A round robin on room acoustical simulation and auralization

Fabian Brinkmann, Lukas Aspöck, David Ackermann, Steffen Lepa, Michael Vorländer, and Stefan Weinzierl

---

ARTICLES YOU MAY BE INTERESTED IN

Introduction to the Special Issue on Room Acoustic Modeling and Auralization
The Journal of the Acoustical Society of America **145**, 2597 (2019); https://doi.org/10.1121/1.5099017

An investigation of listener envelopment utilizing a spherical microphone array and third-order ambisonics reproduction
The Journal of the Acoustical Society of America **145**, 2795 (2019); https://doi.org/10.1121/1.5096161

Machine-learning-based estimation and rendering of scattering in virtual reality
The Journal of the Acoustical Society of America **145**, 2664 (2019); https://doi.org/10.1121/1.5095875

Perceptual evaluation of headphone auralization of rooms captured with spherical microphone arrays with respect to spaciousness and timbre
The Journal of the Acoustical Society of America **145**, 2783 (2019); https://doi.org/10.1121/1.5096164

Passive volumetric time domain simulation for room acoustics applications
The Journal of the Acoustical Society of America **145**, 2613 (2019); https://doi.org/10.1121/1.5095876

A framework for auralization of boundary element method simulations including source and receiver directivity
The Journal of the Acoustical Society of America **145**, 2625 (2019); https://doi.org/10.1121/1.5096171

A round robin on room acoustical simulation and auralization

Fabian Brinkmann,[1,a)] Lukas Aspöck,[2] David Ackermann,[1] Steffen Lepa,[1] Michael Vorländer,[2] and Stefan Weinzierl[1]

[1]*Audio Communication Group, Technical University of Berlin, Einsteinufer 17c, Berlin, D-10587, Germany*
[2]*Institute of Technical Acoustics, Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, Kopernikusstraße 5, Aachen, D-52074, Germany*

A round robin was conducted to evaluate the state of the art of room acoustic modeling software both in the physical and perceptual realms. The test was based on six acoustic scenes highlighting specific acoustic phenomena and for three complex, "real-world" spatial environments. The results demonstrate that most present simulation algorithms generate obvious model errors once the assumptions of geometrical acoustics are no longer met. As a consequence, they are neither able to provide a reliable pattern of early reflections nor do they provide a reliable prediction of room acoustic parameters outside a medium frequency range. In the perceptual domain, the algorithms under test could generate mostly plausible but not authentic auralizations, i.e., the difference between simulated and measured impulse responses of the same scene was always clearly audible. Most relevant for this perceptual difference are deviations in tone color and source position between measurement and simulation, which to a large extent can be traced back to the simplified use of random incidence absorption and scattering coefficients and shortcomings in the simulation of early reflections due to the missing or insufficient modeling of diffraction.

I. INTRODUCTION

Room acoustical simulation shows an increasing number of applications covering not only the classical tasks of acoustical and electro-acoustical planning,[1] but also fields such as architectural history,[2,3] music research,[4] game audio,[5] or virtual acoustic reality in general.[6] This applies to wave-based simulations as well as to simulations based on geometrical acoustics (GA)[7] or hybrid approaches.[8] Many of these applications make use of the possibility to generate binaural signals based on including head-related transfer functions (HRTFs) into the numerical signal chain, a process which was coined auralization.[9] At the same time, there is no undivided confidence in the reliability of room acoustical simulations when it comes, for example, to the design of new performance venues for music and speech, where acoustic scale models are still an important tool with specific advantages.[10] The application and further development of room acoustic simulations is thus crucially dependent on the availability of a procedure to objectively assess the accuracy of these applications—the more so since background theories such as GA make obvious simplifications, which are valid only in a limited frequency range.

There have been different attempts to validate the mentioned modeling approaches and related software implementations. Two databases with analytically defined test scenarios were established by Otsuru *et al.*[11] and Hornikx *et al.*[12] They are intended for cross validation of wave-based simulation algorithms and not dependent on measured reference data. This is an approach that guarantees a perfect reference, but a viable option only for very simple scenes for which analytic solutions are available.

In three round robin experiments conducted between 1994 and 2002 (RR-I to RR-III)[13–16] the results of different room acoustical simulation algorithms were compared to measurements of a smaller lecture hall (RR-I), a multipurpose hall (RR-II), and a music studio (RR-III). In this series of tests, different information was provided to the participants at different phases. In phase I of RR-I and RR-II, the participants had to estimate the geometry and boundary conditions themselves from architectural plans and written information ("3 mm carpet"); in phase II, the data were harmonized based on a common three-dimensional (3D) model and boundary conditions estimated by room acoustical measurements. In RR-III, absorption and scattering coefficients for one wall and the ceiling of the room were measured in the reverberation room, and taken from tabulated data otherwise. Measured and simulated room impulse responses (IRs) were compared based on room acoustical parameters, i.e., audio features extracted from energy decay representations, such as the early decay time (EDT) and other parameters suggested in ISO 3382-1.[17]

The biggest challenge in working with measured references in these tests has been to guarantee an exact match of the measured situation and input parameters of the numerical model. This applies to the geometric model of the acoustic scene, the behaviour of the sources and receivers as an integral part of the acoustic transfer path, and—above all—the

a)Electronic mail: fabian.brinkmann@tu-berlin.de

acoustic boundary conditions. For complex rooms such as concert venues or lecture halls, a comprehensive specification of absorption and scattering for all boundaries is practically impossible because neither can all different surfaces with their different types of installation be measured in the laboratory nor are any (standardized) full-range measurement techniques available to determine them *in situ*.[18] Fitting the input parameters according to measurements of the reverberation time, on the other hand, may be a pragmatic and often applied solution in room acoustic planning. As a procedure for the evaluation of numerical simulations, however, it would contain an element of circular reasoning if both the premises (the boundary conditions) and the success of the simulation were determined by the measurement of the same room acoustical parameters, or by ones that strongly correlate with each other. Hence, although RR-I to RR-III imitated a "real-world" acoustical planning scenario and gave an impression of how reliable different room acoustics simulation softwares are as planning tools, they could hardly give concrete insights into the strengths and weaknesses of the algorithms themselves.

A reliable reference for room acoustical simulations with respect to geometry and boundary conditions can only be provided if the scene is sufficiently simple so that the relevant measuring methods do not reach their limits. This approach was followed by Tsingos *et al.* when setting up the "Bell Labs Box," i.e., a 16 m³ rectangular enclosure with one baffle inside, in order to compare measured and simulated IRs and validate a proprietary simulation algorithm.[19] The planned extension of the test system toward different and more complex configurations and an evaluation of different numerical simulations, however, has not yet taken place.

The round robin on room acoustical simulation and auralization presented here represents a combination of both approaches and extends them to an evaluation in the physical *and* perceptual realm. The test was based on a database of measured IRs established for this purpose.[20] It contains 3D room models, source and receiver directivities, and one-third octave absorption and scattering coefficients for 11 acoustic scenes (Table I). Eight of these scenes are simple configurations for which all parameters could be measured in the laboratory with high precision. They were designed to isolate specific acoustical phenomena such as single and multiple reflections on finite and infinite plates, scattering, diffraction, the seat dip effect, or a coupled room. Three of the scenes are complex, real-world rooms similar those used in RR-I to RR-III for which only a best possible, practical estimate of the parameters could be given. A selection of these scene descriptions was provided to developers of room acoustic simulation software who were given six months to simulate IRs based on the provided data.

To evaluate the results of the numerical simulations in the physical domain, measured and simulated IRs were compared based on temporal and spectral features. Moreover, dynamic auralizations of the simulated scenes based on binaural room impulse responses (BRIRs) were evaluated against their measured counterparts with the simulation using HRTFs corresponding to the binaural receiver, which

TABLE I. Overview of the 11 scenes contained in the database: Scenes 1–8 are designed scenarios to isolate acoustical phenomena and scenes 9–11 are representative room acoustic scenarios. Most scenes include multiple source/receiver positions and configurations (e.g., different surfaces materials). The column "Algorithms" shows the number of participating teams in the physical/perceptual evaluations of the round robin. Gray entries were not considered in the round robin.

Number	Scene	Algorithms
1	Single reflection (infinite plate)	5/-
2	Single reflection (finite plate)	4/-
3	Multiple reflections (parallel finite plates)	5/-
4	Single reflection (reflector array)	3/-
5	Diffraction (infinite wedge)	3/-
6	Diffraction (finite body)	-/-
7	Multiple diffraction (seat dip effect)	-/-
8	Coupled rooms	6/-
9	Small room (seminar room)	6/4
10	Medium room (chamber music hall)	6/4
11	Large room (auditorium)	6/4

was also used for the measurements. The listening test yielded measures for the plausibility and authenticity of the simulation, as well as difference ratings for a selection of specific perceptual qualities.

II. METHOD

A. Scene descriptions

For the round robin, 9 of the 11 scenes of the database were selected (Table I), each of which was supposed to be simulated with different settings (boundary conditions, source and receiver positions). A short overview of these configurations will be given below, while a more comprehensive description, including all scene configurations, is available in the supplemental material[21] and documentation of the database itself.[20]

Scene 1 realizes a single reflection on quasi infinite rigid, absorbing, and diffusing baffles for incident and exit angles of 30°, 45°, and 60°. Scene 2 is a single reflection on a finite quadratic plate with edge lengths of 1 m and 2 m, and incident/exit angles of 30°, 45°, and 60°. Receiver positions behind the plate are included to assess diffraction around the reflector. Scene 3 constitutes a flutter echo between two finite reflectors with edge lengths of 2 m and a single source-receiver configuration. Scene 4 realizes a single reflection on an array of nine reflectors with edge lengths of 68 cm (spaced 13 cm apart) for incident and exit angles of 30°, 45°, and 60°, as well as a reflection point on the center of a reflector and a reflection point between four reflectors. Scene 5 features the diffraction around a quasi infinite wedge (4.75 m × 2.07 m) for four different source and receiver heights below and above the upper edge of the wedge. Scene 8 establishes the double sloped energy decay of a reverberation chamber coupled to a laboratory room. Different degrees of coupling were realized by two opening angles of the connecting door (4.1° and 30.4°) and source positions inside both rooms. Scenes 9–11 are complex real-life environments of different size where omnidirectional source and receiver configurations according to ISO 3382-1[17] were

J. Acoust. Soc. Am. **145** (4), April 2019

Brinkmann *et al.*   2747

included, as well as a binaural receiver and directional sources.

In all scenes, IRs were measured with a Genelec 8020c studio monitor and 1/2 in. pressure microphones (G.R.A.S. 40AF, Bruel and Kjær type 4134). BRIRs were measured with QSC-K8 PA speakers and the FABIAN head and torso simulator[22] (HATS). The QSC-K8 speakers were chosen due to their higher sound power enabling auralizations with a higher signal-to-noise ratio; the FABIAN HATS was chosen due to the ability to automatically measure BRIRs for different head orientations. For the complex rooms (scenes 9–11), IRs were additionally measured with an ISO-3382-1[17] compliant dodecahedron speaker for the ISO-compliant analysis of room acoustical parameters.

Five sources were used for the binaural measurements of which four were arranged in a semicircular setup to mimic the positions of a string quartet, and the fifths source was placed in the center of the virtual quartet to mimic the position of a singer. The receiver was placed at a distance of 3–4 times the critical distance to emphasize the influence of the room (cf. Table II). BRIRs were measured for head-above-torso orientations to the left and right in the range of $\pm 44°$ with a resolution of $2°$, allowing for a perceptually transparent switching of BRIRs for different head orientations[23] within the typical range of motion.[24]

B. Simulation algorithms

Six teams using five different simulation algorithms participated in the round robin: BRASS (Brazilian Room Acoustic Simulation Software) is a ray tracing algorithm developed in the academic environment, which clusters reflections up to fifth order to provide accurate early reflections without deploying an image source model.[25] EASE V4.4 is a commercial tool for the simulation of room acoustical and electro-acoustical environments, which uses image sources for the direct sound and early reflections and ray tracing for the late reverberation.[26] ODEON combined 14 is a commercial tool for room acoustical simulation based on a hybrid ray tracing approach for detecting early specular reflections and calculating late reverberation.[27] RAVEN (Room Acoustics for Virtual Environments) is a hybrid algorithm developed in the academic environment that uses image sources for the direct sound and early reflections, as well as ray tracing for the late reverberation.[28] RAZR is an open source academic algorithm for the simulation of rectangular rooms through a combination of image sources and a feedback delay network for late reverberation.[29]

TABLE II. Selected properties of the small, medium, and large room: Approximate volume V and reverberation time T_m (averaged across 500 Hz and 1 kHz octaves), as well as the corresponding Schroeder frequency $f_s = 2000\sqrt{T/V}$ and critical distance $d_c \approx 0.057\sqrt{V/T}$ for each room. The distance from the binaural receiver to the center QSC-K8 speaker is given by d.

	V/m^3	T_m/s	f_s/Hz	d_c/m	d/m
Small	145	2.0	234	0.49	4.00
Medium	2,350	1.3	47	2.42	9.95
Large	8,650	2.1	31	3.66	11.33

All algorithms consider frequency dependent absorption and scattering coefficients, air absorption, and arbitrary receiver and source directivities—with the exception of RAZR, which assumes omnidirectional sources and does not account for scattering. The simulation of diffraction is only implemented in ODEON, and is only activated in case of a blocked direct sound path by estimating diffraction paths around objects. ODEON is also the only algorithm that considers the energy loss of specular reflections caused by diffraction around finite objects by adjusting the scattering coefficient depending on the incident angle and size of the reflecting surface (cf. Ref. 27, pp. 79 and 83). Moreover, ODEON takes into account angle dependent absorption by modifying random incidence coefficients based on the mid-range absorption between 1 and 4 kHz and idealized absorber models (cf. Ref. 27, p. 74).

All simulations were carried out in the groups or companies of the software developers themselves. A second contribution using ODEON 12 came from the Department of Industrial Engineering, University of Bologna (V12). Please note that RAVEN is developed at Rheinisch-Westfälische Technische Hochschule (RWTH) Aachen University, which was also involved in acquiring the acoustic scenes that were used in the round robin. However, neither did RAVEN play a role in the generation of the reference database nor were results from RAVEN adjusted according to measured data. To avoid a bias, the RAVEN simulations were conducted by a person who was not aware of the measurement results nor was he involved in the round robin otherwise.

More teams showed interest in contributing to the round robin. Developers of wave-based algorithms, however, were not ready to provide results for the entire audible bandwidth, and commercial algorithms sometimes missed an interface for including other than the stock HRTFs in their simulations. RAZR was allowed to participate despite the high degree of simplification of the underlying algorithm due to the open nature of the call that initiated the round robin. In retrospect, the results turned out to be particularly interesting because they were in some properties well comparable to results from the remaining algorithms.

In the following, the terms *algorithm* and *software* will refer to the combination of the actual simulation software and the people that used it to simulate the IRs.

C. Task and data processing

The participants were instructed to simulate IRs without changing the source and receiver directivities or the surface properties (absorption, scattering). For the simple scenes used (scenes 1–5), the boundary conditions could be reliably determined by laboratory measurements, so no modification would be reasonable. For the complex scenes (scenes 8–11), the measured absorption and scattering coefficients can only be considered as best possible estimations, so the simulation could probably have been improved by fitting the boundary conditions according to the measured results of room acoustical parameters. In this case, the task of the round robin corresponds to the predictive situation of a new room acoustic design where no such measurements are available. To ensure this, the measured IRs were not available to the participants

at the time of the simulations. In contrast, the room geometry provided by the 3D model could be simplified if required by the specific simulation algorithm since the authors consider such a pre-processing, which is always necessary for high resolution architectural models, as part of the simulation itself.

The source directivities of the Genelec 8020c and QSC-K8 were provided by means of IRs and third octave spectra on a 1×1 equal angle sampling grid. Head-related impulse responses (HRIRs) of the head and torso simulator that was also used for measuring BRIRs were obtained on a 1×1 equal angle sampling grid from the FABIAN database.[30,31] The frequency response of the sources is contained in the measured IRs and BRIRs and the corresponding directivities, while the frequency responses of FABIAN's DPA 4060 microphones are removed from the HRIRs provided and from all measured BRIRs.

The software teams reported that they used the provided directivities without changes, with the exception of RAZR that used omnidirectional sources with only the on-axis frequency response taken from the provided data. The ODEON contribution of the developers' group used a spatial resolution of $10°$ for the source directivities and $3°$ for the binaural receiver. Both ODEON contributions converted the directivity information to octave values and restricted the range to center frequencies between 64 Hz and 8 kHz. The contributions from ODEON and RAZR also converted the absorption coefficients to octave values in the ranges from 63 Hz to 8 kHz and 250 Hz to 4 kHz, respectively. While RAZR neglected the provided scattering coefficients, the ODEON teams obtained a nominal mid-frequency scattering coefficient by averaging between 400 Hz and 1.25 kHz (developers) and by using the provided values at 800 Hz (University of Bologna). None of the teams reported to have changed the 3D models, with the exception of RAZR, which generated rectangular rooms with equivalent volumes maintaining the ratios of the main room dimensions. Most of the teams used a transition order of two and three to combine early and late reflections, with the exception of BRASS, which clustered ray traced reflections up to order five to obtain image-source-like components, and the ODEON contribution from the University of Bologna that used a transition order of ten.

D. Physical evaluation

For the physical evaluation, IRs for the omnidirectional receiver were processed in MATLAB using methods from the open source project ITA-Toolbox.[32] The measured and simulated IRs were temporally aligned, normalized to the root mean square of the IR, and truncated to a length of 46 ms using a two-sided Hann window (3 ms fade in, 10 ms fade out). In case of scene 8, IRs were truncated to 2.2 s with a 50 ms fade out, and the energy decay curve (EDC) was calculated for the 1 kHz octave band.

For scenes 9–11, room acoustical parameters were calculated according to ISO 3382-1[17] based on IRs measured with the dodecahedral loudspeaker and omnidirectional microphone for two source and five receiver positions. The parameters for the measured and simulated IRs were calculated using the *ita_roomacoustics* routine. For the simulated RIRs, the calculation of the EDC was based on the entire RIR, while the measured RIRs were truncated after detecting the noise floor according to ISO 3382-1.[33] For the sake of brevity, only T_{20} results are presented as averages over all source/receiver combinations.

Measured and simulated BRIRs were further analyzed with respect to differences in perceived tone color. This was assessed by means of energetic differences in 37 auditory filter bands between 80 Hz and 16 kHz using the gammatone filterbank from the auditory toolbox.[34]

E. Auralization

Dynamic auralizations, considering head movements of the listeners in a horizontal range of $\pm44°$, were obtained by dynamic convolution of the measured and simulated BRIRs with anechoic audio content: The BRIRs for all head orientations and sources were stored in SOFA files[35] and loaded by a customized version of the Sound Scape Renderer[36] (SSR) used for convolution. The BRIRs were selected according to the current head orientation of the listener as provided by a Polhemus Patriot head tracker (precision $0.003°$). Pure Data[37] was used to start and stop the anechoic audio content according to open sound control messages triggered via MATLAB-based user interfaces. Pure Data and the SSR ran on a Linux-based desktop computer where the audio routing was done by the Jack Audio Connection Kit (JACK). The user interfaces ran on a separate laptop computer with Windows. The setup made it possible to switch between auralizations rendered from measured BRIRs, and BRIRs from different acoustic simulation algorithms at any time, whereby the audio content was restarted. For playback, Sennheiser HD 800 headphones were used at a playback level of 70 dB(A) (measured with pink noise). To minimize the influence of the headphone, a compensation filter was designed using regularized inversion.[38]

Because the BRIRs differed in level across algorithms and compared to the measured data, they had to be normalized. The gain for normalization was obtained by averaging the logarithmic magnitude response of the binaural transfer functions (center source and neutral head orientation of FABIAN) between 200 Hz and 1 kHz and across the left and right ears. One gain value was applied to all BRIRs of each algorithm, assuming that the algorithms preserved the level difference between the sources and ears, which was confirmed by an analysis of the level across source positions. Afterward, the authors made manual adjustments in the range of ±0.5 dB to optimize the loudness matching between algorithms and across measured and simulated data by means of informal listening (cf. audio examples provided in Sec. III B 3).

F. Perceptual evaluation

The perceptual evaluation was done based on two measures for the overall perceived difference between measurement and simulation (*authenticity*[39] and *plausibility*) and a differential diagnosis using the *Spatial Audio Quality Inventory* (SAQI), a qualitative test including 48 perceptual

J. Acoust. Soc. Am. **145** (4), April 2019

Brinkmann *et al.* 2749

qualities relevant for the quality of virtual acoustic environments.[40]

A measure for *authenticity*, indicating the existence of any audible difference between measurement and simulation,[39] was obtained by implementing a two interval, two alternative forced choice test (2I/2AFC) as a double-blind and criterion-free procedure. On a user interface with three buttons A, B, and X, the subjects were asked "Does X equal A of B?" Reference and simulation were randomly assigned to the buttons, and the participants could listen to A, B, and X in any order and as often as they wanted before making their choice.

To analyze the significance of the results, the type I error level (concluding that there is a difference although there is none) and the type II error level (concluding that there is no difference although there is one) were both set to 0.05. Since testing for authenticity requires proving the null hypothesis (no audible difference), which is not possible with inferential statistics, a minimum-effect test was conducted based on a practically meaningful detection rate of 0.9.[41] Hence, the alternative hypothesis to be rejected for assuming authenticity was $H_1(p_{2AFC} \geq 0.9)$, and the null hypothesis (no difference) was $H_0(p_{2AFC} = 0.5)$ with 0.5 as the two alternative forced choice test (2AFC) guessing probability. According to the desired error levels and effect size, $N = 13$ trials had to be conducted per participant[42] with authenticity to be assumed in the case of less than $N_{min} = 10$ correct answers. Note that $N_{crit.} = 10$ refers to a detection rate of about 75%, which equals a guessing probability of 50% and is the definition of the just noticeable difference (JND).

A looped pink noise pulse between 100 Hz and 20 kHz and a duration of 1 s (20 ms squared sine ramps) followed by 1.5 s silence was used as audio content due to its high potential to reveal possible flaws of the simulations that are related to timbral and spatial perceptions. The bandwidth was chosen according to the operating range of the measurement equipment and frequency range where absorption and scattering coefficients were provided. The pulse was auralized by the rightmost source as viewed from the direction of the binaural receiver (position of the cello in the virtual string quartet).

As a somewhat less strict criterion *plausibility* was determined, indicating whether BRIRs can be identified as "simulated" according to artefacts in the stimulus itself, i.e., without immediate comparison to an external reference. The test was implemented as a yes–no task. After each presentation, participants were asked "Was this an audio example from a real room?", and the answers were analyzed with signal detection theory (SDT).[43] This allows to obtain a criterion-free measure for the sensory difference d' between auralizations based on measured and simulated BRIRs, with $d' = 0$ indicating that differences were inaudible and $d' > 0$ indicating that differences are audible. The sensory difference can be converted to the easier to interpret 2AFC detection rate by $p_{2AFC} = \Phi(d'/\sqrt{2})$, where $\Phi(\cdot)$ is the cumulative standard normal distribution.

In analogy to authenticity, plausibility was tested separately for each participant. To analyze the significance of the

results, the type I error level (wrongly concluding that a simulation is *not* plausible) and the type II error level (wrongly concluding that a simulation is plausible) were again balanced and set to 0.05. According to the desired error levels, the meaningful d'_{min} to be rejected in a minimum-effect test,[41] is $d'_{min} = 0.82$ [cf. Eq. (13) in Lindau and Weinzierl[43]]. It corresponds to a 2AFC detection rate of $p_{2AFC} = 0.72$, which is similar to the critical value of the test for authenticity.

For the test, auralizations of 3–5 s duration were presented to the participants. The presentation order was randomized, and participants did not know whether an auralization was based on measured or simulated BRIRs, but were informed that the test conditions were approximately evenly distributed across $N = 100$ test trials (5 source positions × 20 audio contents). To avoid possible familiarization, 20 different monophonic audio contents were used exactly once with each of the 5 sources. These included an artificial noise signal, female/male speech and singing in different languages, solo instrument recordings, and excerpts of different pop songs. A visual impression of the room was provided by a 55 in. curved screen with a two picture slide show. One picture showed the entire room with an empty stage, and one was taken from the virtual listening position with loudspeakers on the stage (cf. SuppPub3,[21] Fig. S3–34).

In addition to the two overall measures for the perceived difference between measurement and simulation, ten perceptual qualities from the SAQI were selected based on informal prior listening according to their relevance and with an eye on completeness. The selection covers sound source related aspects (*source position*, *source extension*, *distance*, *localizability*), coloration (*tone color bright/dark*), the response of the acoustic environment (*duration of reverberation*, *envelopment by reverberation*), the temporal behaviour (*crispness*), and also includes the holistic measures *difference* and *clarity*. Some of the original SAQI items were combined to limit the duration of the listening test, such as *source position*, condensed from *horizontal* and *vertical direction*, and *source extension*, condensed from *depth*, *width*, and *height*. The participants received written circumscriptions (from Lindau *et al.*[40]) and oral explanations of the qualities before the test started.

Two types of audio content were selected for SAQI testing: The pink noise pulse already used for testing authenticity was believed to best reveal artifacts for most selected qualities, and an anechoic recording of Mozart's string quartet No. 1 (bars 1–6) was taken as typical real-life content. The four tracks of the string quartet recording were assigned to the four sources arranged on stage in a semi-circular setup, and the noise pulse was played only by the rightmost source of the virtual string quartet.

Auralizations based on simulated BRIRs were compared to their measured counterparts in an interface with four continuous sliders. The scale labels were displayed above and below the sliders. Two buttons positioned below each slider, labeled A, B, were used to start the auralizations with A starting the reference and the four simulations randomly assigned to four B buttons. While the audio content was held constant for each rating screen, the qualities to be rated were

2750   J. Acoust. Soc. Am. **145** (4), April 2019

Brinkmann *et al.*

presented in randomized order. The participants could listen as long as they needed, and switch between the four conditions on each rating screen.

Twenty-nine participants (8 female, 21 male, mean age 34 years) took part in the listening test. Twenty-four participants had already done listening tests before, 14 were experienced with room acoustical simulation, and 11 were experienced with binaural synthesis. On average, the subjects were concerned 2 h per day with listening, playing, or working with audio. After the participants had been informed about the purpose of the experiment, the test for plausibility was conducted first, followed by the test for authenticity and the SAQI. The order of the three tests was identical for all participants because previous exposure to the test environment should generally be avoided concerning the plausibility measure.[43]

The plausibility and authenticity tests employed the medium size room only since informal prior listening had shown that the overall quality of each algorithm did not differ substantially among the three acoustic environments. For these two tests, each participant evaluated only one randomly assigned simulation algorithm, i.e., each algorithm was tested by 7 subjects (the 29th subject was discarded in this case). Each subject was presented the whole set of rooms and algorithms with varying audio content during the SAQI test.

Each test included a separate training to familiarize the participants with the interface, stimuli, and test procedure. Subjects were encouraged to move their heads and compare the auralizations at different head orientations, as this might provide additional cues. The entire test took 90 min on average, including general instructions, training, and short breaks between the three sections. Throughout the session, the experimenter was sitting behind a screen, not visible to the participant to avoid potential distractions. The test was conducted in a quiet environment with a reverberation time of $T_m = 0.77$ s.

III. RESULTS

In two sections, exemplary results for the comparison of measurements and room acoustic simulations are shown both for the simple scenes (scene 1–8), highlighting the modeling of specific acoustical phenomena, and for the complex scenes (scenes 9–11), highlighting the performance of room acoustical simulation and auralization software in real-world situations. The results are anonymized, with letters A to F assigned to the participating simulation algorithms. Only a selection of the results are discussed, while a comprehensive overview of all results, including the exact source and receiver positions for every scene is given in the supplemental material (SuppPub1–3).[21] Since some software teams contributed to selected cases only, the number of participants differs from scene to scene.

A. Simple scenes

1. Specular reflections

Modeling a specular reflection is a simple task for an algorithm based on GA, in which case the addition of reflected energy to the direct sound results in a comb filter-like magnitude spectrum. Figure 1(a) shows the results for a reflection on a quasi infinite rigid surface (scene 1, floor of the hemi anechoic chamber) for incident and exit angles of $\gamma = 45°$. The line of sight distance between source and receiver was 4.2 m, and the source/receiver were 3 m away from the point of reflection. The comb filter effect is visible for all algorithms with small differences in the frequencies of notches and peaks due to minor deviations in the positioning of the sources/receivers between measurements and simulations. When the rigid surface is replaced by an absorber, results show that for all algorithms the comb filter effect becomes weaker for higher frequencies due to the increasing absorption (cf. SuppPub2,[21] Figs. S2-3 and S2-4).

In scene 2, a reflection on a finite medium density fibreboard plate with an edge length of 1 m and 25 mm thickness was measured. Figure 1(b) shows results in the frequency domain for incident and exit angles of $\gamma = 45°$. The distance between source and receiver was 5.7 m, and the source/receiver were 4 m away from the point of reflection. Due to the limited size of the reflector, most of the energy below $\sim$300 Hz is diffracted around the plate and the comb filter is less pronounced in this case. This was only correctly modeled by C, which includes a first-order edge diffraction model, whereas the remaining algorithms show a pronounced but "wrong" comb filter effect also for low frequencies and a largely correct simulation only for frequencies above 600 Hz. Results of the reflection on an array (scene 4,
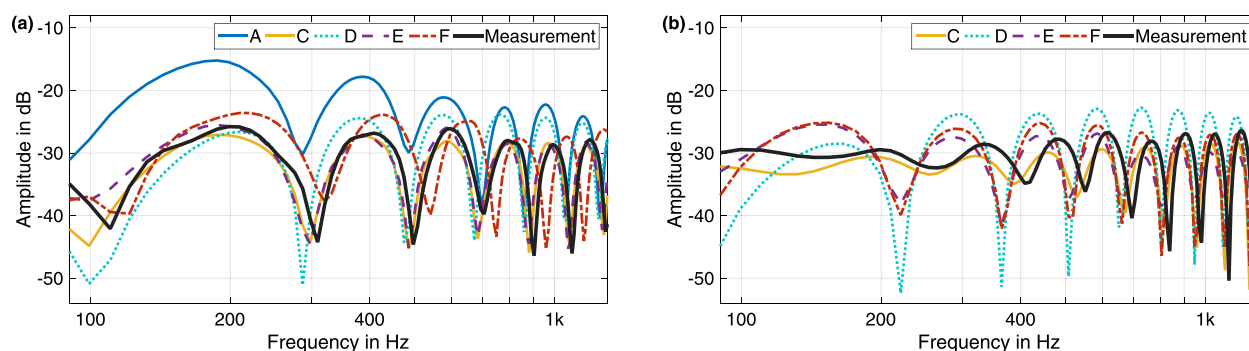

FIG. 1. (Color online) Specular reflections: Magnitude spectra of measured and simulated IRs for the reflection on a quasi infinite (a) and finite rigid plate (b). Both cases are for incidence/exit angles of $\gamma = 45°$ [scene 1 and 2; source position LS02; receiver position MP02; cf. SuppPub1 (Ref. 21), Figs. S1-1/2 and S1-7/8 for scene geometry].
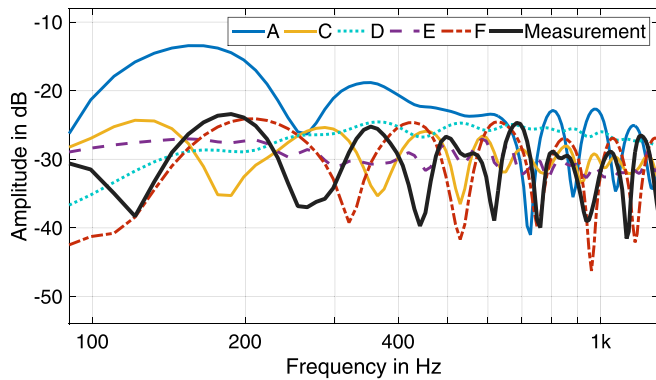
FIG. 2. (Color online) Diffuse reflections: Magnitude spectra of measured and simulated IRs for the reflection on a one-dimensional diffusor and incidence/exit angles of $\gamma = 45°$ [scene 1; source position LS02; receiver position MP02; cf. SuppPub1 (Ref. 21), Fig. S1-20/21 for scene geometry].

cf. SuppPub1,[21] Fig. S1-17) show that for complex reflector structures, even more substantial deviations from the measurement can be observed in the frequency domain for all software, in particular for *D*, which failed to include a reflection at all in case of the *off center* setup (cf. SuppPub2,[21] Figs. S2-23, S2-25, S2-27). In both situations shown in Fig. 1, software *D* shows a slightly distorted spectral shape in favor of high frequencies, whereas software *A* shows an emphasis of low frequencies for the infinite plate.

2. Diffuse reflection

To investigate in how far the algorithms can handle diffuse reflections, a one-dimensional diffusor consisting of periodically arranged wooden beams was placed on the floor of the hemi anechoic chamber (scene 1). In contrast to all other scenes, no scattering data were provided in this case. Instead, the participants were asked to model the scattering according to the demands of their software, which they did by using the geometrical diffusor model rather than assigning scattering coefficients derived from the provided dimensions of the diffusor to a single surface. Results for incident and exit angles of $\gamma = 45°$ are given in Fig. 2. The distance between source and receiver was 4.2 m, and the source and receiver were 3 m away from the point of reflection. No participant was able to match the measured frequency response, which can be described by an irregular comb filter. Software

C and *F* result in a frequency response similar to the measurement, but an inaccurate temporal modeling of the diffuse reflections (cf. SuppPub2,[21] Fig. S2-9) leads to a misalignment of peaks and notches in the frequency response.

3. Diffraction

In case the direct sound path is close to objects or edges, diffraction adds energy to the direct sound. This causes a temporal broadening of the main impulse and/or an isolated reflection, which leads to a weak and irregular comb filter structure in the magnitude spectrum. This can be observed in the measurement depicted in Fig. 3(a) for a source 4 m in front and a receiver 1 m behind a medium density fibreboard panel with edge lengths of 1 m (scene 2). The source is visible from the receiver, and the direct sound path has a distance of 0.7 m to the reflector panel. While simulations from *D*, *E*, and *F* show a small extent of irregularity, *C* disregards this effect completely. Although all algorithms come relatively close the measurement, as the influence of the diffraction wave in the illuminated region is small, slightly audible coloration artifacts can be expected.

If the source is not visible to the receiver, i.e., if the direct sound path is blocked by an object, significant energetic contributions come from diffraction around objects and/or transmission through objects. Figure 3(b) shows results for the diffraction around a quasi infinite medium density fibreboard wedge with a height of 2.07 m and a thickness of 25 mm (scene 5). The source and receiver were positioned 3 m in front and behind the partition at a height of 1.23 m. Because only two participants were able to simulate first-order diffraction, no letters are assigned to the simulation results in order to keep the anonymity of the participants. The results show that both programs are able to match the general trend of the measured curve where the diffracted energy arriving at the receiver decreases with increasing frequency. Apparently, the reflections on the rigid floor in front of and behind the partition, which create the comb filter structure in the measured frequency response, are not modeled. When comparing the two simulation results, a similar result can be observed for frequencies above 250 Hz while the curves substantially deviate for frequencies lower frequencies, reaching a difference of more than 10 dB for 100 Hz.
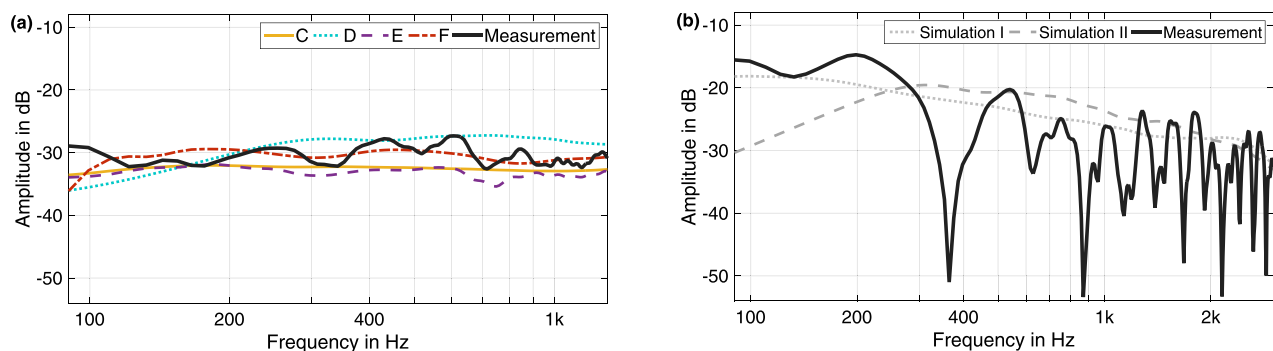

FIG. 3. (Color online) Diffraction: Magnitude spectra of measured and simulated IRs for grazing sound incidence at a finite rigid plate (a) [scene 2; source position LS05; receiver position MP04; cf. SuppPub1 (Ref. 21), Fig. S1-7/8 for scene geometry] and diffraction on a quasi infinite wedge (b) [scene 5, source position LS01; receiver position MP01; cf. SuppPub1 (Ref. 21), Fig. S1-20/21 for scene geometry].

4. Coupled volumes

Coupled volumes, as they are used, for example, in concert hall design to achieve a variable reverberation time, typically lead to a double sloped EDC.[44] Measured and simulated EDCs for the 1 kHz octave are shown in Fig. 4 for a reverberation chamber, which was coupled to the laboratory by a door with an opening angle of $\phi = 30.4°$ (scene 8). Source and receiver were both located inside the laboratory with distances of 2.4 m and 2.2 m, respectively, to the door. The double sloped decay is clearly visible in the measured data, where the transition between decay rates of the reverberation chamber and laboratory room appears at approximately $t = 0.3$ s. When analyzing the results of this scene, it has to be considered that the EDC simulation is sensitive to the ratio of the reverberation times of both individual rooms, thus, highly depends on the provided boundary conditions. In most cases, the simulations exhibit only a weakly double sloped EDC with the exception of A that seems to correctly simulate both decay rates but fails in the correct simulation of the transition time. EDCs evaluated for different octave bands and a door opening angle of $\phi = 4.1°$ show the same trends (cf. SuppPub2,[21] Figs. S2-34–S2-37).

B. Complex scenes

1. Room acoustical parameters

Figure 5 shows the reverberation time T_{20} estimated from measured and simulated RIRs and averaged across ten source and receiver positions. Figure 5 also shows the Eyring reverberation times[45] calculated based on the room volumes provided in Table II and the absorption coefficients provided to the software teams.

In contrast to the simple scenes, the differences between measurement and simulation here refer both to deficits of the simulation algorithms and the possibly incorrect estimation of absorption coefficients with the *in situ* measurements conducted. Both uncertainties also occur in room acoustical design practice; the results are thus a valid indication of reliability of room acoustical simulation as a planning and design tool. As a result of both sources of error, a trend for overestimating the actual reverberation times at low

FIG. 5. (Color online) Reverberation time: T_{20} calculated from measured and simulated RIRs. Results are averaged across ten source/receiver positions and evaluated for six octave bands. To improve the readability, the simulation results are shifted in horizontal direction, and the reference values are connected by lines.

frequencies and underestimating them at high frequencies can be observed. The simulations resulted in reverberation times that are closer to the Eyring estimates than to the measured values in most cases. Because the simulations and the Eyring estimates are based on the provided absorption data, this might indicate that differences between measurements and simulations are dominated by uncertainty in the absorption coefficients. The differences between measurement and simulation are particularly high for the 125 Hz and 250 Hz octave bands, where the measured reverberation times are, on average, overestimated by 58% (125 Hz) and 35% (250 Hz). For the mid-frequency range (500 Hz–2 kHz), there is not systematic deviation; the differences between simulation and measurement are, however, still above the JND in most cases. A systematical overestimation of the absorption coefficients at 1 kHz, which was observed in RR-I,[13] does not appear in the three scenes tested here.

The results for additional room acoustical parameters and all source and receiver positions show similar trends (cf. SuppPub3,[21] Figs. S3-1–S3-31). While the EDT is overestimated at low frequencies and underestimated at high
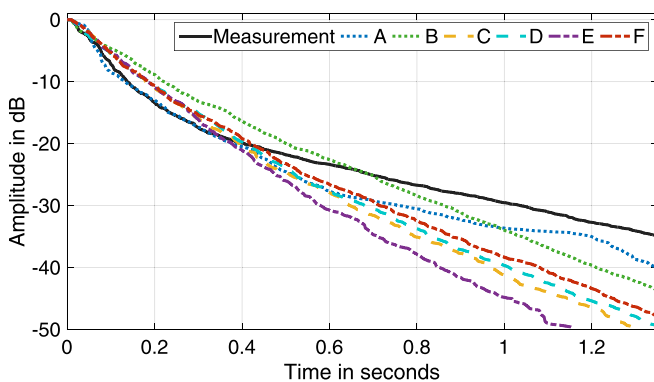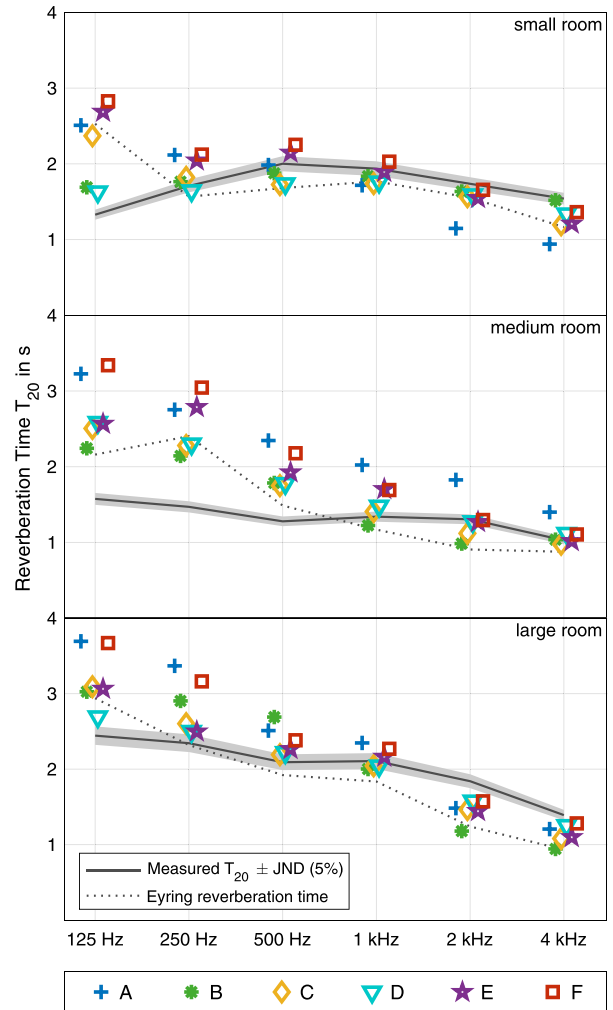

FIG. 4. (Color online) Coupled volumes: Measured and simulated energy decay curves of the coupled rooms for the 1 kHz octave band and a door opening angle $\phi = 30.4°$ [scene 8, source position LS02; receiver position MP03; cf. SuppPub1 (Ref. 21), Fig. S1-28/29 for scene geometry].

J. Acoust. Soc. Am. **145** (4), April 2019
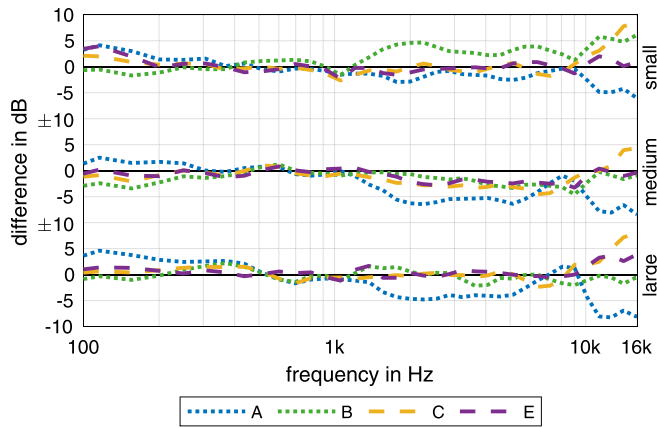
Brinkmann *et al.* 2753

FIG. 6. (Color online) Energetic differences between simulated and measured BRIRs in auditory filter bands averaged across source positions and ears for the three complex rooms [cf. SuppPub1 (Ref. 21), Figs. S1-15–S1-17 for scene geometry].

frequencies, the opposite holds for the clarity (C_{80}) and definition (D_{50}). Room acoustic parameters for individual source/receiver positions were analyzed for 1 kHz. The correlation between values based on measurements and simulations, however, was statistically non-significant with the exception of C_{80} and D_{50} where correlations between 0.7 and 0.9 were observed for the large room and all algorithms. For the EDT, a correlation of -0.7 occurred for D in the small room, indicating a reversed spatial dependency in this case (cf. SuppPub3,[21] Figs. S3-1–S3-2).

2. Spectral differences

Spectral differences between measured and simulated BRIRs with neutral head orientation are shown in Fig. 6. Averaged results are given because correlations among ears and source positions were high ($\emptyset = 0.8$), although average errors are between 2 dB and 2.5 dB. Software A always shows a bass boost and a lack of energy at high frequencies, while software C exhibits a high-frequency boost in all cases. Overall, smallest differences were observed for E ($\emptyset = 1.3$ dB), followed by B and C ($\emptyset = 2.2$ dB), and A ($\emptyset = 3.2$ dB).
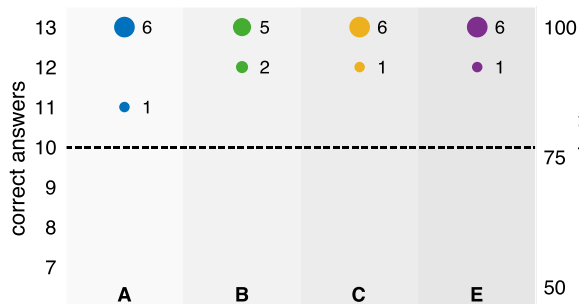

FIG. 7. (Color online) Results of the test for authenticity: Numbers of correct answers (left y axis) and corresponding detection rates in percent (right y axis). The size of the dots and the numbers next to them show how many participants had identical results. Results on or above the dashed line indicate significant differences, i.e., non-authentic simulations. Correct answers of 50% denote guessing, and 75% denotes the threshold of perception.

3. Perceptual evaluation

Taking into account the previously analyzed differences in T_{20} and the magnitude spectra, *authenticity* of the simulated rooms can presumably not be reached. This is proved by the results from the test for authenticity, assessing the perceptual identity of measured and simulated BRIRs of the medium room (scene 10) in a 2AFC listening test paradigm (cf. Fig. 7). Apparently, all participants could reliably identify differences between reality and simulations with detection rates of $p_{2AFC} \geq 0.98$, and the number of correct answers clearly exceeding the critical value of $N_{crit.} = 10$ for all simulation algorithms. Thus, none of the auralizations managed to be indistinguishable from the measured reference. This means that at the time being, blind simulations starting without *a priori* knowledge about reverberation times, etc., cannot lead to authentic results.

Results for the evaluation of *plausibility*, testing the credibility of simulations vs measurements with respect to an inner acoustic reference, are given in Fig. 8. The simulations were perceived as plausible in most cases, indicated by sensitivity values below the critical value. However, slight differences between the algorithms emerge: Simulation B was perceived as plausible by all participants ($\hat{d}'_{mean} = 0.07$), one participant detected artifacts in simulations C and E ($\hat{d}'_{mean} = 0.07$, and 0.3), and simulation A was perceived as implausible by 3 participants ($\hat{d}'_{mean} = 0.75$).

Differences in *specific auditory qualities* were measured using selected attributes of the SAQI (Fig. 9). Median values and 95% bootstrap confidence intervals (CIs; non-parametric resampling, bias corrected and accelerated CI calculation[46]) are given because the ratings were not normally distributed in the majority of cases. The auralizations of simulated BRIRs were directly compared to their measured counterparts, thus, a rating of 0 indicates no perceivable difference, and a rating of ± 1 stands for maximum differences.

The cases where the CIs do not overlap zero are taken as an indication of significant deviations between measurement and simulation (cf. Fig. 10). Here, differences become obvious between the different algorithms, between the two
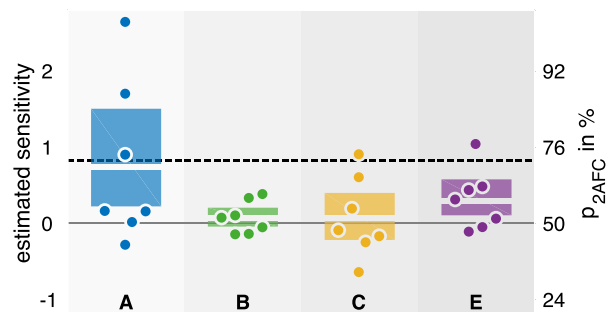

FIG. 8. (Color online) Results of the test for plausibility: Estimated individual sensitivities $\hat{d}'$ (left, y axis) and corresponding 2AFC detection rates p_{2AFC} (right, y axis) are given by the points (offset in horizontal direction to improve readability). Individual sensitivities on or above the dashed line indicate non-plausible simulations. Correct answers of 50% denote guessing, and 75% denotes the threshold of perception. The boxes show the group mean and 90% bootstrapped CIs [non-parametric resampling, bias corrected, and accelerated CI calculation (Ref. 46)]. A tabular overview of the individual results is given in SuppPup3 (Ref. 21), Fig. S3-35.
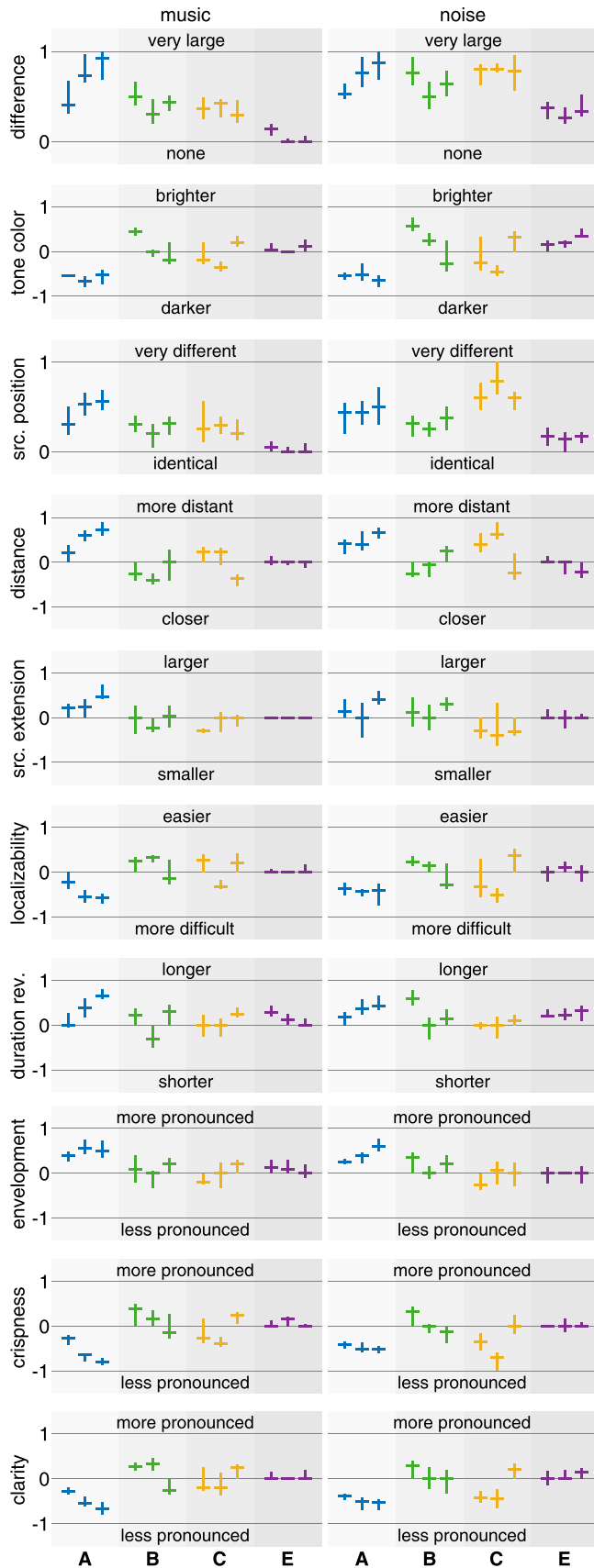
FIG. 9. (Color online) Differences in specific auditory qualities, measured with attributes of the SAQI, showing the median of differences between simulation and measured reference (horizontal lines) with 95% bootstrap CIs (vertical lines). The ratings were given for music (string quartet, left) and pulsed pink noise (right) as audio content and for the small, medium, and large rooms (from left to right).

audio contents, and, to a lesser degree, between the three different rooms. Whereas the softwares B, C, and E show significant deviations in 12%–30% of the experimental trials, this is the case in 88% of the cases for A. Also visible is a large difference between the two audio contents with the pulsed pink noise making the differences in most qualities more noticeable.

These visual inspection observations are confirmed by a three-factorial analysis of variance (ANOVA for repeated measurements) to test for significant differences concerning the normally distributed rating item *difference* with the factors *algorithm* (A,B,C,E), *room* (small, medium, large), and *content* (music, noise). It shows a highly significant main effect for the factor *algorithm* $[F(3) = 128.9, p < 0.001, \eta_p^2 = 0.82]$. Bonferroni-corrected *post hoc t*-tests showed that E performed significantly better than the remaining algorithms (all $p < 0.001$), while A was significantly worse than the others (all $p < 0.001$). Variation of the presented *room* led to a small but significant main effect $[F(2) = 3.2, p = 0.048, \eta_p^2 = 0.1]$, with larger perceived deviations from the reference for the large room compared to the other two (estimated marginal means: $\epsilon_{small} = 0.49$, $\epsilon_{medium} = 0.48$, $\epsilon_{large} = 0.52$; standard errors ≈ 0.025). Moreover, perceived differences turned out to be significantly larger for pink noise compared to the musical content across all algorithms and rooms $[F(1) = 78, p < 0.01, \eta_p^2 = 0.74]$. A detailed report of the ANOVA statistics is given in SuppPub3,[21] Figs. S3-36–S3-39.

To highlight the qualitative pattern of perceptual differences between simulation and measurement, a three-way multivariate analysis of variance (MANOVA for repeated measurements) was carried out for all attributes except *difference*. An inspection of the model residuals proved that the requirement of normality was met (SuppPub3,[21] Fig. S3-40). Here, the factor *content* had a multivariate main effect [*Pillai's Trace* = 0.753, $F(9,20) = 6.79$, $p < 0.001$, $\eta_p^2 = 0.75$] with always larger perceived deviations for the noise signal, although not every univariate main effect is significant. The factor *algorithm* also generated a multivariate main effect [*Pillai's Trace* = 1.715, $F(27,234) = 11.572$, $p < 0.001$, $\eta_p^2 = 0.57$] with significant univariate main effects for *all* qualities (all $p < 0.01$). Finally, a multivariate main effect [*Pillai's Trace* = 0.952, $F(18,98) = 4.944$, $p < 0.001$, $\eta_p^2 = 0.48$], encompassing five significant univariate main effects was occurring for the factor *room* (all $p < 0.01$). Noteworthy, the factor *algorithm* explains considerably more variance than the *room* ($\eta_p^2 = 0.57$ vs $\eta_p^2 = 0.48$), and also causes the largest range in the estimated marginal means $[\mu(\Delta\epsilon) = 0.47$ vs 0.1 *room*, and 0.07 *content*, $\mu(\cdot): =$ average across qualities], showing that the *algorithm* has the strongest influence on the perceived differences between simulations and reference. The interactions *algorithm × content*, and *algorithm × room* are significant for all qualities as well [*Pillai's Trace* = 0.908, $F(27,234) = 3.763$, $p < 0.001$, $\eta_p^2 = 0.3$, and *Pillai's Trace* = 1.504, $F(54,990) = 6.133$, $p < 0.001$, $\eta_p^2 = 0.25$], demonstrating that no single algorithm clearly outperforms the others with respect to all rooms, content types, and perceptual qualities.

To give an overview of the size of the simulation-related deviations in the various qualities, the estimated

J. Acoust. Soc. Am. **145** (4), April 2019
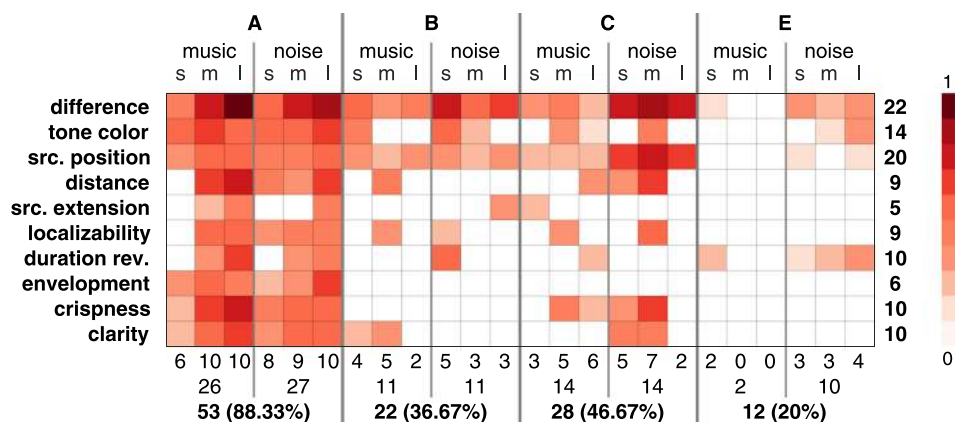
Brinkmann *et al.* 2755

FIG. 10. (Color online) Results of the SAQI test: Degree of deviations by algorithm, audio content, room size, and perceptual quality: White areas denote CIs overlapping with 0, shaded areas denote CIs *not* overlapping with 0, in which case the shading denotes the absolute median ratings in the range between 0 and 1 as indicated by the color bar. Numbers indicate the sum of significant deviations across rows and columns. Results for the small, medium, and large rooms are indicated by the letters *s*, *m*, and *l*.

marginal means for the software algorithms and their variation across rooms are given in Table III. The overall picture is in line with previous observations: Smallest differences were observed for E, medium differences for B and C, and largest differences for A. These differences are quite consistent across the three rooms. A detailed report of the MANOVA statistics is given in SuppPub3,[21] Figs. S3-40–S3-51.

Finally, a mixed regression model[47] was estimated with *difference* as dependent and the other nine qualities as independent variables. This was done to assess the importance of each perceptual dimension for the degree of overall perceived differences as expressed in the *difference* score. For this purpose, absolute values were taken, thus, assuming that positive and negative deviations (e.g., *tone color*: darker–to brighter) would equally contribute to the perceived *difference*. To test for multicollinearity, bivariate Pearson correlations between absolute scores of ratings of all attributes were calculated, with $r = 0.24$ on average and $r \leq 0.51$ in all cases. Thus, no qualities were excluded from the regression model and only removed in case of non-significant contributions to the prediction. The model included a random intercept term for *participant* in order to control for individual rating thresholds, and assumed a first-order auto-regressive residual covariance matrix due to repeated measurements. An inspection of the model residuals showed that the requirement of

normality was met (cf. Table. IV). The final model (with *duration of reverb* and *crispness* removed due to non-significant influence) accounts for $R^2 = 55.9\%$ of the variance (marginal $R^2 = 41.3\%$, Ref. [48]) and is shown in Table IV. *Tone color* has the largest influence on the *difference*, followed by *source position* and *localizability*.

C. Primary research data

The database of acoustical scenes (Table I), including all data (3D models, absorption and scattering coefficients, source and receiver information) provided to the participants is available as an electronic publication.[20] It now contains also the reference measurements that were not available for the participants. A description of all scene configurations and a comprehensive compilation of all results of the physical and perceptual evaluation is available in the supplemental material.[21] These also include examples of the audio stimuli of the listening tests with a static version of the originally dynamic binaural auralizations for neutral head orientation. Auralizations based on the measured data are directly followed by auralizations based on the simulated data as verbally announced. The audio files are diffuse field compensated and should be played back via headphones.

TABLE III. Estimated marginal means ϵ of perceived differences between measurement and simulation for ten perceptual qualities (SAQI attributes), and their range $\Delta\epsilon$ across rooms (in parentheses). The marginal means were obtained by ANOVA. The last row shows the mean absolute values.

Quality	A	B	C	E
Difference	0.68 (0.29)	0.53 (0.19)	0.56 (0.06)	0.24 (0.10)
Tone color	−0.54 (0.13)	0.18 (0.62)	−0.05 (0.56)	0.15 (0.18)
Source position	0.48 (0.16)	0.32 (0.11)	0.47 (0.07)	0.14 (0.06)
Distance	0.46 (0.43)	−0.12 (0.28)	0.13 (0.56)	−0.07 (0.17)
Source extension	0.21 (0.33)	0.06 (0.28)	−0.12 (0.09)	0.03 (0.06)
Localizability	−0.37 (0.16)	0.10 (0.25)	−0.01 (0.50)	0.07 (0.13)
Duration rev.	0.34 (0.41)	0.09 (0.54)	0.05 (0.13)	0.20 (0.04)
Envelopment	0.40 (0.29)	0.06 (0.28)	−0.03 (0.11)	0.04 (0.14)
Crispness	−0.48 (0.26)	0.07 (0.24)	−0.20 (0.62)	0.08 (0.05)
Clarity	−0.44 (0.17)	0.07 (0.28)	−0.12 (0.41)	0.04 (0.06)
∅	0.44 (0.26)	0.16 (0.31)	0.17 (0.31)	0.11 (0.10)

TABLE IV. Mixed regression model showing the influence of the different qualities on the perceived overall *difference*. (Left) Standardized model estimates (beta weights). All included qualities have significant contributions with $p < 0.05$. (Right) Distribution of the model residuals and the corresponding normal probability density function with identical mean and standard deviation.

Quality	Beta weight	
Tone color	0.294	
Source position	0.162	
Localizability	0.139	
Clarity	0.084	
Distance	0.083	
Crispness	0.082	
Envelopment	0.081	

IV. DISCUSSION

When interpreting the present results, readers should be reminded that the participating software developers were not allowed to change the input data (source/receiver directivities and absorption/scattering coefficients) with the exception of the geometrical resolution of the 3D models. This *blind* evaluation was considered the best way to assess quality differences between simulation algorithms because the fitting of parameters would compensate for shortcomings of the numerical simulation and make deviations from the reference—and most likely also between algorithms—appear smaller than they really are.[18]

In the *physical* evaluation of the participating algorithms, simulated IRs were compared with the measured reference. Only in the case of specular reflections on a quasi infinite surface (scene 1), the algorithms were able to match the spectral and temporal behaviour of the reference well. Small deviations already occurred for reflections on a quasi infinite two-dimensional diffusor (scene 1), where no algorithm was able to exactly match the comb filter structure of the magnitude spectrum. Even if this might have only minor consequences for the room acoustical parameters of complex environments,[15,16] it is likely to introduce coloration in the modeling of early reflections. For the reflection on the finite plate (scene 2), only one algorithm with an implementation of first-order edge diffraction was able to approximate the diffraction around the plate for wave lengths, which are large compared to the dimension of the plate. No algorithm, however, accurately modeled the coloration due to grazing sound incidence at the same plate. Because almost all relevant acoustic environments contain reflectors or objects of limited size, this is another source that might introduce severe coloration in modeling early reflections. Whereas the precise modeling of scattering effects of diffusing structures will remain a challenge in GA, there are approaches to account for diffraction in image source models and ray tracing,[7] which the authors believe deserve more attention to improve the tested algorithms. Modeling diffraction becomes even more important in case the direct sound path is blocked by an object. Results of scene 5 showed that by modeling only a single diffraction path the spectral shape of the transfer function is not well preserved. Neglecting diffraction entirely will be even worse, keeping in mind the increased relevance of dynamic simulations for virtual acoustic reality, where sudden jumps in loudness, tone color, or source position are likely to be perceived if the listener passes objects blocking the direct sound path. The double sloped decay of a coupled room (scene 8) could also not be modeled by most simulation algorithms, which might cause differences in the perceived reverberation tail. The observed errors could be caused by an insufficient number of rays in the ray tracing and/or, again, by omitting diffraction around the area that couples the two volumes. At this point, analytical and stochastic models for the energy decay of coupled volumes (cf. Luizard *et al*.[49] for an overview) could be used as a reference to improve the behaviour of GA-based simulation algorithms.

The simple scenes discussed above (scenes 1–8) could be accurately described by means of a 3D model, source and receiver directivities, as well as absorption and scattering coefficients or complex impedances. For complex scenarios, however, the acoustic surface properties (absorption, scattering) were estimated based on narrow band *in situ* measurements[50] or from material descriptions and pictures. As a consequence, differences between reference and simulation may either stem from shortcomings of the algorithms or uncertainties in the description of the boundary conditions. Since this is the case also in real-life applications, the discussion of the results for scenes 9–11 gives an impression of the general reliability and the systematic errors that occur in the application of these algorithms for acoustic planning tasks.

A comparison of the temporal structure of the simulated and measured IRs for the three rooms (SuppPub3,[21] Figs. S3-12, S3-22, S3-32) reveals that not all strong individual reflections are correctly modeled. This is due to the missing or insufficient representation of diffraction phenomena, possibly in combination with the impact of angle dependent surface properties, which are not considered by any of the algorithms. At least for reflections that occur before the perceptual mixing time,[51] the difference between the measured and simulated reflection patterns is likely to be audible.

Considering the calculated room acoustical parameters according to ISO 3382-1,[17] there is no systematic deviation between measurement and simulation for values in the medium frequency range (500 Hz–2 kHz); in many cases, the deviation is within the JND, which can be considered as a critical perceptual threshold. While a systematic overestimation of the reverberation time at 1 kHz was, unlike in the past,[13] no longer observed, probably due to the better *in situ* measurements or improved databases with tabulated absorption coefficients, there is still a systematic overestimation of low-frequency reverberation as well as a tendency to underestimate the reverberation time above 2 kHz. Both effects can be traced back to inaccurate absorption coefficients in connection with the geometry used, whose resolution seems to be optimal only for the middle frequency range.[18]

In addition to the physical evaluation, a *perceptual* evaluation of the different simulation algorithms was conducted based on two overall measures for the degree of perceived difference between simulation and reference, and on a qualitative description of the differences based on attributes from the SAQI. The test for *authenticity* showed that differences between simulations under test and the reference were always audible. This finding corresponds to the physical evaluation where no algorithm met the investigated room acoustical parameters within the tolerance of the JND in all frequency bands. Considering the high sensitivity of the test and human auditory system in general, it seems unlikely that the simulation of a complex acoustic environment will be able to achieve authenticity in a blind comparison in the foreseeable future. Even if fitting the input data would be allowed it could be argued that, on the one hand, the accuracy of wave-based simulations that numerically solve the wave equation will always be limited by the quality of input data describing the sound source and the boundary conditions. This will remain a challenging and, currently, at least partially unsolved task.[18] On the other hand, the limited accuracy of the modeling of diffraction and scattering in GA

J. Acoust. Soc. Am. **145** (4), April 2019

Brinkmann *et al.*   2757

was shown to introduce errors that are very likely to be audible, even if providing accurate input data.

While authenticity, denoting that the simulation sounds exactly like the real room, is a very strict criterion, *plausibility*, denoting the occurrence of obvious artifacts in the simulation which can be detected even without comparison with an explicit reference, can be considered a minimum quality criterion for virtual acoustic realities. In the present test, three out of four algorithms were able to provide plausible auralizations for most of the participants, and only one algorithm produced detection rates well above the threshold of perception. The averaged detection rates of $0.52 \leq \bar{p}_{2\text{AFC}} \leq 0.58$ for B, C, and E are comparable to those found for non-individual binaural simulation based on measured BRIRs [$\bar{p}_{2\text{AFC}} = 0.51$ (Ref. 43) and $\bar{p}_{2\text{AFC}} = 0.55$ (Ref. 52)].

The perceived *difference* between reference and simulation was mostly caused by differences in *tone color* and perceived *source position*, as could be demonstrated by regression analysis of the ratings of the SAQI attributes (Table IV). The deviations in *tone color* can be attributed to the inadequate modeling of early and late reflections for the reasons discussed above. Interestingly, the systematic low-frequency overestimation and high-frequency underestimation of T_{20} in the simulated IRs did not lead to a corresponding rating of tone color (cf. Figs. 5 and 9). In fact, the majority of the simulations by algorithms B, C, and E were rated as brighter than the measurement, indicating that the bass ratio, i.e., the ratio of reverberation times at low and medium frequencies, is not a reliable indicator for tone color in this case. It seems that the missing or insufficient modeling of diffraction for early reflections is what leads to an unnaturally bright sound impression. This was only not the case for software A, which showed a strong low-frequency boost already for the single reflection on the quasi infinite surface [Fig. 1(a)].

Differences in *source position* are most probably a by-product of spectral differences leading to mislocalizations in elevation.[53,54] Although it was not distinguished between localization errors in horizontal and vertical direction in the listening test, a vertical displacement is much more likely because the low-frequency interaural time difference, which dominates horizontal localization,[55] was well preserved in the simulations (except for C), which were controlled by a signal-related analysis (cf. SuppPub3[21] Fig. S3-33).

Compared to previous attempts, the current round robin has given a much more detailed insight into the performance of room acoustical simulation algorithms. This was made possible by the creation of a database of acoustical scenes with well controlled information on geometry and boundary conditions, highlighting different acoustic phenomena, and by conducting a technical *and* perceptual evaluation of the generated auralizations. This procedure entailed a larger effort on the side of the developers to calculate the required IRs, and is one reason why the number of participating software teams was lower than in previous attempts.[13–16] Some of the features that would allow easier accessibility for benchmarking tasks like the current one, however, would also be valuable extensions of the software packages for practical application. These include interfaces to import external HRTF sets into the software, the scripting and automation for different source and receiver combinations, different project variants, or different HRTF orientations for dynamic binaural synthesis. It should also be noted that the computation times for simulating an IR strongly vary with the software, which indicates opportunities for performance optimization in some cases. Such an optimization might be a prerequisite for the future implementation of computationally more demanding models for diffraction and scattering.

Since the database and the reference measurements are now open for free access, they can also be used by the developers of room acoustical simulation software themselves to evaluate the performance of new modeling approaches. The improvement of simulation software, as well as the extension of the database by further acoustic scenes, will be a reason for the authors to repeat this test in the future.

V. CONCLUSION

In the first round robin on room acoustical simulation and auralization, the simulation results for six simple scenes and three complex rooms provided by six teams using five different acoustic simulation algorithms were compared against measured data with respect to physical and perceptual properties. The results demonstrate that most present simulation algorithms based on GA generate obvious model errors once the assumptions of an infinite reflective baffle are no longer met. As a consequence, they are neither able to provide an exact pattern of early reflections, nor do they provide an exact prediction of room acoustic parameters outside a medium frequency range of 500 Hz–2 kHz.

In the perceptual domain, the algorithms under test could generate mostly plausible but not authentic auralizations. That means the difference between simulated and measured IRs of the same scene was always clearly audible. Most relevant for this perceptual difference are deviations in tone color and source position between measurement and simulation, which to a large extent can be traced back to errors in the simulation of early reflections, due to the simplified use of random incidence absorption and scattering coefficients and the missing or insufficient modeling of diffraction. Hence, room acoustical simulations are, unlike measurement-based auralizations,[39] not yet suitable to accurately predict the perceptual properties of sound sources in virtual acoustic environments at the current state of the art. Moreover, significant differences between different simulation algorithms have to be expected.

These conclusions hold for the conducted blind comparison task with initial parameter estimates, as is the case in the acoustic design of not yet existing venues. As soon as this estimate can be fitted to the measurement of a (partially) existing environment, modeling errors will become smaller automatically.

From a methodological point of view, we are convinced that the combination of an open database containing acoustic scenes[20] and a repeated comparison of different simulation algorithms against this reference could provide good prerequisites for the further improvement of room acoustical simulation. Since room acoustical simulation will be more and

more important for the generation of virtual acoustic realities, this evaluation should be based on physical as well as perceptual criteria.

ACKNOWLEDGMENTS

[1] H. Kuttruff, "Digital simulation of concert hall acoustics and its applications," Acoust. Bull. **16**(5), 5–8 (1991).

[2] J. H. Rindel, "The ERATO project and its contribution to our understanding of the acoustics of ancient theatres," in *The Acoustics of Ancient Theatres Conference*, Patras, Greece (2011).

[3] S. Weinzierl, P. Sanvito, and C. Büttner, "The acoustics of Renaissance theatres in Italy," Acta Acust. united Acust. **101**(3), 632–641 (2015).

[4] Z. Schärer Kalkandjiev and S. Weinzierl, "The influence of room acoustics on solo music performances: An empirical investigation," Psychomusicol.: Music, Mind, Brain **25**(3), 195–207 (2015).

[5] C. Schissler, A. Nicholls, and R. Mehra, "Efficient HRTF-based spatial audio for area and volumetric sources," IEEE Trans. Vis. Comput. Graph. **22**(4), 1356–1366 (2016).

[6] M. Vorländer, *Auralization. Fundamentals of Acoustics, Modelling, Simulation, Algorithms and Acoustic Virtual Reality*, 1st ed. (Springer, Berlin, 2008).

[7] L. Savioja and U. P. Svensson, "Overview of geometrical room acoustic modeling techniques," J. Acoust. Soc. Am. **138**(2), 708–730 (2015).

[8] A. Southern, S. Siltanen, D. T. Murphy, and L. Savioja, "Room impulse response synthesis and validation using a hybrid acoustic model," IEEE Trans. Audio Speech Lang. Process. **21**(9), 1940–1952 (2013).

[9] M. Kleiner, B.-I. Dalenbäck, and P. Svensson, "Auralization—An overview," J. Audio Eng. Soc. **41**(11), 861–875 (1993).

[10] J. H. Rindel, "Room acoustic modelling techniques: A comparison of a scale model and a computer model for a new opera theatre," Build. Acoust. **18**(3-4), 259–280 (2011).

[11] T. Otsuru, T. Sakuma, and S. Sakamoto, "Constructing a database of computational methods for environmental acoustics," Acoust. Sci. Technol. **26**(2), 221–224 (2005).

[12] M. Hornikx, M. Kaltenbacher, and S. Marburg, "A platform for benchmark cases in computational acoustics," Acta Acust. united Acust. **101**(4), 811–820 (2015).

[13] M. Vorländer, "International round robin on room acoustical computer simulations," in *15th International Congress on Acoustics*, Trondheim, Norway (1995), pp. 689–692.

[14] I. Bork, "A comparison of room simulation software—The 2nd round robin on room acoustical computer simulation," Acta Acust. Acust. **86**, 943–956 (2000).

[15] I. Bork, "Report on the 3rd round robin on room acoustical computer simulation—Part I: Measurements," Acta Acust. Acust. **91**, 740–752 (2005).

[16] I. Bork, "Report on the 3rd round robin on room acoustical computer simulation—Part II: Calculations," Acta Acust. Acust. **91**, 753–763 (2005).

[17] ISO 3382-1, "Measurement of room acoustic parameters—Part 1: Performance spaces" (International Organization for Standards, Geneva, Switzerland, 2009).

[18] M. Vorländer, "Computer simulations in room acoustics: Concepts and uncertainties," J. Acoust. Soc. Am. **133**(3), 1203–1213 (2013).

[19] N. Tsingos, I. Carlbom, G. Elko, R. Kubli, and T. Funkhouser, "Validating acoustical simulations in the Bell Labs Box," IEEE Comput. Graph. Appl. **22**(4), 28–37 (2002).

[20] L. Aspöck, F. Brinkmann, D. Ackerman, S. Weinzierl, and M. Vorländer, "GRAS—Ground truth for room acoustical simulation" (2018), available at http://dx.doi.org/10.14279/depositonce-6726 (Last viewed March 27, 2019).

[21] See supplemental material at http://dx.doi.org/10.1121/1.5096178 for detailed scene descriptions, additional results and details of the statistical analyses.

[22] A. Lindau, T. Hohn, and S. Weinzierl, "Binaural resynthesis for comparative studies of acoustical environments," in *122th AES Convention*, Vienna, Austria (2007), Convention Paper 7032.

[23] A. Lindau and S. Weinzierl, "On the spatial resolution of virtual acoustic environments for head movements on horizontal, vertical and lateral direction," in *EAA Symposium on Auralization*, Espoo, Finland (2009).

[24] W. R. Thurlow, J. W. Mangels, and P. S. Runge, "Head movements during sound localization," J. Acoust. Soc. Am. **42**(2), 489–493 (1967).

[25] J. C. B. Torres, L. Aspöck, and M. Vorländer, "Comparative study of two geometrical acoustic simulation models," J. Brazilian Soc. Mech. Sci. Eng. **40**(6), 300 (2018).

[26] Ahnert Feistel Media Group, "EASE—User's guide and tutorial (version 4.3)," available at http://www.afmg-support.de/SoftwareDownloadBase/AFMG/EASE/EASE_4.3_Tutorial_English.pdf (Last viewed April 3, 2019).

[27] Odeon A/S, "ODEON room acoustics software user's manual" (2016), available at https://odeon.dk/wp-content/uploads/2017/09/ODEON_Manual.pdf (Last viewed April 3, 2019).

[28] D. Schröder, "Physically based real-time auralization of interactive virtual environments," Ph.D. thesis, RWTH Aachen, Aachen, Germany, 2011.

[29] T. Wendt, S. van de Par, and S. D. Ewert, "A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation," J. Audio Eng. Soc. **62**(11), 748–766 (2014).

[30] F. Brinkmann, A. Lindau, S. Weinzierl, S. v. d. Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "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 (2017).

[31] F. Brinkmann, A. Lindau, S. Weinzierl, G. Geissler, S. van de Par, M. Müller-Trapet, R. Opdam, and M. Vorländer, "The FABIAN head-related transfer function data base" (2017), available at http://dx.doi.org/10.14279/depositonce-5718.2 (Last viewed March 27, 2019).

[32] M. Berzborn, R. Bomhardt, J. Klein, J.-G. Richter, and M. Vorländer, "The ITA-Toolbox: An open source MATLAB toolbox for acoustic measurements and signal processing," in *Fortschritte der Akustik—DAGA 2017*, Kiel, Germany (2017), pp. 222–225.

[33] M. Guski and M. Vorländer, "Comparison of noise compensation methods for room acoustic impulse response evaluations," Acta Acust. united Acust. **100**, 320–327 (2014).

[34] M. Slaney, "Auditory toolbox. Version 2," Technical Report No. 1998-010, Interval Research Corporation (1998).

[35] AES Standards Committee, *AES69-2015: AES Standard for File Exchange—Spatial Acoustic Data File Format* (Audio Engineering Society, Inc., New York, 2015).

[36] M. Geier, J. Ahrens, and S. Spors, "The Sound Scape Renderer: A unified spatial audio reproduction framework for arbitrary rendering methods," in *124th AES Convention, Preprint 7330*, Amsterdam, The Netherlands (2008).

[37] M. S. Puckette, "Pure data," in *Int. Computer Music Conf. (ICMC)*, Thessaloniki, Greece (1997).

[38] A. Lindau and F. Brinkmann, "Perceptual evaluation of headphone compensation in binaural synthesis based on non-individual recordings," J. Audio Eng. Soc. **60**(1/2), 54–62 (2012).

[39] F. Brinkmann, A. Lindau, and S. Weinzierl, "On the authenticity of individual dynamic binaural synthesis," J. Acoust. Soc. Am. **142**(4), 1784–1795 (2017).

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

[41] K. R. Murphy and B. Myors, "Testing the hypothesis that treatments have negligible effects: Minimum-effect tests in the general linear model," J. Appl. Psychol. **84**(2), 234–248 (1999).

[42] L. Leventhal, "Type 1 and type 2 errors in the statistical analysis of listening tests," J. Audio Eng. Soc. **34**(6), 437–453 (1986).

[43] A. Lindau and S. Weinzierl, "Assessing the plausibility of virtual acoustic environments," Acta Acust. united Acust. **98**(5), 804–810 (2012).

[44] N. Xiang, Y. Jing, and A. C. Bockman, "Investigation of acoustically coupled enclosures using a diffusion-equation model," J. Acoust. Soc. Am. **126**(3), 1187–1198 (2009).

[45] H. Kuttruff, *Room Acoustics*, 5th ed. (Spon, Oxford, UK, 2009).

[46] J. Carpenter and J. Bithell, "Bootstrap confidence intervals: When, which, what? A practical guide for medical statisticians," Statist. Med. **19**(9), 1141–1164 (2000).

[47] J. J. Hox, *Multilevel Analysis. Techniques and Apllications, Quantitative Methodology*, 2nd ed. (Routledge, New York, 2010).

[48] S. Nakagawa and H. Schielzeth, "A general and simple method for obtaining R^2 from generalized linear mixed-effects models," Methods Ecol. Evol. **4**, 133–142 (2013).

[49] P. Luizard, J.-D. Polack, and B. F. G. Katz, "Sound energy decay in coupled spaces using a parametric analytical solution of a diffusion equation," J. Acoust. Soc. Am. **135**(5), 2765–2776 (2014).

J. Acoust. Soc. Am. **145** (4), April 2019

Brinkmann *et al.* 2759

[50] E. Brandão, A. Lenzi, and S. Paul, "A review of the *in situ* impedance and sound absorption measurement techniques," Acta Acust. united Acust. **101**(3), 443–463 (2015).

[51] A. Lindau, L. Kosanke, and S. Weinzierl, "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 (2012).

[52] C. Pike, F. Melchior, and T. Tew, "Assessing the plausibility of non-individualised dynamic binaural synthesis in a small room," in *AES 55th International Conference*, Helsinki, Finland (2014).

[53] J. Blauert, *Spatial Hearing. The Psychophysics of Human Sound Localization*, revised ed. (MIT Press, Cambridge, MA, 1997).

[54] R. Baumgartner, P. Majdak, and B. Laback, "Modeling sound-source localization in sagittal planes for human listeners," J. Acoust. Soc. Am. **136**(2), 791–802 (2014).

[55] F. L. Wightman and D. J. Kistler, "The dominant role of low-frequency interaural time differences in sound localization," J. Acoust. Soc. Am. **91**(3), 1648–1661 (1992).