



The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms[★]

Sarah Roßkopf^{1,*} , Leon O.H. Krocze¹ , Felix Stärz², Matthias Blau^{2,4} , Steven Van de Par^{3,4} , and Andreas Mühlberger¹ 

¹Department of Psychology, Clinical Psychology and Psychotherapy, University of Regensburg, Universitätsstr. 31, 93053 Regensburg, Germany

²Institut für Hörtechnik und Audiologie, Jade Hochschule, Ofener Str. 16, 26121 Oldenburg, Germany

³Acoustics Group, Carl von Ossietzky Universität Oldenburg, Carl-von-Ossietzky-Str. 9-11, 26129 Oldenburg, Germany

⁴Cluster of Excellence “Hearing4All”

Received 21 February 2024, Accepted 19 September 2024

Abstract – Virtual Reality (VR) enables the presentation of realistic audio-visual environments by combining head-tracked binaural auralizations with visual scenes. Whether these auralizations improve social presence in VR and enable sound source localization comparable to that of real sound sources is yet unclear. Therefore, we implemented two sound source localization paradigms (speech stimuli) in a virtual seminar room. First, we measured localization continuously using a placement task. Second, we measured gaze as a naturalistic behavior. Forty-nine participants compared three auralizations based on measured binaural room impulse responses (BRIRs), simulated BRIRs, both with generic and individual head-related impulse responses (HRIRs), with loudspeakers and an anchor (gaming audio engine). In both paradigms, no differences were found between binaural rendering and loudspeaker trials concerning ratings of social presence and subjective realism. However, sound source localization accuracy of binaurally rendered sound sources was inferior to loudspeakers. Binaural auralizations based on generic simulations were equivalent to renderings based on individualized simulations in terms of localization accuracy but inferior in terms of social presence. Since social presence and subjective realism are strongly correlated, the implementation of plausible binaural auralizations is suggested for VR settings where high levels of (social) presence are relevant (e.g. multiuser interaction, VR exposure therapy).

Keywords: Binaural auralizations, Virtual reality, Sound source localization, Subjective audio realism, Eye-tracking, Social presence, Head related impulse responses, Gaze, Attention

1 Introduction

There is a wide range of applications for virtual reality (VR) scenarios in many different scientific fields. For instance, there is broad evidence for the usefulness of VR in psychotherapy (for a review: [2, 3]). What unites many disciplines that deal with VR is the interest in creating a realistic and convincing virtual environment. An important concept besides realism is presence. Presence is defined as the subjective experience of “being there” – in the virtual scene [4]. There are several internal and external factors on presence (for a review see [5]). Internal factors are among other psychological variables, such as mood or state. In general, higher levels of arousal contribute to presence [5]. In addition, external factors such as features of the VR, con-

tent, and tasks contribute to presence [6]. In the field of acoustics, research is conducted on various methods to synthesize acoustic virtual environments [7]. These methods, further called auralizations, aim at creating a realistic spatial auditory impression (for a review see [8]). High degrees of realism of head-tracked binaural auralizations regarding acoustical properties such as reverberance or source distance could be confirmed [9]. Furthermore, in an audiovisual virtual seminar room scenario, participants were not able to reliably distinguish between real sound sources (loudspeakers) and binaural auralizations [10]. In this study, we will use these head-tracked binaural auralizations and will refer to them as (plausible) binaural auralizations in the following.

Importantly, studies that include spatial audio and improved audio quality (i.e. a higher number of audio channels) have reported increased levels of presence [11, 12]. Furthermore, task-relevance of the three-dimensionality of audio influence presence [13]. Presence is even suggested

[★]A part of the here reported data has already been published in the proceedings of the Forum Acusticum 2023 at Torino [1].

*Corresponding author: sarah.rosskopf@ur.de

to serve as a measurement variable for audio quality in VR since ratings are sensitive to changes in bass or sound pressure level [14]. A psychologically highly relevant and specific aspect of presence is social presence, which means the feeling of being together with another person in the virtual environment [15]. It was found that the higher realism of virtual agents enhances social presence (for a review see [16]). Taken together, this suggests that a positive effect of spatial audio and therefore acoustic realism on social presence can be assumed. The extent to which plausible binaural auralizations are creating (social) presence and how they influence the perceived acoustic realism in audiovisual VR is an open question. An increase in social presence would be especially valuable for applications in the field of virtual exposure therapy for example social anxiety disorder (for overview see [17]).

A further goal of this study was to examine the extent of realism of plausible binaural auralizations in terms of sound source localization accuracy in VR. Close-to-real acoustic properties and indistinguishability of plausible binaural auralizations from real sound sources have already been shown [9, 10]. Especially for natural acoustic stimuli such as speech or music in reverberant acoustic environments such as seminar rooms authentic simulations can be achieved [18]. An open question is whether the plausible binaural auralizations also allow close-to-real sound source localization in an audiovisual virtual seminar room scenario and whether this affects the subjective experience. Since visual cues are known to influence sound source localization, known as the ventriloquist effect (see [19]), effects of visual VR have to be considered. A further open question is whether individualized binaural auralizations are needed or whether binaural auralizations based on a default specification are equivalent in this context. Evidence on the need for individual head-related impulse responses (HRIRs) in the rendering of spatial audios is inconsistent (see [20]). While in earlier research in particular, sound source localization performance was found to be worse when using generic HRIRs (e.g. [21]), later research indicates an equivalency of individualized and generic HRIRs especially for lifelike stimuli such as speech [9, 22]. However, it was also found that even small deviations in sound source localization induced by non-individualized in comparison to individualized HRIRs binaural auralizations of free-field drone noise impaired perceived realism in VR [23]. There is also broad evidence that the effect of individualized versus non-individualized HRIRs is differentially depending on individual characteristics such as hearing experience [24] or particularly deviating head shapes such as in children [25].

When comparing sound source localization accuracy, not only differences between binaural auralizations have to be taken into account. Also, the peculiarities of audiovisual virtual reality affect sound source localization. Not only the perception of visual objects and rooms is compressed to less than three-quarter of the original size using a head mounted display (HMD) [26, 27] but wearing the HMD itself decreased sound source localization accuracy of pink or white noise played by loudspeakers [28, 29]. Providing visual information about possible source positions could in

part compensate for the negative effects of the HMD [28, 30]. It is also well documented that when an audiovisual scene is being integrated, visual cues dominate the auditory ones (ventriloquist effect [31]). Furthermore, we could previously show that visual virtual room information altered distance perception [32]. When an audiovisual congruent room was displayed, participants were more accurate in terms of distance estimations than when an oversized virtual room was displayed. Also, the selection of a suitable measurement tool for sound source localization in VR is an important issue. Auditory distance perception is often examined based on verbal reports (of explicit scales or unitless implicit scales) or motoric tasks such as walking (for a review see [33]). The influence of the method itself to measure egocentric distances as well as the interaction with the virtual environment must be taken into account. When participants had to estimate egocentric distances (to visual targets) in VR, verbal answers were found to be more accurate than visually guided walking, where participants indicated distances by walking up to turned off visual targets [34]. Visual, haptic, and locomotive feedback can improve egocentric distance perception in VR [35], but influences also depend on the technologies used [27]. In real-life scenario, the movement of visual markers to adjust estimated sound source positions – further called placement paradigm – was found to provide more accurate estimations than verbal reports [36]. Furthermore, the placement paradigm enables continuous measurement of distance and azimuthal sound source perception. The placement task was therefore used in several studies and was found to be a sensitive tool for investigating differences in distance perception [32, 37].

One further advantage of the use of VR is that relatively complex scenes can be created and that high flexibility for different measurement tools and strategies is provided. By tracking gaze behavior, a comparable naturalistic task for sound source localization is provided. Humans tend to look at people who are speaking. This gaze behavior is modulated by audiovisual speech integration [38]. Speaker-directed gaze orientation is not only part of multimodal social attention but also influences auditory perception. Acoustic cues are derived more accurately when presented frontally or slightly lateral to the head [39]. Additionally, shifting the gaze toward a sound source enhances cue discrimination even when the head is not moved [40]. Gaze behavior can also be used as a measure of sound source localization [41]. Eye-tracking paradigms provide a naturalistic and implicit tool for measuring attentional resources and can be seen as a task that has a high relevance for real-world situations (validity) and is at the same time as standardized as possible [1].

This study aimed at investigating sound source localization with two different paradigms. First, a placement paradigm was used to precisely and continuously measure sound source localization accuracy. Second, an eye-tracking paradigm was used to gain evidence of the usability of this unobtrusive and naturalistic measurement method and on the localizability of different Audio Conditions in a more complex seminar room scene including virtual agents. The data of 25 participants on the eye-tracking paradigm has already

been published in the proceedings of the Forum Acusticum 2023 at Torino [1]. Since we regard externalization of auralizations as a prerequisite for more or less accurate sound source localization, we also investigated in-head localization or externalization of Audio Condition.

We compared the plausible binaural auralizations to real audio sources (loudspeakers) and an anchor (state-of-the-art 3D-audio-sound implemented within the VR engine – Steam Audio v 4.1.4, Valve Corporation, Bellevue, WA, USA [42]). The anchor was selected due to its practical relevance and external validity as a common and easy-to-implement audio presentation technique in research where audio effects are typically not the main target of investigation (e.g. VR exposure therapy). Notably, while the steam audio engine can be seen as such a practical technique, it also incorporates important features such as room geometry, surface material and head tracking. Based on these thoughts, the steam audio engine is an ideal comparison to test whether more sophisticated yet technically complex auralizations (such as stimuli rendered with RAZR) can yield improvements in VR experience (i.e. with respect to social presence and subjective realism) which justifies the increased technical complexity. We, therefore, investigated whether the accuracy of sound source localization of plausible binaural head-tracked binaural auralizations equals that of real sound sources and is superior to the state-of-the-art game-engine anchor. Furthermore, we were interested in whether similar effects for subjective experience in terms of social presence and perceived spatial audio quality can be achieved and whether these dimensions are correlated.

We hypothesized that all plausible binaural auralization methods are equivalent to real audio sources regarding sound source localization. This was investigated in the placement paradigm with respect to distance and azimuthal localization. In the eye-tracking paradigm, sound source localization accuracy was investigated in terms of how often participants directed their gaze toward the virtual agent at the sound source position. If one comparison results in significant differences, the hypothesis must be declined. Next, we hypothesized that the binaural auralizations simulated with RAZR [43] which are based on generic HRIRs are equivalent to simulated BRIRs (RAZR) using individual HRIRs (simIndivHRIRs) and also to binaural auralizations based on measured BRIRs (measHATS). We therefore hypothesize that the binaural auralizations are all equivalent regarding sound source localization. Third, we expected similar effects (equivalency of binaural auralizations and real audio source; equivalency of binaural auralizations) for subjective experience measured via two rating items, first concerning social presence, and second concerning subjective realism. Since the anchor was found to be perceived less frequently externalized in pilot tests when it was played among the other audio conditions, we expect inferior sound source localization of the anchor condition. Lastly, we hypothesize superiority of the binaural auralization and the loudspeakers over the anchor concerning social presence and subjective realism, since they were precisely tailored to each other and the experimental room. All hypotheses were preregistered (<https://osf.io/9yqf7>; <https://osf.io/2q4y3>).



Figure 1. Auditory measurement system of individual and generic (HATS) head-related impulse responses (HRIRs).

2 Methods

The goal of the study was to compare three different plausible binaural auralization with loudspeakers and an anchor in audiovisual virtual environments. In our investigations, two different measurement paradigms were implemented in a virtual seminar room (see Figs. 3 and 4). Main outcome variables were sound source localization accuracy and subjective experience in VR (ratings of social presence and realism).

2.1 Sample

Healthy adult individuals with self-reported unimpaired hearing, normal or corrected to normal vision, and German-speaking experience of a minimum of 5 years were included in the study. Our sample ($N = 49$) consisted of 38 female and 11 male participants aged between 19 and 46 ($M = 23.2$, $SD = 4.6$). The majority of participants were students ($n = 44$). All participants gave written informed consent. The study was in line with the Declaration of Helsinki and approved by the local ethics committee (University of Regensburg).

2.2 Room and visual virtual setup

The experiment took place in a seminar room of the University of Regensburg (room size: $10.6 \text{ m} \times 7.1 \text{ m} \times 3.3 \text{ m}$, reverberation time: 0.91 s). The room consists of four concrete walls, one of which is equipped with a large mirrored window, an acoustically optimized ceiling, and a carpet on concrete floor (see Figs. 3 and 4). For the visual virtual room, we created a photorealistic model of the

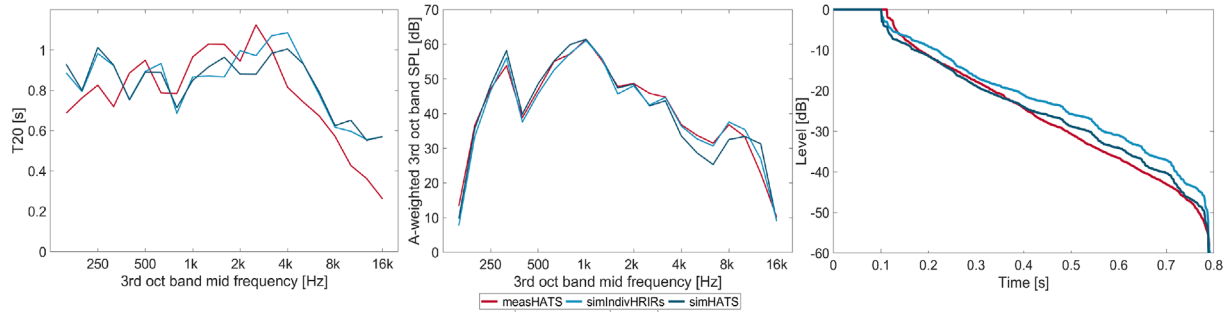


Figure 2. Objective frequency dependent data derived from BRIRs (frontal-close speaker see Fig. 3). Left: reverberation time (T20), middle: A-weighted third-octave band sound pressure levels after convolving the speech stimulus used in the listening test with the BRIRs, right: energy decay curves.



Figure 3. Setup of placement paradigm. On the left: Position of participant and loudspeakers in the room. In the middle: Placement task (Agent had to be placed at perceived sound source). On the right: Source and listener positions.



Figure 4. Setup of eye-tracking-paradigm. On the left, the position of participants and loudspeakers (in block 1) in the room are shown. In the middle, a visual virtual room with the eye-tracking task for sound source localization is depicted (participants had to look at the perceived sound source). On the right: source and listener positions.

seminar room with the Unreal Game Engine (v 4.27, Epic Inc.) and Blender (v 2.79, Blender) using textures based on high-resolution photographs (as already described in the previous study [1]). The visual virtual environment was presented via an HMD (Vive Pro Eye, HTC). This device was also used for the measurement of gaze by eye-tracking. For audiovisual virtual reality, an inaudible workstation with passive cooling was used (Silentmaxx PC Kenko S-770i). The visual representation of the virtual room (in terms of HMD position and direction) was matched to the real room via an in-house-developed two-point calibration technique using custom-made mounts for the HTC motion controller [44]. Since the headphones were

mounted to the HMD this also implied partial calibration of the headphone position (only pitch, yaw and roll data). Data on our calibration technique was collected and a very high accordance of real and virtually visible positions could be affirmed (see [44]). Virtual agents were created using MakeHuman (v 1.2) and Blender (v 2.79).

2.3 Auditory setup

2.3.1 Audio conditions

We compared five different audio presentation modes (for an overview see Tab. 1). The first Audio Condition involved loudspeakers in the room as real sound sources,

Table 1. Overview of investigated audio conditions. For plausible binaural auralizations detailed information on used binaural room impulse response (BRIR) sets, used head-related impulse response (HRIR) set and headphone equalization (HPEQ), spatial resolution, and frequency independent direct-to-reverberant energy ratio (DRRs) in dB of BRIRs are given.

Name of audio condition			Acoustic specifications				
Plausible binaural auralizations	BRIR	Loudspeaker directivity genelec 8030b	HRIR	HPEQ	Spatial resolution of BRIRs		DRR
					Elevation	Azimuth	
measHATS	Measured in real room	Implicitly included	Kemar HATS	Individual	–	–90° to 90° in 5° steps	1.65
simIndivHRIRs	Simulated using RAZR	Measured	Individual	Individual	–30° to 30° in 7.5° steps	–90° to 90° in 5° steps	–0.54
simHATS	Simulated using RAZR	Measured	Kemar HATS	Individual	–30° to 30° in 7.5° steps	–90° to 90° in 5° steps	0.39
Comparison conditions			Description				
Loudspeaker Anchor	Loudspeakers as real sound sources in the room (Genelec 8030b) head-tracked binaural 3D auralizations created by Steam Audio v 4.1.4 (Valve Corporation)						

which provided the best possible comparison condition. Next, we used head-tracked binaural auralizations based on three different BRIR sets, for which high plausibility was found [10]. More precisely, the second Audio Condition, referred to as measHATS, was a binaural auralization based on BRIR sets that were measured in the real room using a commercial head-and-torso-simulator (HATS; Kemar type 45BB, GRAS Sound and Vibration A/S, Holte, Denmark). Therefore, the source directivity of the loudspeakers (also used as comparison condition) were implicitly included. The head-above-torso orientation of the HATS was varied between -90° and 90° in 5° steps, resulting in 37 azimuthal orientations. The spatial resolution is based on the proposed minimum BRIR grid resolution by Lindau et al. [45]. It was shown that for the majority of listeners ($>95\%$) the here used spatial resolution of BRIRs is sufficient to create a plausible binaural simulation for natural stimuli such as music in reverberant spaces [46]. MEMS microphones (TDK type ICS-40619, TDK InvenSense, San Jose, CA, USA) inserted into the ear canals of the HATS using PIRATE earplugs [47] were used for the measurements. BRIRs were measured using multiple exponential sweep stimuli (for further details see [9]). The elevation angle was fixed at 0° , implying that the sound source remained static even when participants raised or lowered their head. The third and fourth Audio Condition were binaural auralizations based on BRIR sets which were simulated using RAZR (v 0.962b, [43]). The simulated reverberation time T_{20} was fitted to previously measured monaural impulse responses of the seminar room. Additionally, the source directivity of the loudspeakers (same as for measHATS and comparison condition loudspeaker) was included in the simulation via a database with directivities measured by ourselves (see [9]). The simulated room impulse responses were combined with measured HRIRs. The measurement system for the HRIRs (see Fig. 1) is a replication of the setup constructed and used at Jade Hochschule Oldenburg, for further details see [9]. Simulated BRIRs were obtained for 37 azimuthal head-above torso orientations (-90° to 90° in 5° steps) and nine elevation angles (-30° to 30° in 7.5° steps). In Audio Condition number

three, further referred to as simIndivHRIRs, individually measured HRIRs were used for the rendering of BRIRs. In Audio Condition number four, subsequently referred to as simHATS, generic HRIRs were used, measured with the above-described HATS. The simulated ear height of the plausible auralizations depended on the experimental task. In the placement task, the simulated and real ear height was set to 1.30 m since participants were seated in an auditorium (using a height-adjustable chair). During the eye-tracking paradigm, the simulated ear height was set to 1.60 m, since participants took up the lecturer position in front of the auditorium. No adjustment of participants' real ear height was made here, as the natural standing position of participants in front of an auditorium was targeted. Last, the fifth audio mode, subsequently called anchor, consisted of head-tracked binaural 3D auralizations created by a state-of-the-art audio engine (Steam Audio v 4.1.4, Valve Corporation, Bellevue, WA, USA) implemented in the Unreal Engine. Real-time ray tracing was used for modeling physics-based reverb. We used the above described virtual room as static geometry, and specified the room acoustic via predefined acoustic material properties (e.g. carpet for the floor). Frequency-dependent occlusion and sound propagation via nearest-neighbor option were chosen. The volume attenuation was adapted on a one-individual perceptive level towards the loudspeaker condition, to avoid salient loudness differences.

2.3.2 Technical setup

We compared binaural auralizations to real sound sources in the room. Two-way active loudspeakers (Genelec 8030b, Genelec Oy, Isalmi, Finland) were used as real sound sources in the room. All other Audio Conditions were presented using a headphone amplifier (Lake People G103P, Lake People Electronic GmbH, Konstanz, Germany) and extra-aural headphones (AKG K1000, AKG Acoustics GmbH Vienna, Austria), which were mounted to the HMD with custom-made 3D printed supports [48]. Compared to circumaural headphones, the spectral influence of the extra-aural headphones on the sound

field produced by a real loudspeaker is smaller [49] especially for speech stimuli [50] but it cannot be completely excluded. Likewise, the presence of the HMD may introduce subtle spectral colorations in comparison to not wearing an HMD. Nonetheless, no differences regarding plausibility could be found for binaural auralizations based on measurements with or without HMD in a prior study [51]. For playback on loudspeakers and headphones, an external audio interface (RME Fireface, UC, Audio AG Haimhausen, Germany) was used.

2.3.3 Audio Stimuli

The stimuli consisted of dry recordings of female speech and were derived from a German learning program (studio21 A1 und A2, Cornelsen Verlag [52]). The stimuli were loudness normalized (based the integrated loudness function from the Matlab “Audio ToolboxTM”) following EBUR 128 [53] and Hann windowed (10 ms) to avoid cutting artifacts. For the placement paradigm, German greetings (e.g. “Hallo”) were used. For the eye-tracking paradigm, typical language course statements from one word (e.g. “station”) to five-word sentences (e.g. “What is it called in German?”). The order of stimulus presentation was pseudo-randomized via randomization lists. For the presentation of stimuli, we created five different randomization lists per paradigm, each beginning with a different audio mode (lists were counterbalanced across participants). Three blocks per list were created within which combinations of stimuli, speaker position, and Audio Condition were repeated equally often. The stimuli were pseudo-randomized within the three blocks with the following constraints: not more than three repetitions of the same Audio Condition, same position, and same utterance.

2.4 Design

We manipulated the Audio Condition with five levels: real sound source, measured HATS BRIRs, simulated BRIRs based on individual HRIRs, simulated BRIRs based on HATS, and anchor. We further varied the position (angle, distance) from which the sounds were played (loudspeakers) or simulated. With two different tasks, we investigated the influence of binaural auralizations on sound source localization, social presence, and subjective realism in virtual reality.

2.5 Procedure

The experimental procedure comprised two appointments. On the first appointment, participants gave written informed consent and filled in questionnaires on demographic data, their hearing experience, and their degree of social anxiety (SPIN, [54]). Furthermore, the individual measurement of participants’ HRIRs and an individualized headphone equalization (HPEQ) was conducted. The second appointment consisted of two parts, first the placement paradigm, second the eye-tracking paradigm. For the general preparation of both experimental parts, the current status of hearing impairment (self-report) was collected and

impaired participants (e.g. current otolaryngological symptoms) were excluded. Next, participants were informed about and prepared for the case of motion sickness. Further, they reported about their affective state via the positive and negative affect schedule (PANAS, [55]). Next, the HMD was fitted to participants by adjusting the position to the one used during the individual HPEQ measurement, via recorded strap positions (individually adjusted length of side strap and length of Velcro fastener on crown). Participants entered the seminar room “blind-folded” (by wearing the HMD) and were guided to the starting position by the experimenter and virtual footprints. This procedure allowed participants to remain unaware of the positions of the loudspeakers. Afterward, the first experimental part consisting of the placement task was conducted (see Sect. 2.6.1). Presence during the placement paradigm was measured via the multimodal presence scale (MPS; [56]). After completing the placement task, participants took a break outside the experimental room and VR. Then, participants accomplished the second experimental part, the eye-tracking paradigm, after which again, presence was measured via MPS. Additionally, post-assessment questionnaires on both experimental parts and mood (PANAS) were conducted. Neither before, during nor after the experiment did the participants see the room or the source positions since they always wore the HMD when being there.

2.6 Sound source localization measurement

2.6.1 Placement paradigm

2.6.1.1 Specific setup

Participants had to place a virtual agent at the position in the virtual room, where they assumed the sound source. The placement was accomplished by pointing the HTC Vive motion controller towards a selected position at which then, the agent appeared and which had to be confirmed by another button press. The agents could be placed continuously on the floor of the virtual room without any prior restrictions. More precisely, the x - and y -axis, but not the z -axis coordinates of the position of the virtual agents could be altered to indicate sound source location. The agent only appeared when a specific position has been selected (is hidden when the sound is played) to prevent a possible bias of visual information on auditory localization (ventriloquist effect [31]). If a sound was in-head localized, participants were instructed to place the agent close to their own chair (radius of 80 cm) resulting in the agent disappearing only a sphere being left.

The gaze direction of agents was shifted during positioning so that agents always kept facing toward the participants. The sound sources also faced the participants. The loudspeakers in the room and the simulated sound sources were at a height of 1.15 m, corresponding to the height of the mouth of the virtual agent. There were four different source positions; speaker 1: 2.80 m, 0°; speaker 2: 4.80 m, 0°; speaker 3: 2.45 m, 45°; speaker 4: 2.15 m, 90° (see Fig. 3). During the experiment, the position of participants,

virtual agents (position after placement), and the sound source positions were tracked. Participants were seated in the auditorium of a seminar room. Speech stimuli, that were supposed to evoke an orientation reaction (German Greetings) were played from the source positions. At the end of this experimental part, a brief assessment of perceived elevation of sound sources was conducted to gain preliminary insights whether Audio Conditions were perceived elevated. Therefore, participants had to adjust the height of a loudspeaker icon to the height at which they perceived a sound source. For all Audio Conditions, the same source position (speaker 1) and sound stimuli were used (one trial per condition).

2.6.1.2 The procedure of placement paradigm

Participants were seated on a height-adjustable chair placed in the auditorium of the room. The height of participants' ears was adjusted to 1.30 m. Then, the HMD displayed an exact replication of the seminar room. The participant was handed the controller with which the rest of the experiment could be conducted. Furthermore, participants were instructed to move their head only on a horizontal plane of -90 to 90° .

To learn the handling of the controller and the placement task practice trials were conducted where participants had to, first, place an agent on a visual target, second, place the agent at the position in the seminar room, where they assumed the sound source, third, indicate in-head localization of sounds, and last, use the interface for ratings. The practice trials were repeated until participants succeeded in placing the agent on a target with no more than 20 cm deviance and when predefined rating buttons were selected.

Following the practice trials the main task was started. It consisted of three blocks with 100 trials in total. After each block, participants were instructed to take a break. Each trial had the same procedure. Before the sound presentation, participants had to orient towards a fixation cross on the frontal wall. If the rotation of the HMD deviated more than 10° from the fixation cross, a red text was displayed, instructing participants, to "Please look straight forward." If verified, the sound was played at the predefined (via randomization list) location. Then, participants placed a female agent at the location in the room, where they assumed the sound source. Participants were able to change the position as often as they liked and there was no time limit for the task. Finally, participants had to confirm the chosen position with another button press and the next trial started after a delay of 3 s. After a fifth of trials, (in total 20 trials specified via a randomization list), a rating followed the localization task. Each rendering at each position had to be rated once concerning the (social) presence and the subjective realism. The rating interface was implemented in the script and was handled via the controller. The first item of the rating was the social presence rating: "Ich habe den Eindruck, dass die Begrüßung gerade von einer anwesenden Person stammen könnte." which translates as "I had the impression that the greeting a moment ago could have come from a present person." The second item of the rating was: "Der Klang war so wie in einem

Seminarraum." which translates as "The sound was like being in a seminar room." We used the second item as indicator for subjective realism of Audio Conditions. The ratings consisted of the statement and a 9-point Likert scale beneath it. The furthest left button was labeled "stimme nicht zu" [I disagree], and the furthest right button was labeled "stimme zu" [I agree]. We previously validated in a pilot study with five participants whether laypeople naïve towards room acoustic research can comprehend and answer these items. We tested six different room acoustic quality and realism items adapted from the Room Acoustical Quality Inventory (RAQI) [57] and three social presence items adapted from the Multimodal Presence Scale (MPS) [56] and chose the items which were the easiest to answer.

After all 100 trials had been presented, participants accomplished the task of indicating the perceived elevation of sound sources. Then, participants were guided out of the experimental room and took off the HMD. Last, participants completed a questionnaire on presence, the MPS, and a post-experiment questionnaire.

2.6.2 Eye-tracking paradigm

2.6.2.1 Specific setup

Participants were instructed that they would participate in a virtual language learning course. Participants were positioned in front of the auditorium at the lecturer position with a virtual visual notebook in front of them. All instructions, rating scales, and vocabulary stimuli were presented on the screen of the virtual notebook. The auditorium of the seminar room was filled with 16 virtual female agents. The virtual agents were animated sitting on a chair and breathing. They were positioned to fill the whole auditorium, see Figure 4. In the notebook, different written words were displayed. Participants were instructed that the words would be read aloud by one of the agents. To measure sound source localization, participants were instructed to look at the location in the room where they assumed the sound source. Gaze behavior was recorded and analyzed during the task. If participants did not externalize a sound, which means that in-head localization of a sound occurred, they were instructed to direct their gaze toward a blue button on the keyboard of the virtual notebook. The 120 trials of the eye-tracking paradigm were conducted in two blocks with varying source positions. The position of in total eight agents exactly matched the (virtual) loudspeaker position in the room (four per block). The real and virtual loudspeakers were directed forward (parallel to the side walls) and accordingly, all virtual agents directed their gaze straight forward. The position of the agents' mouths was at 1.15 m, which corresponded to the height of the acoustic center of the loudspeakers. All agents wore a face mask, to avoid influences of visual cues on localization (ventriloquist effects [31]). Loudspeakers were placed at distances from 2.70 m to 6.80 m, and at azimuthal angles of 2° – 27° (see Fig. 4). At the end of this experimental part, perceived loudness was assessed to gain preliminary insights on possible differences between the Audio Conditions. Therefore, participants were asked

about the perceived loudness of a first sound in comparison to a second sound (0 = much quieter, 1 = a bit quieter, 3 = neutral, 4 = a bit louder, 5 = much louder). The comparison (only one trial per Audio Condition) was always against the loudspeaker condition, but the order was alternated.

2.6.2.2 Procedure

This experimental part started with the five-point eye tracking calibration procedure provided by and presented with the Vive Pro Eye. The producers claim an accuracy of 0.5° – 1.1° for this eye-tracking system [58]. We further validated the accuracy of eye-tracking for each participant. Therefore, 24 visual target icons were presented at the positions where the agents were placed in the main experiment and participants were instructed to look towards the bull's eye of each target and then press a button. Gaze acuity was measured in degree for angle deviance and cm for distance deviance. After this, several practice trials were conducted to ensure understanding and manageability of tasks. Handling of ratings, the eye-tracking task, and what to do when sound was in-head localized were practiced. The instruction was to look towards the spot where participants assumed the sound source. After the practice trials, the first 60 trials (first block) were run. These were followed by a break (about 5 min) during which loudspeakers were rearranged and during which participants had a seat next to the previous position within the experimental room still wearing the HMD. Possible auditory cues during the rearrangement of the loudspeakers were masked by brown noise played via headphones. The loudspeakers were rearranged to increase the overall number of source positions despite a limitation by number of speakers and the interface. Then, the next 60 trials were conducted. All trials started with the visual display of the vocabulary item (word or short sentence) in the notebook (see Fig. 4). The orientation of the participant towards the notebook during the sound onset was controlled. If the rotation of the HMD exceeded 10° , a red text was displayed, instructing participants, to “Please look towards the screen.” If verified, the sound was played back at the designated location. Head movements and gaze toward the source were encouraged as soon as the sound was played. The gaze behavior was recorded and analyzed for three seconds. If no valid fixation on an object in the scene (no adjustment of gaze direction or fixation of the wall) was found, the trial was repeated. If the task was completed, the visual display of the vocabulary item disappeared and after an inter-trial interval of three seconds, the next trial started. After a sixth of the trials, the rating scales were presented in VR and had to be completed. As in the first task, each rendering at each position had to be rated concerning (social) presence and subjective realism. The rating interface was displayed on the virtual notebook, and was handled via the controller. The first item of the rating was the social presence rating: “Ich habe das Gefühl, dass gerade eine anwesende Person zu mir gesprochen hat.” which translates as “I have the feeling that a person present has just spoken to me.” The second item of the rating was the

subjective realism rating, which was the same as in the placement task. At the end of the localization task, the perceived loudness of the auralizations each in comparison to the loudspeaker was assessed. After the VR experiment, participants were guided to the anteroom, and again questionnaires on the experiment (difficulty of task, hypotheses, etc.) and experience in VR (MPS) had to be answered. Then, the final assessment of mood was completed via the PANAS questionnaire.

2.7 Outcome variables

All analyses for hypotheses tests were preregistered (<https://osf.io/9yqf7>; <https://osf.io/2q4y3>).

2.7.1 Sound source localization accuracy

2.7.1.1 Placement paradigm

Deviations between estimated and real angle (in degree) and distance (in cm) were calculated as primary outcome variables for sound source localization accuracy. Per Audio Condition and participant, 20 trials were averaged. The angle deviance was computed using the dot product of the vector between participant and real sound source and the vector between participant and estimated sound source position, converted to degrees for better interpretability (all analyses scripts are accessible in a public repository (<https://osf.io/9yqf7>; <https://osf.io/2q4y3>). The distance deviance (xy plane only) was computed as follows: the Euclidian distance between the participant and the estimated sound source position is subtracted from the Euclidian distance between the participant and the real sound source. As a consequence, a distance deviance of 0 equals perfect distance estimation, and a positive distance deviance indicates an overestimation of distance. Following our preregistered hypothesis rationale, an equivalency of Audio Conditions concerning localization accuracy will only be confirmed, if equivalency can be shown for both dependent variables.

2.7.1.2 Eye-tracking paradigm

As the primary outcome variable for sound source localization accuracy, the rate of fixations on correct agents per participant and Audio Condition was calculated. Only the first fixation was analyzed following a pre-registered analysis plan. Gaze behavior was analyzed offline using a custom Matlab script (v R2022a, The MathWorks, Inc., Natick, MA, USA) which categorized the gaze as fixation or saccade behavior. Fixations were defined using both velocity ($<75^{\circ}/s$) and gaze duration (>200 ms) criteria [59]. Additionally to correct fixations, angle deviance as well as distance deviance of fixated and real source position were calculated as further indicators for sound source localization accuracy (not preregistered).

2.7.1.3 Supplementary indicators

Additionally to these primary outcome variables (angle deviance and distance deviance), we computed following supplementary indicators to classify data concerning sound

source localization (non-preregistered analyses): We analyzed the amount of trials (in %), in which the agent was placed outside the walls of the visual room (Trials outside Room) and classified them as invalid. We further classified data as invalid, if the sound was not perceived externalized or if front-back confusions occurred. As a global indicator for localization accuracy, we computed “overall localization error” (in cm), which is calculated by the Euclidian distance between the estimated and real position of the sound source. As an indicator for systematic angular deviance on the horizontal plane (to left of right side), the azimuthal error was calculated, which ranges from -180° to $+180^\circ$. For the absolute distance error (in cm) as an indicator of the overall accuracy of distance estimation (regardless of whether the distance was overestimated or underestimated) the mean of absolute distance deviance was calculated.

2.7.2 Subjective experience

Ratings of social presence and subjective realism were analyzed to investigate subjective experience during audio-visual presentation. The mean rating value per participant and Audio Condition was computed as the dependent variable.

2.8 Statistical analyses

Statistical analyses were conducted in the R environment (v 4.3.2, [60]) using the packages lme4 [61], lmerTest [62], emmeans [63], and BayesFactor (Bayes-package, [64]). Generalized mixed-effect models of binomial family were computed for categorical data (externalization, front-back confusion, correct fixations), linear mixed-effect models were computed for continuous data (angle deviance in $^\circ$, distance deviance in cm). All models included fixed effects for Audio Conditions. Likelihood ratio tests were conducted to affirm the random effects structure. The final models included random slopes for Audio Conditions by subjects. We used the Bobyqa option of the lmerTest package to optimize models. To test the specified models for the significant effect of the fixed effects, F -values for continuous data and χ^2 values for categorical data are computed with an analysis of variances. Post-hoc simple contrast comparisons were conducted to test for differences following the preregistered hypotheses (<https://osf.io/9yqf7>; <https://osf.io/2q4y3>). Alpha was set to 5%, tests were corrected for multiple comparisons using the Holm method [65]. Since we hypothesized not only differences between but also equivalency of particular Audio Conditions, we conducted tests for equivalency. Following our preregistered analysis plan, if post hoc comparisons revealed no significant difference, Bayes factors were computed to test the probability of an equivalency of Audio Conditions. For higher clarity and better interpretability, Bayes factors for paired t -tests of the respective comparison of Audio Conditions were computed. We used non-informed prior distributions. We defined Bayes Factors greater than three as confirmative following the suggestions of Wagenmakers [66]. All anonymized data, as well as analysis scripts, are accessible in a

public repository (<https://osf.io/9yqf7>; <https://osf.io/2q4y3>).

3 Results

3.1 Placement paradigm

Figure 5 provides an overview of real and estimated source positions from the placement task for each Audio Conditions except the anchor. It can be seen, that front-back confusions occurred in all Audio Conditions, but more often in binaural auralization trials. All in all, the placement patterns appear to be comparable. Overestimation of distances occurred frequently. In Table 2, all outcome variables are reported. Note that per Audio Condition and Participant 20 trials were analyzed. Thus for outcome variables that are reported in %, values $<5\%$ implicate that on average in less than one trial per participant the relevant characteristic of the outcome variable (e.g. internalization) occurred.

3.1.1 Externalization rate

As can be seen in Table 2, in a vast majority of trials the sounds were not in-head localized (meaning high externalization rates) in all Audio Conditions except the Anchor. A mixed-effect logit model revealed a significant main effect Audio Condition, $\chi^2(4) = 119.9$, $p < .001$. Post-hoc comparisons revealed significant differences of Loudspeaker against Anchor ($\beta = -9.11$, $Z = -7.58$, $p < .001$), measHATS against Anchor ($\beta = -11.47$, $Z = -6.85$, $p < .001$), simIndivHRIRs against Anchor ($\beta = -8.84$, $Z = -8.84$, $p < .001$), and simHATS against Anchor ($\beta = -7.96$, $Z = -9.22$, $p < .001$). No significant differences of Loudspeaker against all binaural auralizations could be found as well as no significant differences of comparisons between the different plausible binaural auralizations.

3.1.2 Front-back confusions

Since in more than three out of four trials of the anchor stimuli were not perceived externalized, we did not include this condition in the statistical analysis for differences concerning front-back confusions. A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi^2(3) = 13.92$, $p = .003$. The only significant difference of post-hoc comparisons between Audio Conditions concerning front-back confusions could be found for loudspeakers against simHATS ($\beta = 2.69$, $Z = -2.74$, $p = .036$).

3.1.3 Sound source localization accuracy

Trials in which in-head localization or front-back confusion occurred were not included in this data analysis since they cannot be interpreted consistently in terms of localization accuracy (referred to as invalid trials). For this reason, too, the anchor condition was not included in this data analysis due to too many invalid trials. There were 25 participants (more than half of N) with not a single valid trial for the anchor condition, and in all participants ($N = 49$)

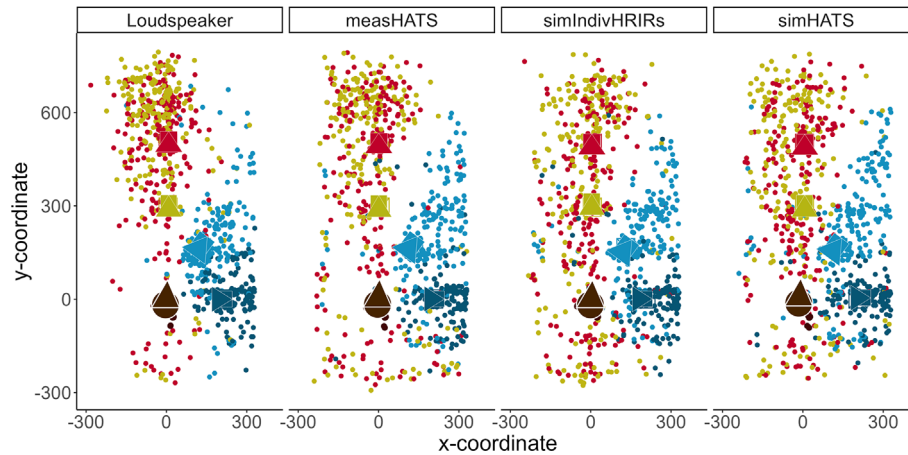


Figure 5. The positions of real sound positions (loudspeaker icons), participants (dark brown, schematic head for frontal direction heading nose), and the estimated sound positions (color depends on the real sound position of that trial) are shown as a function of their y - and x -coordinate in the virtual room. Note that participants' forward orientation is towards the positive x -coordinate.

Table 2. Outcome variables from placement task for audio condition.

Outcome variable	Loudspeaker		measHATS		simIndivHRIRs		simHATS		Anchor	
	M	SD	M	SD	M	SD	M	SD	M	SD
Trials outside room (in %)	2.9	14.1	2.7	12.5	1.9	10.8	1.9	8.2	1.7	5.6
In-head localization Trials (in %)	0.8	2.4	3.1	13.5	3.6	11.7	4.1	11.6	78.4	31.6
Front-back confusions (in %)	4.4	10.0	11.0	17.5	10.7	16.7	12.4	16.8	25.3 ^a	22.4
Invalid trials (in %)	7.0	15.9	12.9	21.3	14.7	23.5	16.2	21.8	73.3	28.6
Overall localization error (in cm) ^b	185.1	122.6	207.7	125.1	196.9	127.7	200.0	121.0	212.4 ^a	111.6
Angle deviance (in °, 0° to 90°) ^b	11.1	12.7	13.8	17.4	14.4	17.8	15.2	19.3	33.2 ^a	27.3 ^a
Azimuthal error (in °, -180° to 180°) ^b	-0.6	16.8	-0.03	22.2	-1.4	22.8	1.5	24.5	-4.7 ^a	42.8
Distance deviance (in cm) ^b	117.1	170.9	126.1	175.9	91.0	185.8	95.7	179.0	-142.0 ^a	150.5
Absolute distance Error (in cm) ^b	163.8	170.9	177.0	175.9	164.7	185.8	165.5	179.0	164.9 ^a	150.5

^a These values cannot be interpreted consistently due to insufficient valid trials.

^b For these analyses invalid trials were excluded.

more than half of the trials for the anchor condition were invalid.

3.1.3.1 Angle deviance

Data from one further participant had to be excluded due to only invalid trials for the measHATS condition. A linear mixed-effect model revealed no significant main effect of Audio Condition, $F(3, 65.54) = 2.72$, $p = .051$. Therefore, Bayes factors are computed to gain evidence for the hypothesis, that the binaural auralizations are equivalent to real sound sources and that all plausible binaural auralizations are equivalent concerning azimuthal localization. The Bayes factor for the hypothesis, that the plausible binaural auralizations are equivalent to loudspeakers is 0.74, which can be interpreted as anecdotal evidence for non-equivalency [67]. The Bayes factor for the hypothesis, that simHATS is equivalent to simIndivHRIRs is 4.72, which can be interpreted as moderate evidence for equivalency of these two conditions. The Bayes factor for the hypothesis, that measHATS and both simulated binaural auralizations are

equivalent is 6.17, which can be interpreted as moderate evidence for equivalency.

3.1.3.2 Distance deviance

The same data as for angle deviance was included for the analysis of distance deviance. A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(3, 65.16) = 8.28$, $p < .001$, on distance deviance. Post-hoc comparisons revealed no significant differences in Loudspeaker against measHATS, but significantly higher distance deviance of the loudspeaker against simIndivHRIRs ($\beta = 34.08$, $t = 3.46$, $p = .005$), and against simHATS ($\beta = 25.50$, $t = 2.51$, $p = .046$). Further, significant higher distance deviances are found for measHATS in comparisons against simIndivHRIRs ($\beta = 37.79$, $t = 4.34$, $p = .001$) and simHATS ($\beta = 29.21.03$, $t = 3.67$, $p = .001$). There was no significant difference between simIndivHRIRs and simHATS. The Bayes factor for the hypothesis, that simIndivHRIRs and simHATS are equivalent is 3.46 (can be interpreted as moderate evidence for equivalency).

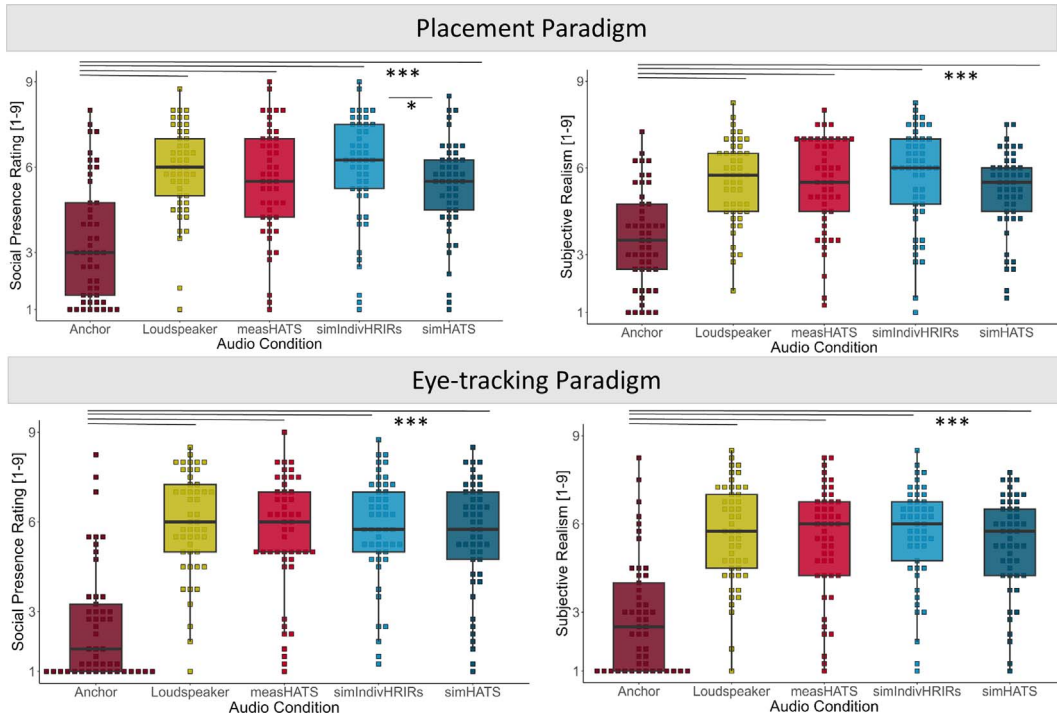


Figure 6. Subjective experience. On the left: Social presence ratings [“I had the impression that the greeting a moment ago could have come from a present person.” (placement paradigm) or “I have the feeling that a person present has just spoken to me.” (eye-tracking paradigm)]; 1 = “I disagree”, 9 = “I agree”] as a function of Audio Condition. On the right: Realism ratings [“The sound was like being in a seminar room.” 1 = “I disagree”, 9 = “I agree”] as a function of Audio Condition and separate for each paradigm.

In prior studies, we constantly found an overestimation of source distances when using visual VR [32, 37]. Therefore, we compared the mean distance error against zero. With a mean average of 75.82 cm ($SD = 113.41$) over all Audio Conditions except Anchor, there is strong evidence for an overestimation of sound distances, $t(195) = 9.36$, $p < .001$.

3.1.4 Subjective experience

3.1.4.1 Social presence

In Figure 6 the influence of Audio Condition on social presence and subjective realism is displayed. Ratings of all participants were included for analysis of rating data ($N = 49$). A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(4, 84.43) = 21.05$, $p < .001$, on social presence. In anchor trials, participants rated their social presence lower ($M = 3.3$, $SD = 2.4$) than in loudspeaker trials ($M = 5.9$, $SD = 2.1$, $\beta = -2.57$), measHATS trials ($M = 5.5$, $SD = 2.4$, $\beta = -2.16$) and both simulated BRIR trials (simIndivHRIRs: $\beta = -2.60$, simHATS: $\beta = -1.96$), all $ps < .001$. One further significant difference could be found. Social presence was rated higher for simIndivHRIRs ($M = 5.9$, $SD = 2.3$) than for simHATS ($M = 5.3$, $SD = 2.2$, $\beta = -0.49$, $t = -2.87$, $p = 0.036$). The Bayes factor for the hypothesis that participants perceive equivalent social presence during loudspeaker and binaural auralizations trials is 0.61, which can be interpreted as anecdotal evidence for non-equivalency. The Bayes factor for the hypothesis that social presence is equivalent for measHATS and both simulated binaural auralizations is 4.63.

3.1.4.2 Subjective realism

Overall, the relationship between Audio Condition and subjective realism is similar to the relationship between Audio Condition and social presence. A linear mixed-effect model revealed a significant main effect of Audio Condition, $F(4, 91.94) = 15.67$, $p < .001$, on subjective realism. Participants rated the realism of the anchor lower ($M = 3.6$, $SD = 2.2$) than the realism of loudspeakers ($M = 5.5$, $SD = 2.0$, $\beta = -1.93$, $t = -7.58$, $p < .001$), of measHATS ($M = 5.4$, $SD = 2.2$, $\beta = -1.86$, $t = -6.8$, $p < .001$), of simIndivHRIRs ($M = 5.7$, $SD = 2.1$, $\beta = -2.10$, $t = -7.40$, $p < .001$) and lower than simHATS ($M = 5.2$, $SD = 2.0$, $\beta = -1.60$, $t = -6.85$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. The Bayes factor for an equivalency of the subjective realism of loudspeakers and plausible binaural auralizations is 5.19. The Bayes factor for an equivalency of subjective realism of binaural auralizations based on measured BRIRs and simulated BRIRs is 6.43. For the hypothesis that simHATS and simIndivHRIRs are equivalent regarding subjective realism, a Bayes factor of 0.08 was found. This can be interpreted as anecdotal evidence for a non-equivalency.

3.2 Eye-tracking paradigm

3.2.1 Eye-tracking accuracy

We could confirm a high accuracy of the eye-tracking data. Per participant the individual deviance of gaze from 24 targets were measured. The median angular deviance was 0.93° (first row: 0.78° , second row: 0.98° , third row:

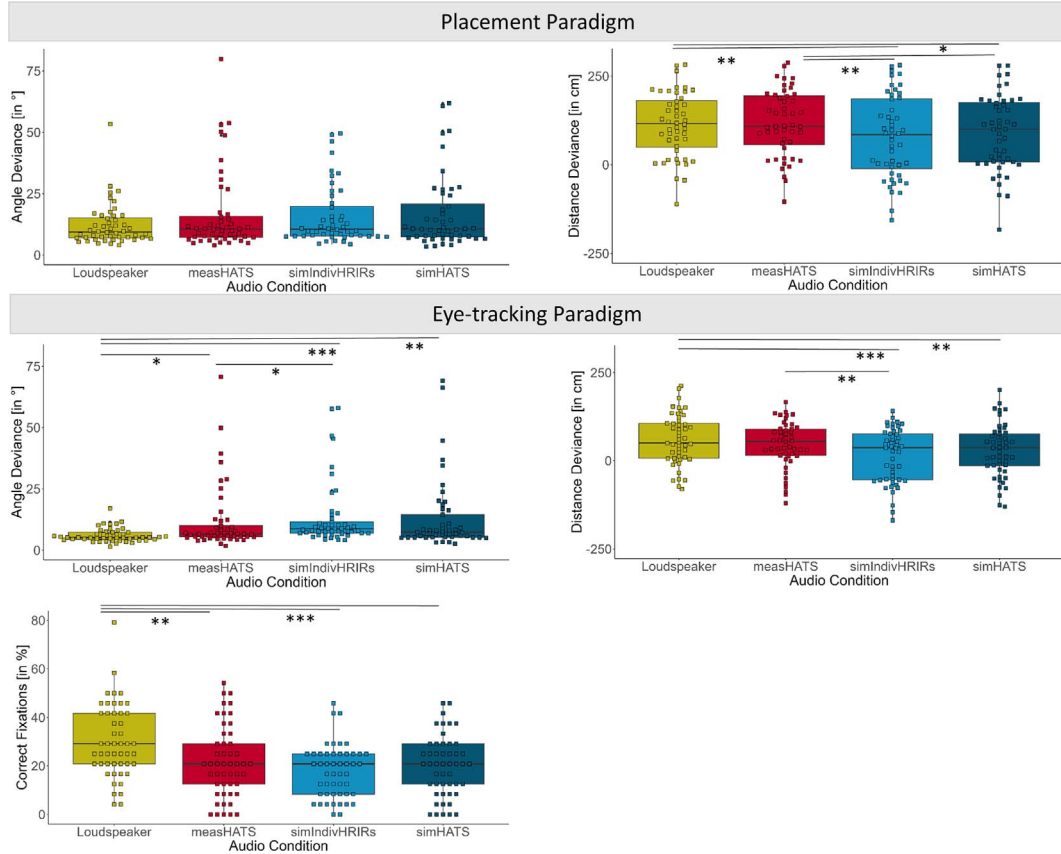


Figure 7. Sound source localization accuracy. On the left: Angle deviation in degree between real and estimated sound position as a function of Audio Condition. On the right: Distance deviance in cm between real and estimated sound position as a function of Audio Condition and separate for each paradigm. For the eye-tracking paradigm, one further figure shows the rate of correct fixations (in %).

1.02, fourth row: 0.96°). The median distance deviance was 5.66 cm (first row: 3.96 cm, second row: 4.94 cm, third row: 6.60 cm, fourth row: 7.71 cm).

3.2.2 Externalization rate

Again, in a vast majority of trials (in %) sounds were perceived externalized for loudspeaker condition ($M = 96.3$, $SD = 10.1$), measHATS condition ($M = 90.4$, $SD = 17.8$), simIndivHRIRs ($M = 90.8$, $SD = 18.4$), and simHATS ($M = 90.6$, $SD = 16.9$), whereas Anchor condition stimuli were mainly not perceived externalized ($M = 18.9$, $SD = 22.7$). A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi^2(4) = 93.73$, $p < .001$, on Externalization. Post-hoc comparisons revealed significant differences of loudspeaker against Anchor ($\beta = -10.79$, $Z = -4.14$, $p < .001$), measHATS against Anchor ($\beta = -11.82$, $Z = -6.17$, $p < .001$), simIndivHRIRs against Anchor ($\beta = -11.28$, $Z = -6.45$, $p < .001$), and simHATS against Anchor ($\beta = -13.0$, $Z = -5.32$, $p < .001$). No significant differences of loudspeaker against all binaural auralizations could be found and no significant differences of comparisons between the different plausible binaural auralizations.

3.2.3 Sound source localization accuracy – rate of correct fixations

Trials in which a sound was not externally perceived were not included in this data analysis. In Figure 7 the rate of trials, in which participants' first fixation was towards the correct agent (agent placed at speaker position) is shown respectively for each Audio Condition. Since again lots of trials for the Anchor condition are invalid ($M = 61.56$, $SD = 25.48$), the Anchor condition was excluded for inference statistics. A mixed-effect logit model revealed a significant main effect of Audio Condition, $\chi^2(3) = 22.77$, $p < .001$, on correct fixation. Post-hoc comparison revealed significant differences of loudspeaker against measHATS ($\beta = 0.49$, $Z = 3.63$, $p = .001$), against simIndivHRIRs ($\beta = 0.65$, $Z = 5.06$, $p < .001$), and against simHATS ($\beta = 0.56$, $Z = 4.27$, $p < .001$). There were no significant differences between the plausible binaural auralizations. For the equivalency hypothesis of simIndivHRIRs and simHATS concerning fixations on correct agents, a Bayes factor of 2.63 could be found (anecdotal evidence for equivalency). A similar degree of evidence could be found for the equivalency hypothesis of measured and simulated Binaural auralizations (Bayes factor = 1.45).

3.2.4 Subjective experience

3.2.4.1 Social presence

In Figure 6, data on subjective experience during the eye-tracking paradigm of 49 participants are shown. A linear mixed-effect model again revealed a significant main effect of Audio Condition on social presence, $F(4, 87.9) = 33.44$, $p < .001$. Participants rated social presence lower for anchor trials ($M = 2.6$, $SD = 2.1$) than for loudspeaker trials ($M = 5.9$, $SD = 2.2$, $\beta = -3.36$, $t = -11.10$, $p < .001$), than for measHATS trials ($M = 5.6$, $SD = 2.3$, $\beta = -2.99$, $t = -8.69$, $p < .001$), simIndivHRIRs trials ($M = 5.8$, $SD = 2.1$, $\beta = -3.21$, $t = -10.07$, $p < .001$) and lower than for simHATS trials ($M = 5.5$, $SD = 2.3$, $\beta = -2.94$, $t = -8.77$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. The Bayes factor for an equivalency of social presence during loudspeaker and plausible binaural auralizations trials is 1.51; for equivalency of measured and simulated binaural auralization trials the Bayes factor is 5.00, and for an equivalency of simIndivHRIRs and simHATS we found a Bayes factor of 0.45.

3.2.4.2 Subjective realism

Similar results could be found for the rating of realism. A linear mixed-effect model revealed a significant main effect of Audio Condition on subjective realism, $F(4, 71.3) = 29.91$, $p < .001$. Participants rated the subjective realism of the anchor lower ($M = 2.9$, $SD = 2.2$) than the subjective realism of loudspeakers ($M = 5.6$, $SD = 2.2$, $\beta = -2.77$, $t = -10.09$, $p < .001$), of measHATS ($M = 5.4$, $SD = 2.2$, $\beta = -2.55$, $t = -7.86$, $p < .001$), of simIndivHRIRs ($M = 5.6$, $SD = 2.1$, $\beta = -2.77$, $t = -9.77$, $p < .001$) and lower than simHATS ($M = 5.3$, $SD = 2.1$, $\beta = -2.43$, $t = -7.99$, $p < .001$). No other significant differences could be found in any other comparison between Audio Conditions. Following Bayes factors were found for the hypothesis of equivalency: of loudspeakers and plausible binaural auralizations - Bayes factor = 3.48, for measured and simulated Binaural auralizations - Bayes factor = 5.93, for simHATS and simIndivHRIRs - Bayes factor = 0.25.

3.3 Comparison of paradigms

Figure 7 shows the angle and distance deviance between the estimated and real sound source position as a function of Audio Condition and used sound source localization measurement paradigm. For the calculation of deviances between the estimated distance or angle and the real distance or angle, further processing of eye-tracking data was carried out. An outlier analysis was done within each found fixation and samples were excluded which were more than two standard deviations away from mean distance to participant. We then took the sample in which the gaze hit was the closest to the participants' position for further comparison. We used the same calculation methods as described for the placement paradigm.

We compared the results gained via the two different paradigms to measure sound source localization. The comparison must be interpreted cautiously since the

continuous data gained from the eye-tracking paradigm is unequivocal to the placement data. First, externalization rates were numerically higher (4.6%) in the placement paradigm. Front-back confusions could not be measured in the eye-tracking paradigm, since trials were repeated until no more fixation on the wall occurred (wall behind participants). Distance deviance seem to be numerically higher in the placement paradigm (about 61 cm higher), as well as angle deviance (about 4.9° more). On an individual participant level, an accordance between sound source localization accuracy in both paradigms was found, almost the same correlations were found for both outcome variables, see Figure 8. Angle localization accuracy in the first paradigm was correlated with angle deviance in the second paradigm, $r(47) = .42$, $p = .003$. Within the loudspeaker condition, a correlation of $r(47) = .51$, $p < .001$ was found, for measHATS a correlation of $r(47) = .37$, $p = .010$, for simIndivHRIRs $r(47) = .37$, $p = .010$, and for simHATS $r(47) = .39$, $p = .006$. Also the accuracy from both paradigms concerning distance estimation was correlated, $r(47) = .42$, $p = .003$. For loudspeaker trials no significant correlation was found, $r(47) = .18$, $p = .229$, whereas for measHATS a correlation of $r(47) = .38$, $p = .008$, for simIndivHRIRs $r(47) = .48$, $p < .001$, and for simHATS a correlation of $r(49) = .43$, $p = .002$, was found.

3.4 Exploratory analyses

3.4.1 Social anxiety, negative affect, and egocentric source distance

Participants' level of social anxiety was assessed. We analyzed whether social anxiety influences the perception of sound source distance. We used "social stimuli" – speech – which are fear-related in persons with social anxiety. The background of this analysis is that fear is known to have a significant influence on human perception, and also on distance perception [68]. However, we found no significant correlation between social anxiety, measured as the sum of the SPIN questionnaire, and perceived egocentric distance of sound sources. Furthermore, we analyzed whether positive or negative affect influenced perceived sound source distance, and again found no significant correlations.

3.4.2 Social presence and subjective realism

Figure 9 shows the correlations between social presence and realism ratings. In the placement paradigm, the ratings of a sound concerning social presence and realism were highly correlated, $r(47) = .76$, $p < .001$. For loudspeaker trials, a correlation of $r(47) = .53$ was found, for measHATS trials $r(47) = .79$, for simIndivHRIRs $r(47) = .85$, for simHATS $r(47) = .72$, $p < .001$, and for the anchor trials a correlation of $r(47) = .74$ was found, all $ps < .001$. Also during the eye-tracking paradigm, these two attributes were highly correlated, $r(47) = 0.87$, $p < .001$. In this task, for loudspeaker trials, a correlation of $r(47) = .74$, for measHATS trials $r(47) = .84$, for simIndivHRIRs trials $r(47) = .88$, for simHATS $r(47) = .83$, and for anchor trials $r(47) = .75$ was found, all $ps < .001$.

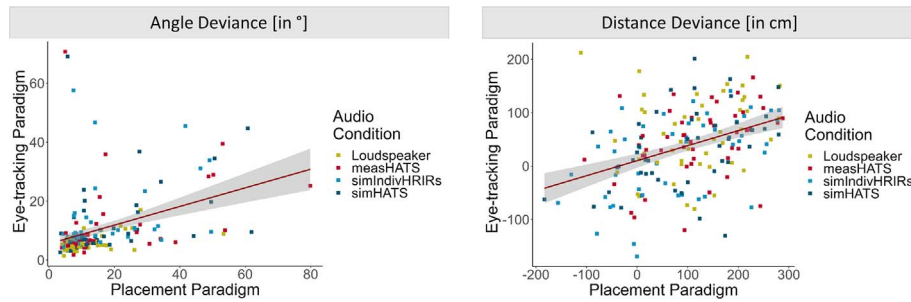


Figure 8. Correlation between placement paradigm (x -axis) and eye-tracking paradigm (y -axis) concerning sound source localization accuracy. On the left: Angle deviation in degree. On the right: Distance deviation in cm.

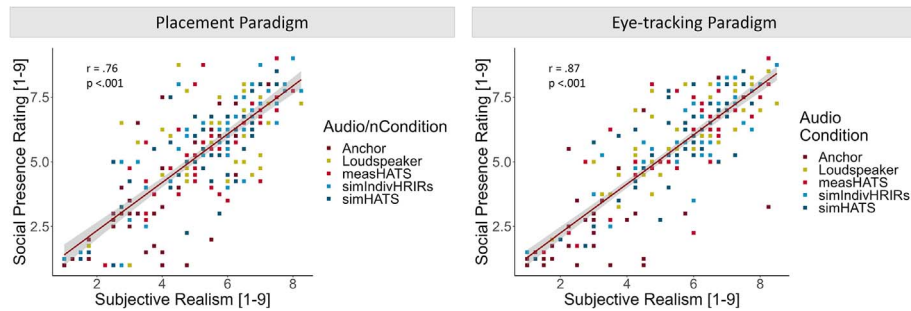


Figure 9. Correlation between social presence and realism rating during placement paradigm (on the left) and eye-tracking paradigm (on the right).

3.4.3 Learning effects

Referring to prior studies, we exploratory investigated whether the accuracy of sound source localization increases over time (placement paradigm). Again, anchor trials and invalid trials were excluded from analysis. A linear mixed model with trial number as fixed-effect and subject as random effect revealed a significant effect of the number of prior trials on distance deviance, $F(1, 3473.4) = 11.50$, $p < .001$. With an intercept of 108.58 and a $\beta = -0.32$, the distance deviance decreased over time. No effect of prior trials could be found on angle deviance.

3.4.4 Additional descriptive analyses

In the [Supplementary material](#), data can be found on follow-up questionnaires (e.g. difficulty of sound source localization, estimated amount of loudspeaker trials and loudspeakers). Furthermore, data on the elevation and loudness perception of Audio Conditions as well as measurement of sound pressure levels is provided in [Supplementary material](#). Shortly, there are perceptual differences between Audio Conditions, both for elevation and loudness perception. Last, data on hearing experience of the participants and its relationship with sound source localization accuracy can be found in [Supplementary material](#). Shortly, only years of practice playing a musical instrument were significantly negatively correlated with distance deviance.

4 Discussion

4.1 Placement paradigm

In the placement paradigm, participants positioned an agent at the estimated sound source. Sound source localization accuracy and subjective experience of plausible binaural auralizations were compared to loudspeakers and an anchor condition. In the context of this experiment, for successful localization it is needed that a sound is perceived externalized. Therefore, we first analyzed the rate of externalized trials per Audio Condition. Indicating plausibility of the used binaural auralizations, high rates were found. Furthermore, there was no significant difference between loudspeakers and binaural auralizations in rates of externalization. In the anchor condition, in over three out of four trials the sound was perceived inside the head (not externalized). This is unexpected since we used a state-of-the-art audio engine (Steam Audio v 4.1.4, Valve Corporation, Bellevue, WA, USA [42]). One possible explanation for this finding is the lack of headphone equalization (only) in the anchor condition. We used extra-aural headphones mounted to the HMD. This allowed a direct comparison to the loudspeaker without listeners being aware whether a headphone rendering was played or the real loudspeaker. For plausible binaural auralizations, we adjusted BRIRs regarding the headphone HMD position in terms of individually measured and computed individual headphone equalizations. This equalization could not be provided for the anchor, since the audio plugin is imple-

mented in the gaming engine. Hence, headphone equalization was found to be crucial e.g. for distance estimation, since an absence of equalization resulted in significantly more in-head localizations [69]. An alternative explanation for the surprisingly infrequent externalization of the anchor could be due to the experimental setup. We presented high-quality renderings beside the anchor and thus the contrast of the anchor to the simulations based on BRIRs which were precisely tailored toward the real room may have influenced the perception of the anchor. The BRIRs included room simulation [43], high quality HRIRs, and source directivity. The reverberation times of all conditions except the anchor were very similar (see Fig. 2) underlining again the sharp contrast to the anchor. It was shown that among several BRIR conditions using different late reverberations only for the most extreme modification (when no late reverberation cues were presented), significantly lower externalization rates were found [70]. The anchor uses ray tracing, which is normally used only for early reflections which could also explain low externalization rates [71]. Further experiments on context and contrast effects on externalization are being planned. However, due to the low rates of externalized trials, the anchor audio condition was excluded from further analyses on sound source localization. To sum up, the anchor differs from the other very similar conditions both in input parameters and on an output level in externalization and subjective experience. No conclusions can be drawn on the low acoustic performance of the anchor since this question was no goal of this study. The results can, nonetheless, be taken as a hint that subjective realism and social presence may be enhanced when spatial audio is perceived externalized.

Another fundamental characteristic of accurate sound source localization is the non-occurrence of front-back confusion. Here, a significant effect of Audio Condition was found. However, only for simHATS stimuli, higher rates of front-back confusion were found in comparison to the loudspeaker condition. For further analyses, only trials in which sounds were perceived externalized and no front-back-confusion occurred, were analyzed. It has to be mentioned that comparably high rates of front-back confusion and also confusion in the real loudspeaker condition were found. A possible explanation for this could be that the stimuli were in parts rather short (on average 1.5 s) and that our participants, naive to room acoustics, did not systematically use head-movements to localize sound source positions.

We were then interested in whether sound source localization of plausible binaural auralizations and loudspeakers were equivalent. Therefore, we analyzed the angle deviance between the real (or simulated) sound source and the estimated sound source as indicator for azimuthal localization. Concerning azimuthal localization, performance was not significantly affected by Audio Condition. We further gained evidence that all plausible binaural auralizations (measHATS, simIndivHRIRs, and simHATS) are equivalent concerning azimuthal localization via Bayes Factors. Nonetheless, we could not find support for an equivalency of real loudspeakers and binaural auralizations concerning azimuthal localization (Bayes Factor < 3).

We then analyzed whether sound distance perception differed between loudspeakers and plausible binaural auralizations. Unexpectedly, the distance deviance was significantly higher in the loudspeaker condition in comparison to binaural auralizations based on simulated BRIRs (both simIndivHRIRs and simHATS). The same was found for the measHATS condition. No difference between measHATS and loudspeakers was found. Furthermore, we gained evidence for the equivalency of simIndivHRIRs and simHATS concerning distance perception. Regarding the good correspondence in perceived azimuth/direction between different binaural auralizations compared to significant differences found in the perceived distance, this finding can be explained considering that perceived azimuth direction is dominated by the direct sound field component based on the precedence effect [72]. For distance perception, on the other hand, the source directivity, which is incorporated in the room simulations made for some auralizations, will influence the direct-to-reverberation ratio (DRR), which in turn could influence perceived distance. Indeed, there were variations in the DRR (see Tab. 1 for frequency-independent DRRs and Fig. 2 for energy-decay-curves) between the binaural auralizations. A detailed investigation of the interplay of DRRs and visual virtual scene and the effect on sound distance perception is outstanding. To interpret these results (lower distance accuracy in real sound source condition compared to simulated condition), it is also worth looking at our previous findings on distance perception in VR. We consistently found evidence for an overestimation of sound sources when using the visual presentation of a room via HMD [1, 32, 37]. One explanation could be the visual distance compression in VR. A visual virtual room is perceived as about a fourth smaller than a corresponding real room [26]. The acoustic perception of a sound source may interact with the compressed visual spatial impression and audiovisual integration may then lead to a distorted perception of distances. Furthermore, our measurement of localization – the placement task – could have contributed to these findings, since numerically higher overestimation was found in the placement paradigm compared to the eye-tracking paradigm. Assuming a visual compression of up to a quarter in VR, participants would have to place the agent much further away to set the desired distance. Based on the assumption that loudspeakers generate the desired sound source distance in the best possible way (and thus represent the baseline), this could mean that the simulated BRIR auralizations generated compressed distances (“too near”), which lead to a better agreement with the visually compressed distances. Furthermore, we found that participants improved the accuracy of distance estimations over time, which was not the case for azimuthal localization. Note that no feedback was given on potential source positions. Potentially, they learned to rely more on the acoustic cues over the course of time or adapted to the visual compression by using different spatial cues. This should be clarified in future studies. It should also be noted, that the majority of participants perceived all audio conditions (except the anchor) as elevated (Supplementary material). The measHATS BRIRs were perceived as most elevated.

These BRIRs were the only audio condition without different vertical orientations, which suggests a possible connection between head tracking and elevated perception of simulated sound sources. Nonetheless, no detrimental effect of the lack of elevation angles in measHATS were found on their azimuthal localizability.

Last, we examined the influence of the Audio Conditions on subjective experience in VR. Not surprisingly, given the high internalization rates, anchor condition stimuli were rated as lowest concerning realism, but also concerning social presence. Social presence was further rated higher in simIndivHRIRs in comparison to the simHATS condition. MeasHATS was found to be equivalent to simHATS and simIndivHRIRs concerning social presence ratings and subjective realism. Last, we gained evidence that for subjective realism, all three plausible binaural auralizations are equivalent to loudspeakers.

4.2 Eye-tracking paradigm

In this task, sound source localization was measured within a comparatively naturalistic paradigm using eye-tracking. We analyzed the rate of trials in which participants looked at the correct virtual agent (agent at sound source position) as a measure of sound source localization accuracy. Comparable to the placement paradigm, we found very low rates of externalization in anchor trials and consequently excluded this Audio Condition from further analyses on sound source localization accuracy.

Comparable to the placement paradigm, our hypotheses that the plausible binaural auralizations are equivalent to loudspeakers concerning sound source localization accuracy could not be confirmed. Higher rates of correct fixations (around 10% more) were found in loudspeaker trials. We further could not gain evidence for an equivalency of the three plausible binaural auralizations (Bayes Factors < 3), which is in contrast to the equivalency between auralizations which we found in the placement paradigm. However, no significant differences were found, indicating a need for further research.

When asking participants about their social presence in VR during the eye-tracking paradigm, an effect of audio condition was found. Participants again reported lower social presence in anchor trials. No other significant differences could be found. However, only for the hypothesis, that measured and simulated binaural auralizations are equivalent, evidence could be found.

Furthermore, differences between Audio Conditions also became apparent in the rating of realism. Again, anchor trials were rated worst. Here, we gained evidence for an equivalency of loudspeakers and all three plausible binaural auralizations concerning subjective realism and for the equivalency of binaural auralizations based on measured and simulated BRIRs. The equivalency of simulated binaural auralizations based on generic BRIRs and individual BRIRs could neither be confirmed nor disproved.

To conclude, the hypothesis that the plausible binaural auralizations (and loudspeakers) are superior to the anchor could be confirmed on all outcome variables. For sound

source localization, loudspeakers were in parts superior to all other conditions. Ambiguous evidence for equality between the plausible binaural auralizations was found. Yet, the binaural auralizations were equivalent to loudspeakers when it came to subjective realism in VR.

4.3 Implications

Our results provide evidence that the here investigated plausible binaural auralizations can be used to create an auditory impression that is very similar to that produced by a real sound source. These binaural auralizations include a room simulation using RAZR [43], accurately measured HRIRs, and acoustic source directivity. In terms of externalization, the renderings are not inferior to the loudspeakers. Nonetheless, loudspeakers are still superior when it comes to sound source localization accuracy. However, clear evidence was gained, that the loudspeakers and the three plausible renderings are equivalent in terms of subjective realism underpinning previous work that found authenticity and plausibility of speech auralizations. In this study, the investigated scene incorporated a visual digital twin of the simulated room, the real room itself, in which participants were present, and multiple real sound sources alongside simulated ones. Therefore, the equivalency concerning subjective realism hints towards transfer-plausibility of the used binaural auralizations [8, 73]. Additionally, binaural auralizations are not inferior to loudspeakers in terms of social presence in VR. One result that requires discussion is the better performance concerning distance estimation for simulated binaural auralizations in comparison to loudspeakers and measured BRIRs. As already briefly described above, one explanation could be that our simulated binaural auralizations generate distance impressions that seem nearer than intended. Potentially, using the measured room impulse responses the impression of distance could be created in a more realistic way than when using the simulated RIRs. In combination with visual distance compression when using an HMD, this could be an explanation for our findings. We used the HTC Vive Pro Eye as HMD. For the HTC Vive, a mean compression rate of 0.6 at a real distance of 5 m was reported [74]. This is a comparatively high compression rate. In another study, in which the plausible binaural auralizations were investigated and in which a room was also visually presented via HTC Vive, potentially similar evidence was found [51]. This implicates potential systematic influences of visual cues on auditory distance perception in VR (referring to ventriloquist effects [31]). Participants reported higher perceived reproduction quality for simulated binaural auralizations (in comparison to loudspeakers and measHATS). The source distance of loudspeakers and measHATS was rated as too distant. This is in line with our findings. Even more interestingly, the source distance for the simulated binaural auralizations was perceived on average as more precise than loudspeaker and measHATS. In contrast, Blau et al. [9] found no difference between measHATS and simulated binaural auralizations in terms of source distance or overall quality. While both studies used comparable BRIR sets, in the 2021

study, no visual virtual reality was presented. Here, no effect of visual compression should have occurred. To add, in the earlier study, participants should compare the binaural auralizations to a non-hidden loudspeaker reference. Furthermore, it is unknown whether participants rated the source distance of binaural auralizations as suboptimal because it is perceived as too close or too far. While the prior studies and this study used comparable binaural auralizations, measurement techniques, comparison conditions, and visual setup differed. More research is needed on audiovisual integration in VR and the influence of context and measurement techniques on source localization.

It should be noted that there is already research on ways to compensate for visual distance compression in VR. One approach used a similar mechanism, namely that audiovisual integration influences distance perception in VR, but in the opposite direction (manipulation of auditory distance to influence visual distance perception). In a recent study [75], the reverberation time of an auditory stimulus was manipulated to alter depth perception. Especially in the near field (up to 5 m), depth perception could be influenced by longer reverberation times. When very long reverberation times were used, participants put more trust in visual information and sensory segregation occurred. Compensation was also attempted by visual manipulations. For example, a kind of geometric minification [76] or the reduction of eye height [77] could decrease (visual) distance underestimation. An interesting and open research approach would be the investigation of the plausible binaural auralizations in comparison to loudspeakers in a visually non-compressed (or compensated) virtual environment.

Besides the question of how accurate audio-rendered sound sources can be localized in comparison to loudspeakers, we investigated whether the three binaural auralizations were equivalent. The measurement of BRIRs in a real room is time-consuming since many different head-above-torso orientations have to be adjusted and extensive acoustical technical equipment is needed. A further disadvantage is the limitation to existing rooms. Rendering of BRIRs of a not yet or not anymore existing room is not possible with this method. Auditory simulations of rooms based on simulated BRIRs can overcome these drawbacks. Furthermore, these simulated BRIRs can quickly be adjusted to new requirements (e.g. change of source positions). We gained evidence, that regarding subjective experience in VR, binaural auralizations based on BRIRs simulated using RAZR [43] are equivalent to measured BRIRs. Furthermore, in the placement paradigm, equivalency could be shown for azimuthal localization. However, distances were estimated significantly worse in terms of overestimation when measured BRIRs were used. Since comparable effects were found for loudspeakers, the same underlying mechanisms (see above) can be assumed. In the eye-tracking paradigm, no significant differences between measured and simulated BRIRs audio conditions were found. However, the evidence for equivalency was small, therefore we could not confirm the equivalency hypothesis. A further concern of this study was investigating how close binaural auralizations based on simulated

BRIRs in combination with generic HRIRs come to the quality of simulated BRIRs which are calculated using individually measured HRIRs. While within the placement paradigm, we gained clear evidence for an equivalency of simHATS and simIndivHRIRs concerning localization accuracy, within the eye-tracking paradigm, only small evidence was gained for the equivalency hypothesis. However, no significant differences were found between simHATS and simIndivHRIRs in any of the localization variables. Looking at both, these results and the effort of individually measured HRIRs, a valid assumption is that for many areas of application, especially when using speech stimuli, the cost-benefit analysis is in favor of using generic HRIRs.

As briefly stated above, an area of application is creating a convincing audiovisual scene for multiuser interaction or virtual exposure therapy. Here, accurate source localization can contribute to higher degrees of realism of the virtual scene. Another important goal besides high realism is a high degree of (social) presence [4]. Concerning the feeling of being with another person present in virtual reality (social presence), higher levels can be reached using plausible auralizations and loudspeakers in comparison to anchor stimuli. In the placement paradigm, the simIndivHRIRs stimuli were rated significantly higher concerning social presence than simHATS. In the eye-tracking paradigm, there is no evidence for the equivalency of simHATS and simIndivHRIRs. This can be seen as an indicator that using individually simulated BRIRs social presence in VR can be more enhanced than when using generic BRIRs (simHATS). Subjective rating of realism also differed between Audio Conditions. Here again, the anchor was inferior to all other Audio Conditions. Interestingly, social presence and subjective realism were strongly correlated. The correlation remained robust even when excluding the anchor condition (which prominently differed from the other conditions in terms of rating scores). This indicates that with a subjectively higher quality of the binaural auralization in a VR scene, higher levels of social presence can be induced (or vice versa). However, it is also possible that participants referred to a related construct, although the two rating items were formulated very differently. A comprehensive investigation of the interplay of subjective realism and social presence is pending, here a broader variety of items as well as inverted items should be implemented. It has to be pointed out, that no further validation of our rating items was conducted to investigate the internal reference to which participants of the main study referred to (only for the pilot study). Therefore, also implicit measures for social presence and subjective realism should be used and psychophysiological data could be included to gain a more robust evidence on this question.

We used two different paradigms to measure sound source localization accuracy of different Audio Conditions in VR. As can be seen in Figure 8, similar patterns for effects of Audio Conditions can be found independently of the measurement paradigm. This indicates towards validity of both paradigms. All in all, both, deviances in distance and angle estimation seem to be in trend higher in the

placement paradigm. Besides the “device” for measurement (controller vs. gaze), the difference between both paradigms is the (non-)existence of predefined offered positions. In the eye-tracking paradigm, 16 different plausible source positions (16 agents in the room) were offered. Contrarily, in the placement paradigm, source positions could be estimated and placed anywhere (on the xy plane). We assume the placement paradigm to be more sensitive to finding differences. The advantage of the eye-tracking paradigm lies in the assumed higher external validity, e.g. when the goal is the assessment of audiovisual quality of a virtual classroom. Since visual feedback on the adjusted source position was only provided in the placement task, higher accuracy can be assumed for the placement task in comparison to eye-tracking. In comparison to traditional pointing methods transferred to a VR setting (e.g. using controllers [28, 30]), by tendency lower levels of azimuthal accuracy were found either for the placement and the eye-tracking paradigm. This may be more due to methodological reasons (sample, outlier exclusion) than acoustic reasons (also for real loudspeaker). A further important difference of the here used placement paradigm in comparison to traditional pointing methods transferred to a VR setting is that a visual virtual room but no possible source positions were displayed, which could also have an effect on azimuthal accuracy. To sum up, both paradigms have drawbacks and merits, the choice should also depend on the intended context. Nonetheless, with both paradigms, we only measured sound source localization in terms of distance and angle estimation. Deviances in elevation perception were not assessed within the paradigms. An additional task on elevation perception revealed in part significant differences between Audio Conditions (see [Supplementary material](#)). This discrepancy must be taken into account. What else needs to be considered, is that we found differences in perceived loudness of the different Audio Conditions. Since loudness is an important factor for distance perception [33], this may have influenced sound source localization. However, only slightly different sound pressure levels were measured (see [Supplementary material](#)). The difference in loudness perception is possibly due to different presentations (via loudspeakers and headphones).

The contribution of a possibly higher measurement error when using gaze behavior instead of placing an agent is not yet clear. It is also noteworthy, that our sample consisted of non-experts in acoustics. Participants were not consistently able to perceive how many different sound sources had been placed in the room, the estimations ranged broadly (see [Supplementary material](#)). Furthermore, subjective assessment of own hearing capabilities was not relied to localization accuracy. Only for years of experience playing a musical instrument, a positive effect was found. To add, participants seem to progress during the tasks. The more trials they completed, the more accurate was localization. This is in line with previous studies on sound localization in three-dimensional virtual environments [32]. Not only that sound localization improve with duration of training [78], but localization training in VR can also

compensate for the shortcomings of generic HRIRs in comparison to individual HRIRs (e.g. [79]). This training effect is even more pronounced when head movements and therefore more information on head-tracked binaural auralizations are targeted. In the eye-tracking paradigm, in which 16 different plausible source positions were offered, less than a third of the fixations were on the correct position (in the best Audio Condition).

5 Conclusion

To conclude, with two different sound source localization paradigms we could show, that the three plausible binaural auralizations allow comparable accuracy. Nonetheless, localization accuracy was not as high as for real sound sources. Concerning externalization rate, social presence and subjective realism, the plausible binaural auralizations are comparable to loudspeakers. The anchor condition (state-of-the-art 3D audio for the Unreal gaming engine) was inferior in all investigated aspects. Of particular interest is the correlation between social presence and subjective realism. Overall, the findings suggest that advanced binaural auralizations may improve emotional processing by increased presence levels that facilitate more seamless social interactions [80]. Research in social or socioemotional contexts or in VR-based psychotherapy can benefit from this. Taking the cost-benefit ratio into account, binaural auralizations based on simulated BRIRS using RAZR [43] in combination with generic HRIRs (simHATS) are considered the most recommendable of the here investigated three plausible binaural auralizations (measHATS, simIndivHRIRs, simHATS) for typical research in the clinical-psychological or other applied psychological fields, as long as the dominant auditory stimulus type is human speech.

Acknowledgments

We would like to especially thank Marieke Bruckmann and Nora Schmid for their help with data acquisition and Andreas Ruider and Alexander May for their technical support.

Funding

This work is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under the project ID 422686707, SPP2236 – AUDICTIVE – Auditory Cognition in Interactive Virtual Environments [81].

Conflicts of interest

The authors declare that there is no conflict of interest.

Data availability statement

The research data associated with this article are available in the public data repository (open science framework – osf) under the references [82, 83]. All scripts required for data analyses as well as the preregistration for hypotheses and statistical tests can be found.

Supplementary material

Supplementary material for this article is available at <https://acta-acustica.edpsciences.org/10.1051/aacus/2024064/olm>.

References

1. S. Roßkopf, L.O.H. Kroczeck, F. Stärz, M. Blau, S. Van de Par, A. Mühlberger: Comparable sound source localization of plausible auralizations and real sound sources evaluated in a naturalistic eye-tracking task in virtual reality, in: Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023, Turin, Italy, 11–15 September, European Acoustics Association, 2024, pp. 1485–1492.
2. T.F. Wechsler, F. Kämpers, A. Mühlberger: Inferiority or even superiority of virtual reality exposure therapy in phobias? – A systematic review and quantitative meta-analysis on randomized controlled trials specifically comparing the efficacy of virtual reality exposure to gold standard in vivo exposure in agoraphobia, specific phobia, and social phobia, *Frontiers in Psychology* 10 (2019) 1758.
3. E. Carl, A.T. Stein, A. Levihn-Coon, J.R. Pogue, B. Rothbaum, P. Emmelkamp, G.J.G. Asmundson, P. Carlbring, M.B. Powers: Virtual reality exposure therapy for anxiety and related disorders: A meta-analysis of randomized controlled trials, *Journal of Anxiety Disorders* 61 (2019) 27–36.
4. M. Slater: Immersion and the illusion of presence in virtual reality, *British Journal of Psychology* 109 (2018) 431–433.
5. J. Diemer, G.W. Alpers, H.M. Peperkor, Y. Shiban, A. Mühlberger: The impact of perception and presence on emotional reactions: a review of research in virtual reality, *Frontiers in Psychology* 6 (2015) 1–9.
6. W.M. Felton, R.E. Jackson: Presence: a review, *International Journal of Human-Computer Interaction* 38 (2022) 1–18.
7. K. Brandenburg, F. Klein, A. Neidhardt, U. Sloma, S. Werner: Creating auditory illusions with binaural technology, in: J. Blauert, J. Braasch (eds), *The technology of binaural understanding*, Springer International Publishing, Cham, 2020, pp. 623–663.
8. A. Neidhardt, C. Schneiderwind, F. Klein: Perceptual matching of room acoustics for auditory augmented reality in small rooms – literature review and theoretical framework, *Trends in Hearing* 26 (2022).
9. M. Blau, A. Budnik, M. Fallahi, H. Steffens, S.D. Ewert, S. Van de Par: Toward realistic binaural auralizations–perceptual comparison between measurement and simulation-based auralizations and the real room for a classroom scenario, *Acta Acustica* 5 (2021) 8.
10. F. Stärz, L.O.H. Kroczeck, S. Rosskopf, A. Mühlberger, M. Blau: Perceptual comparison between the real and the auralized room when being presented with congruent visual stimuli via a head-mounted display, in: Proceedings of the 24th International Congress on Acoustics. International Commission for Acoustics (ICA), Gyeongju, Korea, 24–28 October, 2022.
11. S. Poeschl, K. Wall, N. Doering: Integration of spatial sound in immersive virtual environments an experimental study on effects of spatial sound on presence, in: 2013 IEEE Virtual Reality (VR), Lake Buena Vista, FL, USA, 18–20 March, IEEE, 2013, pp. 129–130.
12. P. Skalski, R. Whitbred: Image versus sound: A comparison of formal feature effects on presence and video game enjoyment, *PsychNology Journal* 8 (2010) 67–84.
13. K. Bormann: Presence and the utility of audio spatialization, *Presence: Teleoperators & Virtual Environments* 14 (2005) 278–297.
14. J. Freeman, J. Lessiter: Here, there and everywhere: the effects of multichannel audio on presence, in: Proceedings of the 2001 International Conference on Auditory Display, Espoo, Finland, July 29–August 1 (2001) 231–234.
15. F. Biocca, C. Harms, J.K. Burgoon: Toward a more robust theory and measure of social presence: review and suggested criteria, *Presence: Teleoperators and Virtual Environments* 12 (2003) 456–480.
16. C.S. Oh, J.N. Bailenson, G.F. Welch: A systematic review of social presence: definition, antecedents, and implications, *Frontiers in Robotics and AI* 5 (2018) 409295.
17. P.M.G. Emmelkamp, K. Meyerbröcker, N. Morina: Virtual reality therapy in social anxiety disorder, *Current Psychiatry Reports* 22 (2020) 1–9.
18. F. Brinkmann, A. Lindau, S. Weinzierl: On the authenticity of individual dynamic binaural synthesis, *Journal of the Acoustical Society of America* 142 (2017) 1784–1795.
19. L. Chen, J. Vroomen: Intersensory binding across space and time: a tutorial review, *Attention, Perception, & Psychophysics* 75 (2013) 790–811.
20. C. Guezenoc, R. Segurier: HRTF individualization: a survey, in: 145th Audio Engineering Society Convention 2018, New York, NY, USA, 17–20 October, 2018.
21. H. Möller, M.F. Sørensen, C.B. Jensen, D. Hammershøi: Binaural technique: do we need individual recordings?, *Journal of the Audio Engineering Society* 44 (1996) 451–469.
22. D.R. Begault, E.M. Wenzel, M.R. Anderson: Direct comparison of the impact of head tracking, reverberation, and individualized head-related transfer functions on the spatial perception of a virtual speech source, *Journal of the Audio Engineering Society* 49 (2001) 904–916.
23. C. Jenny, C. Reuter: Usability of individualized head-related transfer functions in virtual reality: empirical study with perceptual attributes in sagittal plane sound localization, *JMIR Serious Games* 8 (2020) e17576.
24. L. Prud’homme, M. Lavandier: Do we need two ears to perceive the distance of a virtual frontal sound source? *Journal of the Acoustical Society of America* 148 (2020) 1614–1623.
25. H.S. Braren, J. Fels: Towards child-appropriate virtual acoustic environments: a database of high-resolution HRTF measurements and 3D-scans of children, *International Journal of Environmental Research and Public Health* 19 (2021) 324.
26. R.S. Renner, B.M. Velichkovsky, J.R. Helmer: The perception of egocentric distances in virtual environments—a review, *ACM Computing Surveys* 46 (2013) 1–40.
27. L.E. Buck, M.K. Young, B. Bodenheimer: A comparison of distance estimation in HMD-based virtual environments with different HMD-based conditions, *ACM Transactions on Applied Perception* 15 (2018) 1–15.
28. A. Ahrens, K.D. Lund, M. Marschall, T. Dau: Sound source localization with varying amount of visual information in virtual reality, *PLoS One* 14 (2019) e0214603.
29. D. Poirier-Quinot, M.S. Lawless: Impact of wearing a head-mounted display on localization accuracy of real sound sources, *Acta Acustica* 7 (2023) 3.
30. T. Huisman, A. Ahrens, E. MacDonald: Ambisonics sound source localization with varying amount of visual information in virtual reality, *Frontiers in Virtual Reality* 2 (2021) 722321.
31. C.E. Jack, W.R. Thurlow: Effects of degree of visual association and angle of displacement on the “ventriloquism” effect, *Perceptual and Motor Skills* 37 (1973) 967–979.

32. S. Roßkopf, L.O. Kroczeck, F. Stärz, M. Blau, S. Van de Par, A. Mühlberger: The effect of audio-visual room divergence on the localization of real sound sources in virtual reality, in: Fortschritte der Akustik, DAGA, Hamburg, 2023, pp. 1431–1434.
33. P. Zahorik, D.S. Brungart, A.W. Bronkhorst: Auditory distance perception in humans: a summary of past and present research, *Acta Acustica United with Acustica* 91 (2005) 409–420.
34. P. Maruhn, S. Schneider, K. Bengler: Measuring egocentric distance perception in virtual reality: Influence of methodologies, locomotion and translation gains, *PLoS One* 14 (2019) e0224651.
35. H. Adams, J. Stefanucci, S. Creem-Regehr, B. Bodenheimer: Depth perception in augmented reality: the effects of display, shadow, and position, in: 2022 IEEE Conference on Virtual Reality and 3D User Interfaces (VR), Christchurch, New Zealand, 12–16 March, IEEE, 2022, 792–801.
36. P.E. Etchemendy, I. Spiouas, E.R. Calcagno, E. Abregú, M.C. Eguia, R.O. Vergara: Direct-location versus verbal report methods for measuring auditory distance perception in the far field, *Behavior Research Methods* 50 (2018) 1234–1247.
37. L.O.H. Kroczeck, S. Roßkopf, F. Stärz, S. Van de Par, A. Mühlberger: The influence of affective voice on sound distance perception, in: Fortschritte der Akustik, DAGA, Stuttgart, 2022, pp. 1163–1166.
38. T. Foulsham, L.A. Sanderson: Look who’s talking? Sound changes gaze behaviour in a dynamic social scene, *Visual Cognition* 21 (2013) 922–944.
39. J.C. Middlebrooks, Z.A. Onsan: Stream segregation with high spatial acuity, *Journal of the Acoustical Society of America* 132 (2012) 3896–3911.
40. R.K. Maddox, D.A. Pospisil, G.C. Stecker, A.K. Lee: Directing eye gaze enhances auditory spatial cue discrimination, *Current Biology* 24 (2014) 748–752.
41. R. Schleicher, S. Spors, D. Jahn, R. Walter: Gaze as a measure of sound source localization, in: AES 38th International Conference, Piteå, Sweden, June 13–15, Audio Engineering Society, 2010, pp. 1–5.
42. V. Corporation, 2022. Available at <https://valvesoftware.github.io/steam-audio/doc/capi/index.html>.
43. T. Wendt, S. Van de Par, S. Ewert: A computationally-efficient and perceptually-plausible algorithm for binaural room impulse response simulation, *Journal of the Audio Engineering Society* 62 (2014) 748–766.
44. L. Kroczeck, S. Roßkopf, F. Stärz, A. Ruider, M. Blau, S. Van de Par, A. Mühlberger: A room of one’s own: a high-accuracy calibration procedure to align spatial dimensions between a virtual and a real room, 2023. Available at <https://psyarxiv.com/erb49/>.
45. A. Lindau, H.-J. Maempel, S. Weinzierl: Minimum BRIR grid resolution for dynamic binaural synthesis, *Journal of the Acoustical Society of America* 123 (2008) 3498.
46. A. Lindau, S. Weinzierl: On the spatial resolution of virtual acoustic environments for head movements in horizontal, vertical, and lateral direction, in: Proceedings of the EAA Symposium on Auralization, Espoo, Finland, 15–17 June, 2009.
47. F. Denk, F. Brinkmann, A. Stirnemann, B. Kollmeier: The PIRATE: an anthropometric earplug with exchangeable microphones for individual reliable acquisition of transfer functions at the ear canal entrance, in: Fortschritte der Akustik, DAGA, Rostock, 2019, pp. 18–21.
48. F. Stärz, L.O.H. Kroczeck, S. Roßkopf, A. Mühlberger, S. Van de Par, M. Blau: Mounting extra-aural headphones to a head-mounted display using a 3D-printed support, in: Fortschritte der Akustik, DAGA, Hamburg, 2023, pp. 1636–1639.
49. C. Schneiderwind, A. Neidhardt, M. Dominik: Comparing the effect of different open headphone models on the perception of a real sound source, in: Audio Engineering Society Conference (2020), AES International Conference on Audio for Virtual and Augmented Reality. Audio Engineering Society, 2020.
50. P. Lladó, T. McKenzie, N. Meyer-Kahlen, S.J. Schlecht: Predicting perceptual transparency of head-worn devices, *Journal of the Audio Engineering Society* 70 (2022) 585–600.
51. F. Stärz, L.O.H. Kroczeck, S. Roßkopf, A. Mühlberger, S. Van de Par, M. Blau: Comparing room acoustical ratings in an interactive virtual environment to those in the real room, in: Proceedings of the 10th Convention of the European Acoustics Association Forum Acusticum 2023, Turin, Italy, 11–15 September, European Acoustics Association, 2024, pp. 5009–5016.
52. H. Funk, C. Kuhn, L. Nielsen, K. Rische, B. Lex, B. Redecker, B. Winzer-Kiontke: Studio“ 21”: das Deutschbuch: Deutsch als Fremdsprache, Cornelsen Verlag, Berlin, 2013.
53. MathWorks: Audio toolbox user’s guide, 2022.
54. Z. Sosic, U. Gieler, U. Stangier: Screening for social phobia in medical in-and outpatients with the German version of the Social Phobia Inventory (SPIN), *Journal of Anxiety Disorders* 22 (2008) 849–859.
55. H.W. Krohne, B. Egloff, C.-W. Kohlmann, A. Tausch: Untersuchungen mit einer deutschen version der” positive and negative affect schedule” (PANAS), *Diagnostica* 42 (1996) 139–156.
56. G. Makransky, L. Lilleholt, A. Aaby: Development and validation of the Multimodal Presence Scale for virtual reality environments: a confirmatory factor analysis and item response theory approach, *Computers in Human Behavior* 72 (2017) 276–285.
57. S. Weinzierl, S. Lepa, D. Ackermann: A measuring instrument for the auditory perception of rooms: the Room Acoustical Quality Inventory (RAQI), *Journal of the Acoustical Society of America* 144 (2018) 1245–1257.
58. Vive.com: VIVE Pro Eye Support. Available at <https://www.vive.com/us/support/vive-pro-eye/> (accessed July 15, 2024).
59. D.D. Salvucci, J.H. Goldberg: Identifying fixations and saccades in eye-tracking protocols, in: Proceedings of the 2000 Symposium on Eye Tracking Research & Applications – ETRA 00, Palm Beach Gardens, FL, USA, 6–8 November, Association for Computing Machinery, New York, NY, USA, 2000, pp. 71–78.
60. R.T. Development Core : R: A language and environment for statistical computing, 2019. Available at <https://www.R-project.org/> (accessed February 6, 2024).
61. D. Bates, M. Mächler, B. Bolker, S. Walker: Fitting linear mixed-effects models using lme4, *Journal of Statistical Software* 67 (2015) 1–48.
62. A. Kuznetsova, P. B. Brockhoff, R. H. B. Christensen: lmerTest Package: tests in linear mixed effects models, *Journal of Statistical Software* 82 (2017) 1–26.
63. S.R. Searle, F.M. Speed, G.A. Milliken: Population marginal means in the linear model: an alternative to least squares means, *American Statistician* 34 (1980) 216–221.
64. R.D. Morey, J.N. Rouder: BayesFactor version 0.9. 9: an R package for computing Bayes factor for a variety of psychological research designs, 2014.
65. S. Holm: A simple sequentially rejective multiple test procedure, *Scandinavian Journal of Statistics* 6 (1979) 65–70.

66. E.-J. Wagenmakers: A practical solution to the pervasive problems OFP values, *Psychonomic Bulletin & Review* 14 (2007) 779–804.
67. H. Jeffreys: *The theory of probability*, Oxford University Press, Oxford, 1998.
68. K.T. Gagnon, M.N. Geuss, J.K. Stefanucci: Fear influences perceived reaching to targets in audition, but not vision, *Evolution and Human Behavior* 34 (2013) 49–54.
69. K. Sunder, E.-L. Tan, W.-S. Gan: Effect of headphone equalization on auditory distance perception, in: *Audio Engineering Society Convention 137*. Audio Engineering Society, 2014.
70. C. Schneiderwind, M. Richter, N. Merten, A. Neidhardt: Effects of modified late reverberation on audio-visual plausibility and externalization in AR, in: *2023 Immersive and 3D Audio: from Architecture to Automotive (I3DA)*, Bologna, Italy, 5–7 September, IEEE, 2023, pp.1–9.
71. M. Cuevas-Rodríguez, L. Picinali, D. González-Toledo, C. Garre, E. de la Rubia-Cuestas, L. Molina-Tanco, A. Reyes-Lecuona: 3D Tune-In Toolkit: an open-source library for real-time binaural spatialisation, *PLoS One* 14 (2019) e0211899.
72. J. Blauert: *Spatial hearing: the psychophysics of human sound localization*, The MIT Press, Cambridge, Massachusetts, London, England, 1997.
73. S. Wirler, N. Meyer-Kahlen, S. Schlecht: Towards transfer-plausibility for evaluating mixed reality audio in complex scenes, in: *Audio Engineering Society Conference (2020), AES International Conference on Audio for Virtual and Augmented Reality*. Audio Engineering Society, 2020.
74. L. Buck, R. Paris, B. Bodenheimer: Distance compression in the HTC vive pro: a quick revisit of resolution, *Frontiers in Virtual Reality* 2 (2021) 728667.
75. Y.-H. Huang, R. Venkatakrishnan, R. Venkatakrishnan, S.V. Babu, W.-C. Lin: Using audio reverberation to compensate distance compression in virtual reality, in: *ACM Symposium on Applied Perception 2021*, Virtual Event France, 16–17 September, Association for Computing Machinery, New York, NY, USA, 2021, pp. 1–10.
76. B. Li, R. Zhang, A. Nordman, S.A. Kuhl: The effects of minification and display field of view on distance judgments in real and HMD-based environments, in: *Proceedings of the ACM SIGGRAPH Symposium on Applied Perception*, Tübingen Germany, 13–14 September, Association for Computing Machinery, New York, NY, 2015, pp. 55–58.
77. M. Leyrer, S.A. Linkenauer, H.H. Bülthoff, B.J. Mohler: Eye height manipulations: a possible solution to reduce underestimation of egocentric distances in head-mounted displays, *ACM Transactions on Applied Perception* 12 (2015) 1–23.
78. C. Rajguru, G. Brianza, G. Memoli: Sound localization in web-based 3D environments, *Scientific Reports* 12 (2022) 12107.
79. M.A. Steadman, C. Kim, J.-H. Lestang, D.F.M. Goodman, L. Picinali: Short-term effects of sound localization training in virtual reality, *Scientific Reports* 9 (2019) 18284.
80. M. Pfaller, L.O. Kroczeck, B. Lange, R. Fülöp, M. Müller, A. Mühlberger: Social presence as a moderator of the effect of agent behavior on emotional experience in social interactions in virtual reality, *Frontiers in Virtual Reality* 2 (2021) 741138.
81. DFG project homepage. Available at <https://gepris.dfg.de/gepris/projekt/422686707> (accessed February 19, 2024).
82. S. Roßkopf, A. Mühlberger, L.O.H. Kroczeck: The effect of audio rendering on sound source localization, social presence and perceived audio quality in virtual reality, 2023. Available at <https://doi.org/10.17605/OSF.IO/9YQF7>.
83. S. Roßkopf, A. Mühlberger, L. O. H. Kroczeck: Using an eyetracking task to evaluate the influence of binaural audio renderings on sound source localization in virtual reality, 2023. Available at <https://doi.org/10.17605/OSF.IO/2Q4Y3>.

Cite this article as: Roßkopf S. Kroczeck LOH. Stärz F. Blau M. Van de Par S, et al. 2024. The impact of binaural auralizations on sound source localization and social presence in audiovisual virtual reality: converging evidence from placement and eye-tracking paradigms. *Acta Acustica*, 8, 72. <https://doi.org/10.1051/aacus/2024064>.