferences," J. Acoust. Soc. Am. **140**(1), 551–562.

MacCallum, R. C., Widaman, K. F., Zhang, S., and Hong, S. (**1999**). "Sample size in factor analysis," Psychol. Methods **4**(1), 84–99.

Millsap, R. E. (**2011**). *Statistical Approaches to Measurement Invariance* (Routledge, New York).

Muthén, L. K., and Muthén, B. O. (**2012**). *Mplus User Guide. Statistical Analysis with Latent Variables*, 7th ed. (Muthén and Muthén, Los Angeles), available at http://www.statmodel.com/download/usersguide/Mplus%20user%20guide%20Ver_7_r3_web.pdf (Last viewed August 20, 2018).

Pätynen, J., Tervo, S., and Lokki, T. (**2011**). "Simulation of the violin section sound based on the analysis of orchestra performance," in *Proceedings of the IEEE workshop on Applications of Signal Processing to Audio and Acoustics (WASPAA)*, pp. 173–176.

Quiring, R., and Weinzierl, S. (**2016**). "Pitch distributions for individual instruments in the symphonies No. 19 of L.v. Beethoven," available at http://dx.doi.org/10.14279/depositonce-5040 (Last viewed August 20, 2018).

Rumsey, F., Zieliński, S., Kassier, R., and Bech, S. (**2005**). "On the relative importance of spatial and timbral fidelities in judgments of degraded multichannel audio quality," J. Acoust. Soc. Am. **118**(2), 968–976.

Sabine, P. E. (**1928**). "The acoustics of sound recording rooms," Trans. Soc. Motion Picture Eng. **12**(35), 809–822.

Sabine, W. C. (**1906**). "The accuracy of musical taste in regard to architectural acoustics," in *Proceedings of the American Academy of Arts and Sciences 42/2* (1906), pp. 53–58 [reprinted in Sabine, W. C. (1922). *Collected Papers on Acoustics* (ss.: Harvard University Press, Cambridge, MA, 1922), pp. 71–77].

Schmidt, F. L., and Hunter, J. E. (**1999**). "Theory testing and measurement error," Intelligence **27**(3), 183–198.

Shabtai, N. R., Behler, G., Vorländer, M., and Weinzierl, S. (**2017**). "Generation and analysis of an acoustic radiation pattern database for forty-one musical instruments," J. Acoust. Soc. Am. **141**(2), 1246–1256.

Somerville, T., and Gilford, C. L. S. (**1957**). "Acoustics of large orchestral studios and concert halls," Proc. Inst. Elec. Eng. **104**, 85–97 [reprinted in J. Audio Eng. Soc. **7**(4), 160–161 (1959)].
Steenkamp, J.-B. E. M., and Baumgartner, H. (**1998**). "Assessing measurement invariance in cross-national consumer research," J. Consum. Res. **25**(1), 78–90.
Stewart, D. W., Shamdasani, P. N., and Rook, D. W. (**1990**). *Focus Groups: Theory and Practice* (Sage, Newbury Park, CA).
Vigeant, M. C., Wang, L. M., Rindel, J. H., Christensen, C. L., and Gade, A. C. (**2010**). "Multi-channel orchestral anechoic recordings for auralizations," in *Proc. of the International Symposium on Room Acoustics (ISRA 2010)*, Melbourne.
Vooris, R., and Clavio, G. (**2017**). "Scale development," in *The International Encyclopedia of Communication Research Methods* (Wiley, New York).
Vorländer, M., Schröder, D., Pelzer, S., and Wefers, F. (**2014**). "Virtual reality for architectural acoustics," J. Build. Perform. Simul. **8**(1), 15–25.
Watson, F. R. (**1923**). *Acoustics of Buildings* (Wiley, New York).
Wefers, F. (**2010**). "A free, open-source software package for directional audio data," in *Fortschritte der Akustik—DAGA* 2012, pp. 1059–1060.
Weinzierl, S., Vorländer, M., Behler, G., Brinkmann, F., von Coler, H., Detzner, E., Krämer, J., Lindau, A., Pollow, M., Schulz, F., and Shabtai, N. R. (**2017**). "A database of anechoic microphone array measurements of musical instruments," available at http://dx.doi.org/10.14279/depositonce-5861.2 (Last viewed August 20, 2018).
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.