OPEN ACCESS Freely available online

P. N. Porysek Moreta, S. Bech, J. Francombe, J. Østergaard, and S. van de Par,
"Identifying Principal Attributes for Evaluating Audio Quality
of Reproduction Systems With Spatially Dynamic Program Material,"
J. Audio Eng. Soc., vol. 72, no. 9, pp. 552–570 (2024 Sep.).
https://doi.org/10.17743/jaes.2022.0168.

Identifying Principal Attributes for Evaluating Audio Quality of Reproduction Systems With Spatially Dynamic Program Material

PIA NANCY PORYSEK MORETA,[1, 2, *] *AES Student Member*, **SØREN BECH,**[1, 2] *AES Fellow*,
(pipo@bang-olufsen.dk) (sbe@bang-olufsen.dk)

JON FRANCOMBE,[1] **JAN ØSTERGAARD,**[2] **AND STEVEN VAN DE PAR,**[3] *AES Member*
(jofr@bang-olufsen.dk) (jo@es.aau.dk) (steven.van.de.par@uni-oldenburg.de)

[1] *Bang & Olufsen a/s, 7600 Struer, Denmark*
[2] *Aalborg University, Department of Electronic Systems, 9220 Aalborg, Denmark*
[3] *Carl von Ossietzky University of Oldenburg, Department of Medical Physics and Acoustics, 26129 Oldenburg, Germany*

Despite the significant advances in spatial audio technologies, there is a lack of knowledge regarding the impact of spatially dynamic program material on known descriptive attributes of basic audio quality (BAQ). This paper presents an exploratory study that investigates the perception of sound quality in reproduced sound using spatially dynamic program material in comparison with stationary program material. Three reproduction systems (stereo, surround, and surround with height) were compared across two different spatial variations of the same complex audio scene. Using a paired comparison approach combined with the check-all-that-apply (CATA) method, differences in BAQ ratings were collected alongside attributes contributing to the decision of BAQ rating. The study reveals that audio quality and its underlying attributes are primarily impacted by differences in reproduction systems, with spatially dynamic program material showing no direct influence. However, there is slight evidence indicating that the use of spatially dynamic program material may influence the selection of attributes for comparisons of different reproduction systems, leading to less clarity in results. Furthermore, this study can serve as a practical guide as it lists attributes for various loudspeaker comparisons and contextualizes their impact on quality ratings.

0 INTRODUCTION

The objective of this work is to investigate the impact of spatially dynamic program material on the perception of sound quality in reproduced sound. The term "dynamic program material" refers to content distinguished by perceptible changes over time, with a specific focus on spatial alterations, such as the movement of sound sources. Typically, the evaluation of sound reproduction systems relies on nondynamic program material or systems are evaluated with dynamic program material using experimental methods that might not be suitable for use with such program material because of potential biases [1]. There is a lack of previous studies examining the influence of dynamic program material on the evaluation of audio quality in reproduction systems in a controlled environment. This gap in knowledge is particularly noteworthy considering the emergence of "next generation spatial audio technologies," which offer new possibilities for integrating spatially dynamic elements into applications such as audio dramas, TV sports broadcasts [2], and augmented/virtual reality.

To evaluate audio reproduction—whether to assess the quality of a new set of loudspeakers or the performance of an enhancement algorithm—perceptual evaluation is still considered the gold standard [3]. To facilitate perceptual evaluation, standardized procedures are available [4, 5].

Audio program material is required to test the performance of the system, no matter which element of the reproduction chain needs evaluation, and the selection of representative audio material is crucial for obtaining relevant results [4–6]. At the same time, this selection is also subject to certain limitations to avoid bias and errors. Stimuli are often selected to be as low in complexity and short as possible and are chosen based on perceptual constancy [7]. Dynamic aspects, such as spatially dynamic program material including movement, are avoided because they are

*To whom correspondence should be addressed, email: pipo@bang-olufsen.dk. Last updated: August 2, 2024.

considered a cause of confusion [7]. However, recent advancements in production [8, 9], transmission [10], and reproduction technologies for 3D audio (e.g., Dolby Atmos, Sony's 360 Reality Audio format), lead to a growing interest in evaluating these new technologies in conjunction with more complex audio material that may include moving sources.

Perceptual attributes are used to characterize perception and provide more detailed information on audio quality. Numerous studies in the literature have identified various attributes that contribute to the perceived quality of reproduced audio (e.g., [11–13]). There are different methods that are used for attribute elicitation, with outcomes influenced by the choice of reproduction systems and sound material used to exemplify perceptual differences. When analyzing the stimuli selection that was used for elicitation experiments found in the literature (see Mason et al. [14] for an overview), it can be observed that—along with the development toward higher channel reproduction evaluation—there has been an evolution toward the use of more complex and ecologically valid program material. Although earlier research determining the perceptual attributes for reproduced spatial audio used simpler spatial sound events (e.g., speech and isolated guitar in varying rooms) [15], more recent studies used music [16, 17] and ambience scenes [18, 17].

These types of stimuli, while covering some characteristics of current sound reproduction, do not cover the entire spectrum of use case scenarios or technical capabilities. Occasionally, different types of stimuli covering a higher variety of spatially dynamic sound, such as more experimental 3D audio music [13], have been used. In other cases, it is hard to determine the complexity and composition of the used stimuli, as they are not described in detail.

No publicly available previous study has conducted a controlled comparison between stationary and dynamic program material, thus leaving unexplored the influence of spatially dynamic program material on attribute selection. Results of an exploratory study conducted by Gil et al. [19] using a methodology of continuous evaluation (temporal check-all-that-apply [TCATA] [20]) with four perceptual attributes show that some attributes were only selected when the tested program material was spatially dynamic. These findings suggest that certain attributes may hold greater significance than others when employed for evaluating reproduction systems with spatially dynamic audio program material.

Building on these observations, the objective of this paper is to address the following research questions:

a) What are the most important attributes for characterizing the sound quality of reproduction systems when using spatially dynamic program material?
b) Do the important attributes differ between usage of stationary and dynamic program material?

To answer these research questions, a listening experiment was designed. A more detailed description of the considerations and practical details for designing the listening experiment is given in SEC. 1, followed by SEC. 2, where the results of the experiment are presented. Finally, in SEC. 3, the results are discussed and the research questions answered.

1 METHODOLOGY

The following section will outline the experimental design of the listening experiment. Two spatial variations of the same audio scene were created: one stationary and one dynamic. These variations were then used as program material to evaluate the perceived quality difference between three loudspeaker layouts, each featuring a different number of loudspeakers. Different loudspeaker setups were used to create differences in spatial reproduction capabilities. Using more loudspeakers results in a broader and more prominent audio source movement, particularly in terms of width and height dimensions, which is typically linked to improved audio quality [21, 22].

A paired comparison method was applied to compare the reproduction systems. Spatially dynamic and stationary program material were not directly compared by the participants, as the focus of this study was to investigate the influence of spatial dynamic content on the attribute selection describing quality differences and not to detect attributes associated with a change from stationary to dynamic content.

The response variables were basic audio quality (BAQ) difference ratings and the attributes assumed to influence the observed quality differences. The check-all-that-apply (CATA) method [23, 21] was used to investigate which attributes from an available list are perceivable. This method proves particularly useful when there is a known list of generally fitting attributes that needs to be reduced to potential key attributes for a specific use case. As mentioned in SEC. 0, several lists of elicited attributes are available in the literature for evaluating reproduced sound. None of the individual lists included all the attributes that were potentially relevant to this study. Therefore, a pre-experiment was conducted to determine an initial list of attributes to be used in the main experiment.

The following sections will give more details about the materials and methodology of the conducted experiment, starting with the basic information on the physical setup (SEC. 1.1), the creation of stimuli (SEC. 1.2), loudness matching (SEC. 1.3), the selection of initial attributes (SEC. 1.4), and the methodology of the main experiment (SEC. 1.5).

1.1 Physical Setup

An anechoic chamber (6.07 m × 4.95 m × 2.91 m) was used to minimize impact of the reproduction room's acoustics on the evaluated sound field [24, p. 246 f.]. The loudspeakers were arranged in a sphere with a radius of 1.35 m around the listener position. To include a variety of standard reproduction systems used for modern domestic spatial audio reproduction, three loudspeaker layouts (Genelec 8320A and 8331A) according to ITU-R BS. 2051 [25])

Table 1. Loudspeaker layouts and their respective settings of azimuth and elevation.

Loudspeaker layout	Channel	Azimuth	Elevation
0+2+0	Left/Right	±30	0
0+5+0	Left/Right	±30	0
	Center	0	0
	Left/Right side surround	±90	0
4+7+0	Left/Right	±30	0
	Center	0	0
	Left/Right surround	±120	0
	Left/Right rear surround	±135	0
	Left/Right rear surround top front	±45	30
	Left/Right rear surround top back	±135	30

were selected: a 3D surround layout (4+7+0), a 2D surround layout (0+5+0), and a standard stereo layout (0+2+0). Azimuth and elevation coordinates for all layouts are given in Table 1. The loudspeaker array was calibrated as described in Kaplanis et al. [26].

The listening experiment was carried out using an interface created in Cycling 74's *Max* software. The interface was running on a desktop PC (OMEN X 900, HP) in the control room, which was linked to the anechoic chamber's loudspeaker sphere via an audio interface (RME M-32 DA M-16 DA). The audio stimuli that had been pre-processed and rendered were then evaluated using the test interface. The interface was mirrored from the control room monitor screen to a screen within the experimental environment. The screen was placed outside the loudspeaker sphere, approximately 1.4 m away from the listener position, at a comfortable reading height. The participants were seated in the center of the sphere on a fixed chair and could access the interface using the provided mouse and keyboard on a small lap desk.

1.2 Stimuli Creation

Two versions of the same program material were created with different spatial properties. The first was dynamic. The second was stationary, where the same audio objects remained in their start positions. For this, an 18-s-long excerpt of the audio drama scene *The Protest* from the *S3A Object-Based Audio Drama dataset* [27] was selected. This excerpt is part of a protest scene outside a bank and is composed of different sound sources like crowd chatter, a person talking through a megaphone, various background noises (e.g., a horse, breaking glass), and music. This scene included considerable horizontal and vertical movement patterns that could be manipulated using the given metadata. The editing of the dataset was done using the audio workstation *REAPER* in combination with the VST plugin *EAR Production Suite* [28]. To generate the two versions of spatial variation, the movement patterns of the mix were preserved and enhanced for the dynamic version of the scene, while the movement patterns were removed to create the stationary version.

Channels that contained significant audible movement but were not controllable through the provided metadata were excluded. This included reverberation channels and two channels that contained the signal of a helicopter as, despite removing the movement paths, the sound source of the helicopter still conveyed a sense of movement. By excluding these channels, the study aimed to focus specifically on the controlled spatial dynamic elements within the reproduced audio.

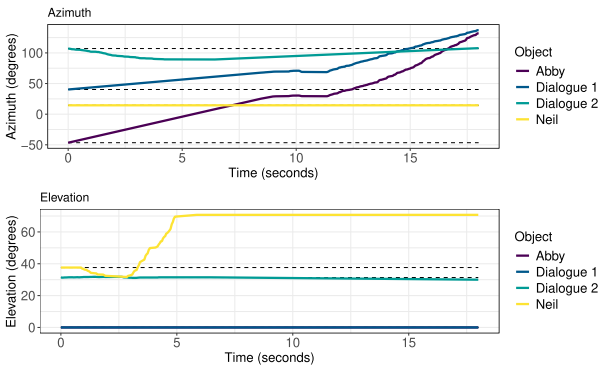


Fig. 1. The movement paths (azimuth and elevation) of four tracks illustrating the most prominent changes in audio object positions within the audio excerpt from the scene titled *The Protest*.

Four channels containing speech and the most pronounced movement patterns within the scene are visualized in Fig. 1. These patterns primarily involve horizontal motion of a crowd of protesters, moving from the listener's right side to the left side (Abby, Dialogue 1, Dialogue 2). One sound source exhibits a vertical movement (Neil, a male speaker). The two resulting scenes (dynamic/stationary) were then each rendered for all three selected reproduction systems at a sampling frequency of 48 kHz. For offline rendering, the *ITU audio definition model renderer* [29] was used. This resulted in six stimuli (two spatial variations of the program item for each of three reproduction systems).

1.3 Loudness Matching

Loudness matching was performed in order to avoid bias in the results caused by loudness differences in the material [30]. The Zwicker method according to ISO 532-1 [31] was used to obtain loudness estimates over the whole

length of each stimulus. A model was used because a) a loudness-matching listening experiment for complex stimuli was considered a complicated task for the listeners and b) an experiment by Francombe et al. [32] facing a similar situation in a loudness-matching experiment showed a non-significant difference in performance of available psychoacoustic loudness models compared with perceptual listening tests.

The target loudness was set so that after applying the loudness matching, the measured equivalent A-weighted sound pressure levels ($L(a)eq$) for each stimulus ranged from 76.1 dB to 77.5 dB measured at the listening position. This indicates that the level range for the test stimuli fell within the range of preferred or most comfortable listening levels for music (as an example, for complex stimuli), which typically range from 64.8 dB to 85.0 dB [24].

1.4 CATA Attribute Selection

The check-all-that-apply (CATA) method is used to identify a subset of important attributes from a predetermined list. The challenge at hand was the abundance of available attribute lists, which required the selection of the most appropriate one. To determine the most suitable list for this experiment, a pre-experiment was required. Three lists of elicited attributes from the literature were considered.

- The Audio Wheel for reproduced sound [11], compiled from various elicitation studies and chosen for its wide coverage of aspects related to reproduced sound.
- The spatial audio quality inventory (SAQI) [12], developed to assess audio quality in virtual environments, making it suitable for spatial sound evaluation.
- The attribute list from Francombe et al. [22], compiled through a paired comparison approach similar to our experiment, included expert and nonexpert listener assessments of various reproduction systems. Although initially separated, the sublists for experts and nonexperts were combined due to overlapping attributes for convenience.

Upon comparison, the different attribute lists demonstrated significant overlap, with several attributes having similar or identical descriptions across the lists. However, there were also noticeable differences, particularly concerning potentially important attributes.

The six stimuli were presented to three expert listeners with Tonmeister training.[1] All attributes from the three previously mentioned publications were included into one table. To facilitate the overview, all attributes were sorted into high-level categories similar to the categories applied by the audio wheel for reproduced sound [11]: *dynamics, timbre, spatial, transparency, artifacts*, and *general*. Attributes from different lists describing the same perception were listed next to each other for direct comparison. The list covered all information of the attributes, including name, definition, range, and further explanation and examples.

The tasks were to a) mark each attribute that best describes the perceptual differences between the presented stimuli and select the description of the attribute that fits the perception most, b) indicate if there are any attributes in the same row (i.e., describing the same attribute) that seem to be put in the wrong category, and c) indicate if any important attributes are missing from the given attribute list.

From the 140 possible attributes from all lists, 40 unique attributes were selected. Only one attribute (*position of sound*) was selected by all three participants. For five attributes (*loudness, envelopment, depth of field, spatial clarity, degree of liking*), different definitions of the same attribute were preferred.

The results revealed that none of the initial three lists exhibited clear dominance for further use in terms of most selections. Therefore, it was decided to merge all selected unique attributes in one list. Where multiple attribute definitions for one unique attribute were selected, either the attribute definition with higher selection count was chosen, or, if the selection was equally distributed, the definition was taken from the list with the most selection agreements (Francombe et al. [13]). Furthermore, the hedonic attributes *overall subjective preference* and *ease of listening* were excluded, resulting in a final list of 38 descriptive attributes that were used in the main experiment (see Table 2).

1.5 Experiment Procedure

The main experiment consisted of three phases: familiarization, the rating and attribute selection stage, and a brief follow-up questionnaire. The experiment obtained approval from the Ethical Committee for the Central Jutland region in Denmark.

1.5.1 Familiarization

Two types of familiarization were conducted prior to the main experiment. First, participants needed to familiarize themselves with the full list of attributes. This was important to ensure a basic knowledge of the available attributes and their definitions. Then, before starting the main experiment block, the participants were asked to familiarize themselves with all the stimuli. The participants were able to switch between the six experiment stimuli on a single page interface. The stimuli were grouped by spatial variation and randomized by reproduction system. The participants were only able to continue with the test after every stimulus was clicked at least once.

The familiarization was performed to ensure that participants were aware of the range of differences in BAQ within the stimuli, so they were capable of using the full range of the scale in the main experiment appropriately, as instructed. BAQ was defined as "*an attribute that is defined as including all aspects of the sound quality being assessed. It includes, but is not restricted to, such things as timbre, transparency, stereophonic imaging, spatial presentation,*

[1] "Tonmeister training" involves specialized education emphasizing critical listening skills alongside technical expertise in audio engineering, production, and mastering.

Table 2. List of attributes used in the study and their descriptions. The list was compiled as described in SEC. 1.4.

Dynamics	
Loudness	Perceived loudness of a sound source. Disappearance of a sound source can be stated by a loudness equalizing zero. Example of a loudness contrast: whispering vs. screaming. Scale: quieter–louder.
Ensemble balance	Relative levels of the different sources. Scale: balanced–unbalanced.
Prominence	The prominence of a sound relative to other sounds in the track. Scale: no prominent sound–a prominent sound.
Dynamic range	Amount of loudness differences between loud and soft passages. In signals with a smaller dynamic range, loud and soft passages differ less from the average loudness. Signals with a larger dynamic range contain both very loud and very soft passages. Scale: smaller–larger.
Timbre	
Overall spectral balance	The magnitude of broad cuts and boosts in the spectrum. Scale: none (perceived as flat)–significant.
Timbral balance: dark–bright	Denotes the balance between bass and treble. Scale: Dark (Excessive bass. Either loud bass or weak treble), Neutral (Bass and treble are perceived equally loud, there is a balance in the reproduction. This also applies if both bass and treble are equally weak or if the bass and treble are both too loud. If it leads to prominent or soft midrange, this is assessed by the midrange strength.), and Bright (Excessive treble. Either loud treble or weak bass). The cause for the sound being dark or light can be deduced from the assessments of Bass strength and Treble strength.
Timbral balance: full	If both low and high frequencies are well represented with good extension, the sound is full. Scale: a little–a lot.
Treble: tinny	Resonances or narrowband frequency prominence in the treble or high frequencies. Scale: none–weak–a lot.
Treble strength	The relative strength of the treble or high frequencies. Scale: Weak (Covered, unsharp), A little under neutral (A soft sound without being dull), Neutral (In the middle of the scale, where you can clearly distinguish instruments), and A lot (Treble Raised. Sharp, hard sound).
Midrange strength	The relative level of the midrange, i.e., the middle frequencies, e.g., sopranos, trumpets, violins, and xylophones. Not to be confused with Canny, which represents prominent narrow frequency (resonances) in the midrange. Scale: soft–loud.
Bass strength	The relative level of bass, i.e., the low frequencies, for example, male voices, bass guitar, bass drum, timpani, and tuba. Should not be confused with bass depth that indicates the low-frequency bass extension. Scale: soft–loud.
Transparency	
Clarity	How clear a sound is to the listener. Scale: muffled–clear.
Spectral clarity	The ability to distinguish different sources based on their spectral content (timbre). Scale: indistinct–precise.
Discernibility	Whether you can pick out individual sounds. Scale: I can pick out the individual sounds–I can't pick out the individual sounds.
Presence	Perception of "being-in-the-scene" or "spatial presence." Impression of being inside a presented scene or to be spatially integrated into the scene. Scale: lower–higher.
Detail	The amount of details within the individual sounds that you can discern. Scale: little detail–lots of detail.
Naturalness	Impression that a signal is in accordance with the expectation/former experience of an equivalent signal. Scale: lower–higher.
Realism	Overall, how realistic it sounds. Scale: unnatural–natural, unreal–real.
Artifacts	
Spectral resonance	Presence of unpleasant resonances/presence of sharp peaks in the spectrum. Scale: resonant–not resonant.
Perceptibility of noise	How perceptible the noise is. Scale: Perceptible–imperceptible.
Spatial	
Spatial clarity	Ease of localization of individual sources. Scale: distinct–indistinct.
Spatial movement	Degree of movement of sound sources. Scale: static–dynamic.
Spatial naturalness	How natural the source position is within the 3D image. Scale: natural–unnatural.
Surrounding	The spread of sound around the space. Scale: unidirectional–surrounding.
Sense of space	The extent to which you feel you are in the same space in which the music/event was performed. Scale: not at all–very much.

Table 2. *Continued*

Envelopment	Degree of being surrounded by a source, scene, or ensemble. Typically, envelopment is associated with a scene. Scale: not enveloping–completely enveloping. Being surrounded by reverberation would be considered highly enveloping. Being surrounded by a large number of dry sources may also be highly enveloping. This may be heard when standing and listening to the rain hitting the pavement. Envelopment may occur with reverberation or other aspects of the scene such as applause in a concert hall, atmosphere or air conditioning (room tone). Holes (an absence of sound from a certain direction) in the reproduction would normally reduce envelopment. Envelopment may be subdivided in horizontal and vertical envelopment.
Spatial openness	How claustrophobic the sound feels. The proximity of the 3D sound field. A sense of air/openness. Scale: suffocating–open.
Spatial balance	How biased the reproduction is to a particular area(s) of the sound field (including a hole in the middle). Scale: even–biased.
Depth of field	Perceived proximity of sources. Scale: close–distant.
Position of sound	Your preference depending on the position of the sound(s). Scale: like–dislike.
Externalization	Describes the distinctness with which a sound source is perceived within or outside the head regardless of their distance. Terminologically often enclosed between the phenomena of in-head localization and out-of-head localization. Examples: Poorly/not externalized = perceived position of sound sources at diotic sound presentation via headphones, good/strongly externalized = perceived position of a natural source in a reverberant environment and when allowing for movements of the listener. Scale: more internalized–more externalized.
Realism of reverb	How realistic the reverb sounds. Scale: natural–unnatural.
Spatial disintegration	Sound sources, which, by experience, should have a united spatial shape, appear spatially separated. Possible cause: Parts of the sound source have been synthesized/simulated using separated algorithms/simulation methods and between those exists an unwanted offset in spatial parameters. Examples: fingering noise and playing tones of an instrument appear at different positions; spirant and voiced phonemes of speech are synthesized separately and then reproduced with an unwanted spatial separation. Scale: more coherent–more disjointed.
Depth	Perceived extent of a sound source in radial direction. Scale: less deep–deeper.
Height	Perceived extent of a sound source in vertical direction. Scale: less high–higher.
Horizontal width	360-degree horizontal width. Scale: point source–surrounding.
	General
Speech intelligibility	Impression of how well the words of a speaker can be understood. Typical of low speech intelligibility: train station announcements. Typical for high speech intelligibility: Newscaster. Scale: lower–higher.
Ease of listening	The effort required to listen to the sound. Scale: easy–hard.

reverberance, echoes, harmonic distortions, quantization noise, pops, clicks and background noise." [33].

1.5.2 Rating and Attribute Selection Phase

The experiment design included six repeated paired comparison judgements; therefore, each experiment session consisted of twelve pages. Each page included the judgement for one stimulus pair. The order of the twelve pages was randomized. Fig. 2 shows the interface that was used for the main experiment, including the scale and all possible attribute selections.

At the start of the trial, participants were automatically presented with stimulus A followed by stimulus B. The assignment of the stimuli to button A or B in each particular pair was randomized. While the stimuli were playing the first time, the participants were not able to switch between them. This was done to ensure the participants listened to both stimuli completely before making a judgement. After the first playback of the stimuli, the participant was able to switch freely between the stimuli. Each page had two sections. In the first section, the participant was asked to rate which stimulus was perceived to have higher BAQ and by how much using the slider on the scale ranging from −50 to 50. A negative value indicated that A was perceived as having better quality than B, whereas a positive value of 50 indicated that B was perceived as having better quality than A. The values −50 and 50 represented the highest possible ratings, respectively. The midpoint of the scale was denoted by a black bar at the 0 point, indicating that A and B were perceived as having equal quality. The task instructions (based on Francombe et al. [13]), read as follows:

"*Bear in mind the full range of stimuli that you heard in the familiarization and try to position the slider appropriately:*

- *When the slider is on the left, A has higher BAQ than B; the further left you move the slider, the greater the difference.*
- *When the slider is on the right, B has higher BAQ than A; the further right you move the slider, the greater the difference.*
- *When the slider is in the center, the position indicates no difference in BAQ between A (left) and B (right)."*

The second section of the page comprised the list of attributes. Only the attribute labels were presented in the interface, while the definitions were accessible on a physical sheet. The position of the attributes on the interface was

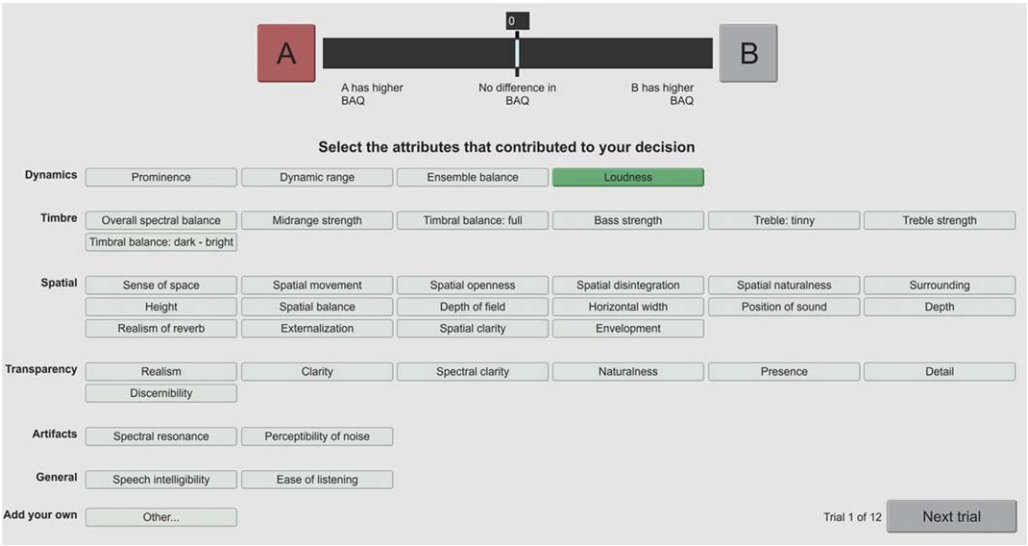


Fig. 2. The experiment GUI featuring the BAQ rating scale and attribute selection. Stimulus A is selected and playing, with both stimuli rated equally (0 on the slider). Below, attribute buttons sorted by categories are shown, with "loudness" currently selected.

randomized by higher level categories (e.g., *dynamics, timbre, spatial*, etc.), and within categories for each participant. They were not randomized for each page to facilitate quick access for the participant and to avoid confusion resulting from an ever-changing list of attributes.

The wording for the task in this section of the page was: *"indicate which of the attributes contributed to your decision for positioning the slider. Note, that it does not matter if the attributes contributed in a negative or positive way, as long as they influenced your decision."* The participants were informed that at least one attribute had to be selected, and the slider had to be moved to proceed to the next page. If the participants felt that one of the percepts leading to their rating was missing from the available list, they had the option to enter a new attribute alongside their definition. This option was included to ensure coverage for potentially overlooked attributes that could be significant.

1.6 Participants

Ten experienced listeners participated in the experiment. An experienced listener was defined as a person who is either trained in critical listening or frequently uses critical listening in their job and has taken part in listening experiments rating sound quality before. All the participants were employees at Bang & Olufsen working within the field of audio and acoustics and are regularly checked for no hearing impairments. The initial group consisted of ten participants within an age range of 25–61 years (mean age 34). Due to inconsistencies, certain participants needed to be removed from the data evaluation. After screening, the group used for analysis in SEC. 2.2 consisted of of nine participants (mean age 35). For the analysis in SECS. 2.3 and 2.4, the group consisted of seven participants (mean age 36). The participants were not paid for their participation.

2 RESULTS

The experiment included the results of ten participants, with each session lasting between 45 and 60 min. To analyze the results, participant consistency was first assessed for each of the tasks individually (SEC. 2.1). This analysis was followed by a separate investigation of the influence of the independent variables (spatial variation and reproduction system) on the dependent variables (BAQ ratings [SEC. 2.2] and attribute selection frequency [SEC. 2.3]). Additionally, the evaluation of the relationship between these variables is presented, before the findings are summarized.

2.1 Participant Performance

The performance of each participant was evaluated for each of the two tasks: quality rating and attribute selection. This assessment involved evaluating the consistency of ratings provided by individual participants. Based on this evaluation, the need to exclude participants to reduce data noise was determined.

2.1.1 Quality Rating Accuracy and Consistency

When conducting a paired comparison task, one method of evaluating the performance of participants is to calculate the circular error percentage. A circular error occurs when a participant or a panel of participants gives intransitive ratings (i.e., if stimulus A is perceived as better quality than B and B as better quality than C, then C can not logically be perceived better than A [34]). A high circular error percentage can indicate various factors such as insufficient attention to the test, a shift in assessment criteria during the test, or general confusion due to the test's difficulty or subtle differences between presented stimuli or attributes. The circular error percentage is calculated as the percentage of intransitive ratings of all possible triads of paired comparisons [22]. Every participant rated four triads of paired comparisons, giving a total of 40 triads in the whole

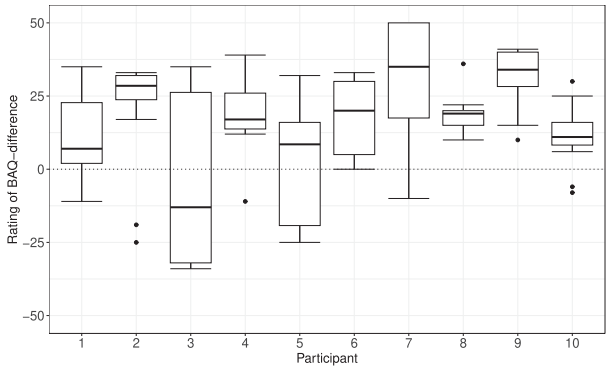


Fig. 3. Box plots of BAQ difference ratings for all conditions and repeats by participant. Each box plot comprises twelve ratings (two repeats of six judgments) ranging from −50 (A was better) to 50 (B was better). After preprocessing, A consistently denotes the reproduction system with a lower channel count, while B represents reproduction systems with a higher channel count.

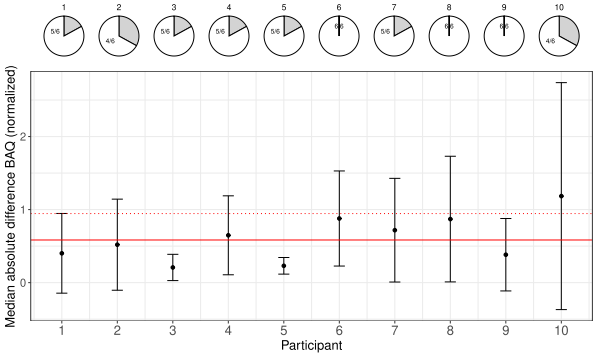


Fig. 4. Median absolute difference in BAQ ratings, z-score normalized, with error bars indicating the median absolute deviation (MAD) for each participant. The solid line represents the overall median, and the dotted line the median MAD around the median difference across all participants. Above the box plots, pie charts illustrate judgment consistency for repeated stimuli (e.g., A perceived as having better BAQ for both first and repeated judgments). Gray areas within the pies represent trials where judgments changed during the repeat. The numbers within the pie plots show the number of consistent judgement from the total number of judgements.

dataset. Only one circular error was detected (from 120 possible trials, i.e., 2.5 %), suggesting that participants were likely attentive and capable of adequately and consistently distinguishing the stimuli.

To evaluate differences between first and repeated ratings for each participant, an analysis of uniform scale use was necessary, thus allowing for meaningful interparticipant comparison of consistency. Fig. 3 shows all BAQ difference ratings for all conditions and repeats by participant as box plots. Each box plot consists of 12 ratings per participant (two repeats of six judgements) with values ranging between −50 (A was better) and 50 (B was better). After preprocessing the data to correct for the random assignment of the stimuli to buttons A and B, stimulus A always refers to the reproduction system with a lower channel count reproduction system. Therefore, the shift to the upper half of the scale, which is observed for most of the box plots, shows that reproduction systems with higher channel count are generally perceived as having better audio quality.

The variation of box plot height among the participants indicates a high variance in scale use between participants. Therefore, to be able to compare the accuracy of scale use between participants, the data were then normalized by using z-score transformation before individual rating absolute differences between first and repeated judgements were calculated.[2] The median absolute difference between ratings of the first and repeated judgement are shown in Fig. 4. The error bars indicate the median absolute deviations (MAD) [35] for each participant. A lower median indicates a generally greater consistency in the participant's responses, while small MAD values indicate that judgment errors were made with a consistent low level of variation.

The results reveal significant variability in MADs among participants, indicating that some participants encountered greater difficulties in consistently rating different stimuli compared to others. The highest MAD was found for participant 10. To assess and evaluate each participant's performance in relation to the overall participant performance, the median and the upper MAD for all participants were calculated and are shown in Fig. 4 by a solid and dashed line, respectively. Here, participant 10 exhibited a significantly higher median, surpassing the upper MAD threshold, indicating notably lower accuracy for this participant compared with the other participants.

The pie plots above the box plots indicate the number of trials where the judgement was consistent for the repeat (e.g., A is perceived as having better BAQ for first and repeated judgements). The gray area in the pies shows the proportion of trials where a change in judgment to the alternative stimulus occurred during the repeat. Most of the participants (1, 3, 4, 5, and 7) changed their opinion at least once. Two participants (2 and 10) altered their judgment twice, while only three participants (6, 8, and 9) remained consistent in all of their decisions.

Overall, these observations suggest that the evaluation of audio quality difference was a challenging task for some participants, as results show difficulties in making consistent judgments. Because participant 10 was identified as an outlier based on their median, while also being part of the group with the highest amount of changes in decision, the participant was excluded from the subsequent analysis.

2.1.2 Attribute Selection Consistency

To evaluate participant performance in the CATA task, attribute selection consistency between the first and repeated judgments was checked. Judgements were considered to be consistent when the same attribute was selected or not selected in both the first and the repeated judgments. Judgments were considered to be inconsistent if attributes were either only selected in the first judgement or only in the

[2]Absolute difference (d) is calculated as $d = \left|\frac{rating_1-\mu}{\sigma} - \frac{rating_2-\mu}{\sigma}\right|$ where μ represents the mean and σ denotes the standard deviation.

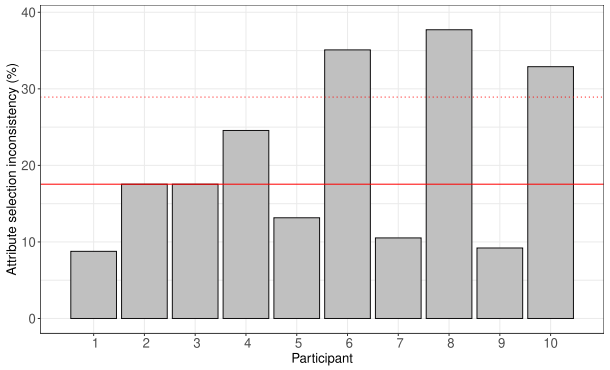


Fig. 5. Percentage of selection inconsistency for each participant. The horizontal line represents the median of attribute selection inconsistency over all participants, and the dotted line indicates the median absolute deviation (MAD) above the median.

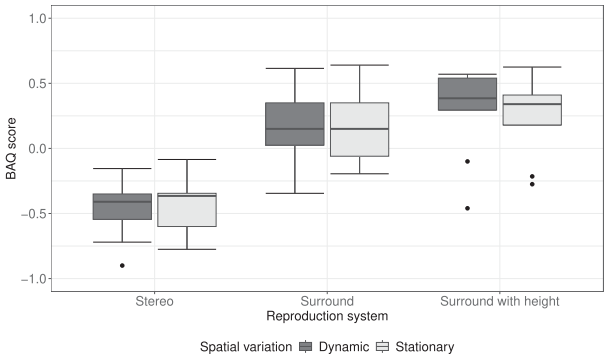


Fig. 6. Box plots showing the calculated BAQ scores (see SEC. 2.2 for a detailed description of the scoring methodology) for three different reproduction systems grouped by spatial variation (dynamic, stationary) across all participants (excluding participant 10).

repeated judgement. Each page allowed a selection of up to 38 attributes, summing up to a total of 228 possible attributes selections. The attribute selection inconsistency percentage was calculated as the percentage of inconsistent selections from a total of 228 possible attribute selections per participant.

Fig. 5 shows the attribute selection inconsistency percentage for each participant and the median and MAD of the percentage across participants. Participants 6, 8, and 10 had noticeably higher inconsistency percentages than the other participants and were identified as outliers with significantly worse performance, as their values lay above the MAD. Based on the presented results, judgements by participants 6, 8, and 10 were determined as unreliable and excluded from the final analysis of the attribute selection.

2.1.3 Summary

The purpose of this section was to evaluate the validity and reliability of the performance of the participants and identify participants that potentially need to be excluded. The consistency of the participants' judgments was evaluated by analyzing the BAQ rating consistency and attribute selection consistency between repeated judgements of the experiment for each participant. Following this analysis, participant 10 was excluded for further evaluation of the BAQ-difference ratings (SEC. 2.2). Participants 6, 8, and 10 were excluded from further analysis of the attribute selection frequency (SEC. 2.3).

2.2 Basic Audio Quality Ratings

BAQ scores for each of the reproduction systems by participant and spatial variation were summarized, applying the method described by Walton et al. [21] converting paired comparison judgements to scale values for each individual system.[3] As explained in SEC. 2.1.1, one participant was identified as unreliable and subsequently excluded from the analysis, resulting in a dataset comprising the responses of 9 participants. The results are shown in Fig. 6. Box plots were plotted for calculated BAQ scores over all participants, grouped by spatial variation. The box plots indicate a relationship between an increase in the number of channels in the reproduction system and a high BAQ score. This result further shows that higher channel reproduction systems were generally perceived as having better audio quality. Although the box plots in between different reproduction systems show differences in ratings, the difference between ratings for spatial variation is very low.

A linear mixed-effects analysis of variance (ANOVA) model was used with the BAQ scores to investigate the observed effects and check for interaction of the independent variables on the BAQ scores. The two fixed independent factors were reproduction system and spatial variation, while the participants and their interaction with spatial variation and reproduction systems were included in the model as random factors. It is worth noting that the model specifically accounted for first- and second-order interactions.

The model complexity was reduced by iteratively eliminating nonsignificant predictors while considering the impact on model fit using a stepwise procedure. The resulting model was composed of the two fixed effects, their interaction, and one random effect (interaction of participants and reproduction system). The model residuals were shown to be approximately normally distributed using a Shapiro–Wilk test ($W = 0.962$, $p = 0.414$) and homogeneous using a Levene's test ($F = 0.176$, $p = 0.970$). The ANOVA results of the fixed effects are shown in Table 3. The results revealed that BAQ ratings were affected by the choice of reproduction system ($F = 17.853$, $p < 0.001$). However, there was no significant impact from spatial variation ($F = 0.000$, $p = 1.000$) or interaction between spatial variation and re-

[3] Values were scaled to lie between ± 1, where -1 corresponds to a full preference for A, and +1 for B. With P_{ij} being the probability of having better BAQ for stimulus i versus stimulus j, it is assumed that $P_{ij} = -P_{ji}$. A negative probability P_{ij} denotes that stimulus j was considered to have better BAQ than stimulus i. BAQ scores are then calculated as $S_i = \sum_{j \neq i} P_{ij}$, where S_i is the BAQ score for stimulus i [21].

Table 3. Analysis of variance (ANOVA) table of the fixed effects of the linear mixed-effects model. Sum sq. = sum of squares, mean sq. = mean square, DF = degrees of freedom. p values below $p = 0.001$ are highlighted in bold.

	Sum sq.	Mean sq.	DF	F	p value
Spatial variation	0.000	0.000	1	0.000	1.000
Reproduction system	0.521	0.261	2	17.853	**0.000**
Spatial variation × reproduction system	0.010	0.005	2	0.345	0.709

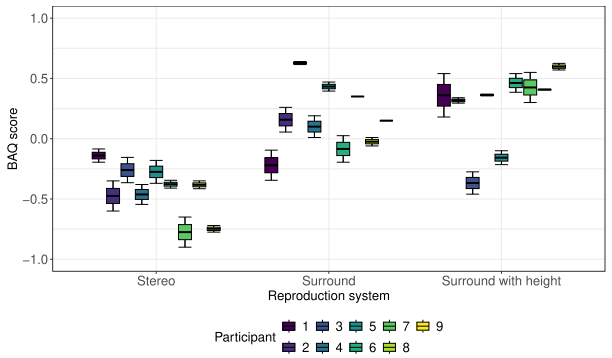


Fig. 7. Box plots illustrate the calculated BAQ scores for three distinct reproduction systems over both spatial variations, grouped by participant (see SEC. 2.2 for a detailed description of the scoring methodology).

production system ($F = 0.345$, $p = 0.709$). Consequently, it was observed that higher BAQ ratings were generally associated with a greater number of channels in the reproduction system.

The reduced model fit was evaluated by marginal R^2 and conditional R^2; 55.08% of the variance in the BAQ ratings was explained by the fixed factors, while the full model explained 91.74% of the variance. This analysis reveals a notable proportion of variance explained by just one random effect (interaction of participants and reproduction system). Consequently, a more in-depth examination of the participant interactions was undertaken. The outcomes are visualized in Fig. 7. Notably, participant ratings appear to be relatively grouped together for "stereo" and "surround." However, for the "surround with height" condition, the system received noticeably lower ratings from participants 3 and 6 compared with the rest of the participants. This division of ratings into two clusters may account for the influence of the participant–reproduction system interaction. Possible explanations for this divergence could include variations in participant preferences, potential confusion stemming from elevated sound sources, modifications in the audio mix, or changes in the perceived direction of sound sources.

In summary, the results indicate that BAQ ratings were primarily influenced by the evaluated audio reproduction systems, demonstrating a trend in which higher channel systems exhibited higher quality. However, it is worth noting that there was also a notable impact of varying participant ratings for the "surround with height" system, highlighting the significance of the participant–reproduction system interaction.

2.3 Attribute Selection Frequency Analysis

The analysis of attribute selection frequency was guided by two objectives: a) to identify important attributes that contribute to the perceived differences in BAQ and b) to investigate how the selection of these attributes are affected by the two independent variables (spatial variation and reproduction system comparison[4]). The analysis in this section was carried out on the reduced dataset with seven participants (see SEC. 2.1.2).

In this study, assessors were directly asked to select attributes contributing to their perception of differences in BAQ between the presented stimuli. Therefore, it can be assumed that any attribute, once selected, effectively describes either the whole or a part of the perceived differences between the presented stimuli [36, pp. 228–231]. Assuming the assessor panel understood the range of possible attributes, one way to identify the most important attributes for this decision process would be to observe the attributes that were selected with the highest frequency across all participants. Although there are alternative methods, such as examining the count of participants selecting a specific attribute, the primary focus in this analysis, when referring to key attributes, will be on the highest frequency approach.

Therefore, the analysis of attribute frequency was structured as follows:

- First, attributes that were rarely used were excluded, using an analysis of the whole dataset. This was done to enhance robustness of subsequent analyses (see SEC. 2.3.1).
- Second, the effect of interaction of spatial variation and reproduction system comparison on attribute selection frequency (see SEC. 2.3.2) was analyzed.
- Last, the influence of both independent factors (spatial variation and reproduction system comparison) on attribute selection was assessed (see SEC. 2.3.3).

2.3.1 Overall Selection Frequency Analysis

The initial step of the process involved eliminating attributes that were rarely or never selected. This step was taken to reduce "noise" in the data and improve the precision of the subsequent analysis. The overall distribution of

[4]It is important to note that "reproduction system comparison" does not represent a new factor. In the BAQ ratings analysis, "reproduction system" was a factor derived from system-wise score calculations. However, for attribute selection frequency analysis, separation by system is not possible because each comparison involves inseparable attribute lists.

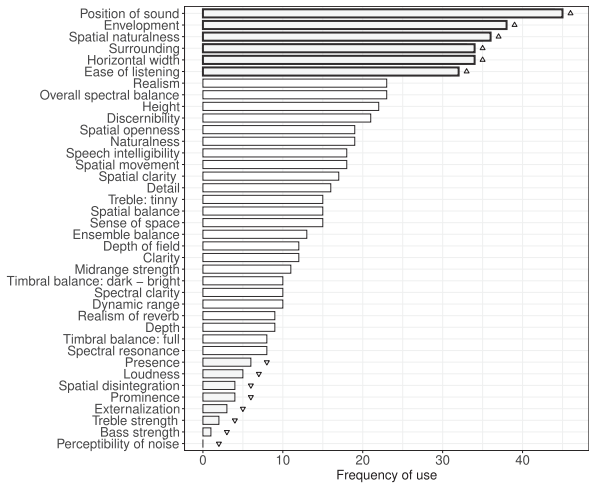


Fig. 8. Selection frequency of all 38 attributes and conditions for 7 participants, sorted from highest to lowest observed frequency. Downward-facing triangles show attributes selected fewer times, and upward-facing triangles show attributes selected greater times than expected, assuming a uniform distribution.

attribute selections was analyzed, and attributes that were rarely selected were excluded. To analyze the selection frequency, goodness-of-fit-analysis was employed [24, pp. 186–187].

Overall, the results showed a nonuniform distribution, indicating a relationship between the audio stimuli and attribute selection ($\chi^2(37) = 278.11$, $p < 0.001$). Standardized residuals were then used to evaluate which attributes had been selected more or less than the expected frequency for a uniform distribution. Significant differences at the 95 % confidence level are indicated by standardized residuals lying outside $\pm$ 1.96 [37, p. 698].

The results are shown in Fig. 8. Here, the selection frequency of each attribute summed over participants and conditions is shown. Attributes that were selected significantly over and below the expected frequency of a uniform distribution are marked with upward-facing triangles (greater than expected frequency) and downward-facing triangles (below expected frequency), respectively. The selection frequency of six attributes was determined as greater than the expected frequency (*position of sound, envelopment, spatial naturalness, surrounding, horizontal width*, and *ease of listening*), while eight attributes were selected below the expected frequency (*presence, loudness, spatial disintegration, prominence, externalization, treble strength, bass strength*, and *perceptibility of noise*).

As a form of sanity check, the number of participants that selected a specific attribute during the experiment was then determined before comparing this ranking to the attributes that were chosen most frequently. This comparison helped to determine if the attributes that were commonly chosen were also favored by the majority of participants. This is crucial because high selection rates for specific attributes could be influenced by a few participants consistently choosing those attributes across all conditions.

Table 4 presents all attributes alongside the count of participants who selected the specific attributes at some point during the experiment. The attributes are ranked in descending order from the highest number of participants selecting a specific attribute to the lowest. The table shows that the majority of attributes selected above the expected frequency of a uniform distribution were also selected by the majority of participants. This observation indicates that multiple participants reached a consensus on the importance of these attributes and the selections were not only influenced by the opinion of certain individuals. To maximize resolution of the statistical tests, the eight attributes that fell below the expected frequency were excluded for further analysis.

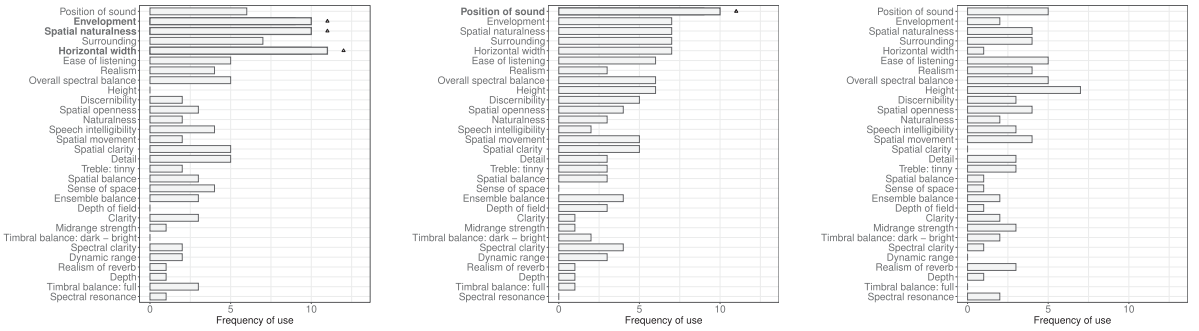
Table 4. Number of participants selecting attributes. Attributes that were selected more often than expected than for a uniform distribution are in bold letters, and attributes that were selected fewer than expected are in italics (see SEC. 2.3.1).

Number of participants	Attributes
7	**Surrounding, spatial naturalness, position of sound**
6	Spatial movement, realism
5	Spectral clarity, spatial clarity, sense of space, **horizontal width, envelopment, ease of listening**, depth of field
4	Timbral balance: dark–bright, spatial openness, realism of reverb, *presence*, overall spectral balance, naturalness, height, discernibility, detail, clarity
3	Treble: tinny, timbral balance: full, speech intelligibility, *spatial disintegration*, spatial balance, *loudness*, *externalization*, ensemble balance, depth,
2	*Treble strength*, spectral resonance, *prominence*, midrange strength, dynamic range
1	*Bass strength*
0	*Perceptibility of noise*

2.3.2 Influence of Interaction Between Spatial Variation and Reproduction System Comparison on Attribute Selection

To evaluate the interaction of the independent variables—spatial variation and reproduction system comparison—the dataset was divided into six subgroups based on the combinations of levels from the two independent variables. For each subgroup, a goodness-of-fit test was conducted to achieve two objectives: 1) identify a relationship between the audio stimuli and attribute selection (indicating a distribution different from a uniform one), and 2) identify attributes that were significantly selected more than expected (standardized residual analysis).

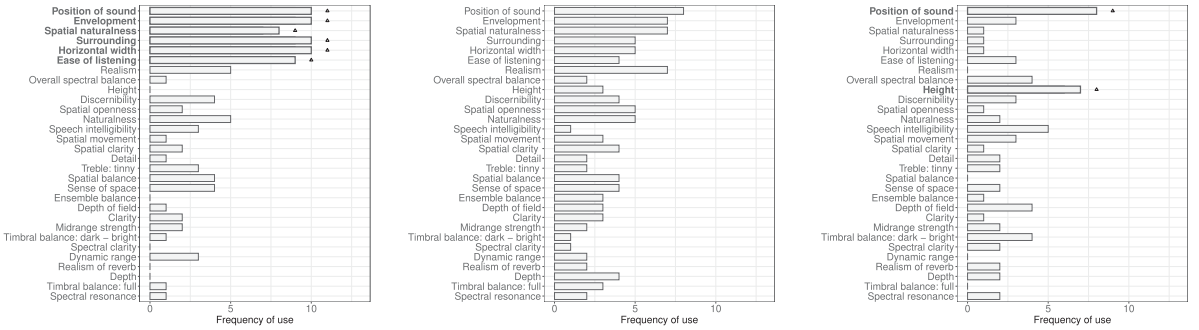
The results are shown in Fig. 9, where the χ^2 statistics with the group sample size (N) and the attributes selected more frequently than expected within each subgroup are marked in bold. It is worth noting that the highest selection count for an attribute at this level was 12. This means that



(a) Dynamic:
Stereo vs. surround
$\chi^2(29, N = 104) = 59.846$, $\mathbf{p = 0.000}$

(b) Dynamic:
Stereo vs. surround with height
$\chi^2(29, N = 112) = 44.429$, $\mathbf{p = 0.033}$

(c) Dynamic:
Surround vs. surround with height
$\chi^2(29, N = 78) = 32.769$, $p = 0.287$



(d) Stationary:
Stereo vs. surround
$\chi^2(29, N = 97) = 83.928$, $\mathbf{p = 0.000}$

(e) Stationary:
Stereo vs. surround with height
$\chi^2(29, N = 108) = 28.667$, $p = 0.482$

(f) Stationary:
Surround vs. surround with height
$\chi^2(29, N = 67) = 39.119$, $\mathbf{p = 0.099}$

Fig. 9. Results of the goodness-of-fit analysis (degrees of freedom and sample size (*N*) in parentheses) and standardized residual tests for each combination of factor levels. Attributes that were selected more frequently than expected assuming a uniform distribution for each subgroup are bold and marked with an upwards triangle. Test statistics for each subgroup are reported. *p* values below $p = 0.1$ are highlighted in bold.

the margin between attributes selected significantly more often than expected and those within the expected range was very narrow.

Attributes that showed significant frequency of use above expected ($p < 0.1$), were observed within four distinct subgroups: dynamic–stereo vs. surround ($\chi^2(29, N = 104) = 59.846$, $p < 0.001$), stationary–stereo vs. surround ($\chi^2(29, N = 112) = 44.429$, $p = 0.033$), dynamic–stereo vs. surround with height ($\chi^2(29, N = 112) = 44.429$, $p = 0.033$) and stationary–stereo vs. surround with height ($\chi^2(29, N = 67) = 39.119$, $p = 0.099$). Most attributes were identified within the stationary-stereo vs. surround group (*horizontal width*, *spatial naturalness*, *envelopment*, *surrounding*, *position of sound*, and *ease of listening*). In contrast, fewer attributes were selected for the same reproduction system comparison under dynamic conditions, with only three attributes (*horizontal width*, *spatial naturalness*, and *envelopment*) selected over the expected frequency.

The results indicate that, under stationary conditions, listeners exhibited a more pronounced consensus across a wider spectrum of attributes. This was evident by a greater number of attributes displaying a selection frequency that significantly deviated from a uniform distribution. In contrast, the dynamic condition saw a reduced degree of consensus, with fewer attributes exhibiting significant deviations from a uniform distribution. This difference suggests that in the dynamic condition, listeners may have found it more challenging to assess and differentiate between the reproduction systems. A similar observation can be made when comparing "surround vs. surround with height." In this case, only attribute selections in the stationary conditions displayed evidence of diverging significantly from a uniform distribution, whereas there was no such tendency in the dynamic condition.

Contrasting findings to the observations explained here can be observed when examining the "stereo vs. surround with height" comparison. In this case, only the attribute selections in the dynamic condition showed a statistically significant deviation from a uniform distribution, while only one attribute was selected above the expected frequency: *position of sound*. This finding suggests that the only aspect participants consistently agreed upon was the dynamic nature of sound positioning across the different reproduction systems, particularly when the sound position was in motion, making it more perceivable.

To gain a more comprehensive understanding, Cochran–Mantel–Haenszel (CMH) analysis [38, p. 576] was applied

as an omnibus test. The CMH analysis is a statistical technique that assesses the overall impact of interaction on attribute selection, specifically identifying changes in the relationship between two variables due to the influence of a third grouping variable.

To determine whether there is a significant influence of factor interaction on attribute selection, CMH analysis was conducted twice, using each factor separately as the grouping variable. The findings suggest that spatial variation influenced how participants chose attributes under different conditions of reproduction system comparison ($\chi^2(58) = 90.834$, $p = 0.003$). Meanwhile, the relationship of attribute selection for different spatial variations remained consistent for different conditions of reproduction system comparison ($\chi^2(29) = 19.363$, $p = 0.911$). These results show that there is an interaction effect on attribute selection, but only when looking at the link between attribute selection for different reproduction system comparisons for different spatial variations.

The CMH analysis, on its own, does not provide insights into the specific changes in attribute selection resulting from the interaction. To achieve a more detailed understanding, further investigation involving significant attribute selection frequency differences was considered. However, because of the scarcity of data points on that level of analysis, these findings were inconclusive.

In summary, the outcomes of the selection frequency analysis within each subgroup revealed a significant influence of stimuli for four conditions. A significant influence of spatial variation on attribute selections under different reproduction system comparisons was also found. In general, dynamic conditions tend to reduce the ability of the panel to clearly determine the attributes that are affecting their decision, unless the most dominant change among the reproduction systems is the shifting of sound position.

2.3.3 Influence of Main Effect of Reproduction System Comparison and Spatial Variation on Attribute Selection

After looking at the influence of factor interaction on the attribute selection, the main effects of spatial variation and reproduction system comparison were analyzed individually. First, the analysis focused on the distribution of attribute selections for each level of the independent factors, summarizing the attributes that were chosen more frequently than expected. To then identify significant differences of attribute selection frequencies for each main effect across their levels, two methods were applied as omnibus tests: Pearson's χ^2 test of independence and a randomization test based on the sum of Q statistics from Cochran's Q test [39].

The omnibus test was applied as described by Meyners et al. [39] and Pike [40, pp. 344–355]. This test combines a randomization test with Cochran's Q test, which is used to assess if the assessor panel revealed significant differences between systems using each of the CATA attributes [39]. The frequencies of attribute selection were randomized across systems while keeping assessor, item, and attribute the same. Cochran's Q test statistic was calculated for each attribute and compared with that for the true data. The p value was calculated as the proportion of the number of randomization where Q was greater than for the true data. Repeats were incorporated by using all judgments for Pearson's χ^2 statistics, emphasizing consistently selected attributes. Therefore, both the first and repeated judgments were also used to calculate the Cochran's Q statistics. Only attributes that were consistently selected in both the first judgment and the repeat, were considered selected.

The suggested effective sample size >24 was not achieved for any of the tested attributes. The effective sample size is derived by the multiplication of assessors showing some rating differences between systems (i.e., neither checking all nor none of the systems) with the number of systems under test [40, p. 354]. Therefore, it should be noted that the post hoc test may not be as robust because of the limited sample size.

Both methods test for significant differences between groups of categorical data and were chosen because no valid test specifically designed for analyzing global attribute selection frequencies currently exists. By using both tests, the study aimed to compare the results and account for any limitations or weaknesses associated with each test, thereby strengthening the overall analysis of the data.

2.3.3.1 Reproduction System Comparison. Results are summarized in Table 5 and Table 6. Analyzing the distribution of attribute selections based on the compared reproduction systems, as shown in Table 5, reveals a pattern consistent to the observations in SEC. 2.3.2.

Every factor-level subgroup ("stereo vs. surround," "stereo vs. surround with height," and "surround vs. surround with height") showed an influence of stimulus on the distribution of attribute selection frequency. The highest number of significantly above the expected frequency selected attributes for each condition was found in the "stereo vs. surround" condition, followed by a similar but reduced number of attributes in the "stereo vs. surround with height" category. Notably, *position of sound* was selected higher than expected in all three groups, indicating its consistent importance across different reproduction system comparisons. However, *height* was exclusively significantly selected in the "surround vs. surround with height" category. These results show that attribute selection frequencies seem to be dependent on the reproduction systems that were compared with each other.

Subsequently, the analysis focused on differences in attribute selection frequencies for each main effect across their respective levels. In the case of Pearson's χ^2 analysis, 45.55% of cells showed frequencies below the expected threshold of 5. This value exceeds the 20% threshold for maximum allowable error in cells. Therefore, it is again advisable to interpret the results with caution.

The selection frequencies of three attributes were found to differ significantly between setups with Pearson's χ^2 ($\chi^2(58, N = 566) = 91.110$, $p = 0.004$). *Height* and *timbral balance: dark–bright* were selected significantly more frequently for the "surround vs. surround with height" con-

Table 5. χ^2 statistics and attributes that were selected more frequently than expected for each reproduction system comparison, summed over spatial variation. Observations for each group are indicated in parentheses. p values below $p = 0.1$ are highlighted in bold.

Stereo vs. Surround $\chi^2(29, N = 201) = 126.611, \mathbf{p = 0.000}$	Stereo vs. Surround with Height $\chi^2(29, N = 220) = 57.091, \mathbf{p = 0.001}$	Surround vs. Surround with Height $\chi^2(29, N = 145) = 9.142, \mathbf{p = 0.002}$
Horizontal width	–	–
Spatial naturalness	Spatial naturalness	–
Envelopment	Envelopment	–
Surrounding	–	–
Position of sound	Position of sound	Position of sound
Ease of listening	–	–
–	–	Height

dition. Conversely, *horizontal width* was selected significantly less for the same comparison but significantly more for "stereo vs. surround with height." This observation is also reflected in the Cochran's Q results ($p = 0.012$), with post hoc analysis indicating a significant difference in *horizontal width* ($p_{horizontalwidth} = 0.022$) between the "stereo vs. surround" and "surround with height" comparisons.

The results show that reproduction system comparison had an influence on the attributes that were selected. This is evident in 1) the kind of attributes that were selected and 2) the number of attributes that were selected. It might also be concluded that the lower number of general selections in the "surround vs. surround with height" scene in combination with the lowest number of attributes selected over the expected frequency (*position of sound* and *height*) indicate a smaller perceived difference between reproduction systems. The results may also suggest that horizontal changes (e.g., "stereo vs. surround") might have a more dominant effect than height changes ("surround vs. surround with height"). This observation could be attributed to the perception that the compared systems with height changes are found to be more similar, a trend also shown by the BAQ rating results (Fig. 6).

2.3.3.2 Spatial Variation. Examining the attributes significantly selected over the expected frequency for each group of spatial variation (see Table 7), the results reveal an almost identical list, with the only differing attribute being *overall spectral balance*. Both the Pearson's χ^2 test conducted for this purpose ($\chi^2(29, N = 566) = 19.413$, $p = 0.910$) and a sign test—equivalent to Cochran's Q test for two groups, as mentioned in Meyners et al. [39]—were performed, and both yielded nonsignificant results. Consequently, spatial variation had no main-effect on attribute selection frequency.

2.4 Relationship of Attributes and BAQ ratings

Both the analysis of BAQ ratings (SEC. 2.2) and attribute selection frequency (SEC. 1.4) revealed that participants were primarily influenced by the change in reproduction system comparison. No main effect of spatial variation was detected for either of the dependent variables. However, it is worth mentioning that the results hint to a subtle interaction effect between the two independent variables on attribute selection frequency.

Given the experimental design, establishing a direct relationship between BAQ scores and the positive and negative influence of selected attributes is not possible. Nonetheless, evaluating the magnitude of quality rating scores for each attribute can provide valuable insights into the overall impact of these selected attributes. To investigate this, ratings were scaled to lie between -1 and 1, and the median absolute magnitudes of the BAQ ratings for each attribute were calculated. As before, only the data of the participants who were found to be consistent in the previous sections (see SECS. 2.1.1 and 2.1.2) were used. Accordingly, the analysis was conducted based on seven participants. Fig. 10 shows the median magnitudes plotted against the frequency of use for each attribute. The figure also shows the grand median, accompanied by matching MADs both above and below the median. In addition, it displays the lower and upper bounds of attribute selection frequency expected for a uniform distribution on the x-axis (as described in SEC. 2.3.1).

Table 6. Attribute selection differences across all reproduction system comparisons using two statistical methods. Provided values include test procedure p values and attributes differing among comparisons. For Pearson's χ^2, upward or downward triangles denote attributes selected significantly more or less for a specific system, respectively. In Cochran's Q, both omnibus test and attribute-specific p values are provided.

Method	Attribute	
Pearson's	Height	Stereo vs. surround ▼ surround vs. surround w. height ▲
χ^2	Horizontal width	Stereo vs. surround ▲ surround vs. surround w. height ▼
$p = 0.004$	Timbral balance: dark–bright	Surround vs. surround with height ▲
Cochran's Q	Horizontal width	$p_{horizontalwidth} = 0.022$
$p = 0.012$		Stereo vs. surround AND surround vs. surround w. height

Table 7. χ^2 statistics and attributes that were selected more frequently than expected for each spatial variation, summed over reproduction system comparison. Observations for each group are indicated in parentheses. p values below $p = 0.1$ are highlighted in bold.

Dynamic $\chi^2(29, N = 294) = 84.776, \boldsymbol{p = 0.000}$	Stationary $\chi^2(29, N = 272) = 77.853, \boldsymbol{p = 0.000}$
Horizontal width	Horizontal width
Spatial naturalness	Spatial naturalness
Envelopment	Envelopment
Surrounding	Surrounding
Position of sound	Position of sound
Ease of listening	Ease of listening
Overall spectral balance	–

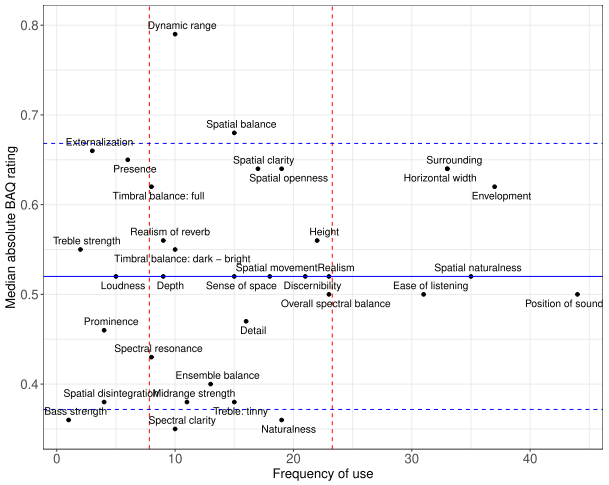


Fig. 10. Median of scaled absolute BAQ ratings summed over each attribute, plotted against frequency of attribute use. Grand median absolute BAQ ratings and MAD above and below the median are shown by the solid line and dotted lines on the y-axis, respectively. The two vertical dotted lines on the x-axis show the lower and upper bounds of attribute selection frequency expected for a uniform distribution.

Attributes selected more frequently than expected from a normal distribution displayed a tendency to concentrate around the grand median of absolute BAQ ratings, predominantly within the upper half of the MAD-region. Attributes within the expected frequency and those selected below it exhibit a more scattered distribution in terms of magnitudes but with a majority falling within the MAD-region around the grand median.

Although attributes that were rarely chosen and had low magnitudes (e.g., *bass strength*) may not be of direct concern, their distribution is still important to consider. High selection frequency often indicates important attributes, but some research scenarios may require examining attributes that were infrequently selected yet had a strong impact on the BAQ rating. Examples of such attributes include *dynamic range* and *spatial balance*, which fall within the expected selection frequency but display the highest magnitude of BAQ ratings. These observations underscore the need for a comprehensive assessment of attribute selection, acknowledging that the significance of an attribute can extend beyond its selection frequency, particularly in research contexts that demand a nuanced approach.

3 CONCLUSION

Upon reviewing previous research, it is evident that there has been growth and development in the implementation of various aspects of movement in modern audio program material. However, only a few studies have explored the use of audio material incorporating these aspects to evaluate audio quality of reproduction systems (e.g., [13]). Additionally, the impact of spatially dynamic properties on the evaluation methods of reproduction systems remains largely unknown. This study aimed to identify key attributes as guidelines for future research in the field of audio quality assessment. By doing so, it sought to contribute new insights, addressing challenges posed by advancements in modern audio technologies, particularly concerning new and complex spatial dynamic audio stimuli. To address these objectives, the following research questions were asked:

a) What are the most important attributes for characterizing the sound quality of reproduction systems when using spatially dynamic program material?
b) Do the important attributes differ between usage of stationary and dynamic program material?

The subsequent section will discuss the findings and provide answers to these research questions.

3.1 What Are the Most Important Attributes for Characterizing the Sound Quality of Reproduction Systems When Using Spatially Dynamic Program Material?

The following attributes were shown to be relevant in the context of audio quality assessment of reproduced sound when using spatially dynamic program material and may warrant additional exploration in future studies. The attributes with significant selection rates were *position of sound, envelopment, spatial naturalness, surrounding, horizontal width, ease of listening*, and *overall spectral balance*.

Moreover, additional attributes selected by the majority of participants but within expected frequency were *spatial*

movement, realism, spectral clarity, spatial clarity, sense of space, and *depth of field*. As it was found that attribute selection was influenced by the reproduction systems being compared, the use of reproduction system comparison-specific additional attributes (*height* and *timbral balance: dark–bright*) for the comparison of two similar higher channel reproduction systems might be considered. Additionally, the use of attributes such as *dynamic range* and *spatial balance* might be of interest, as they were used alongside high-magnitude quality ratings, even though they were selected by fewer participants and less frequently.

3.2 Do the Important Attributes Differ Between Usage of Stationary and Dynamic Program Material?

Statistical tests were employed to examine the influence of the independent variables (spatial variation and reproduction system comparison) as well as their interaction on attribute selections. The analysis revealed no significant impact of spatial variation on attribute selection, indicating that the important attributes did not differ between the usage of stationary and dynamic program material. Attribute selection differences were mainly driven by which reproduction systems were being compared. However, an interaction effect between the two main factors was identified. Evidence suggests that although spatial variation did not exhibit a main effect, it appears to influence attribute selection for different reproduction system comparisons in two ways. Generally, when dynamic audio is introduced, there is a reduced ability of the panel to clearly determine the attributes that are affecting their decisions, evident through a more uniform distribution and fewer attributes being selected more frequently than expected. However, there is one notable exception: the *position of sound*. This attribute stands out as having a higher selection rate when dynamic program material was used, an observation that might be rooted in the fact that it becomes more pronounced with dynamic stimuli.

3.3 Limitations

Although the results of the presented study give a good indication of which attributes are important for describing audio quality differences between reproduction systems using spatially dynamic audio program material, there are still some limitations to the study. The experiment only used one item of program material due to its complexity and length, requiring the streamlining of factors. Although this approach may impact generalizability, it was a necessary trade-off. Realism in movement patterns was also pursued to avoid artificial stimuli.

Only a small sample size of seven experienced listeners, all of whom are employees of Bang & Olufsen, participated in this study. This homogeneity in background may limit the generalizability of the results, particularly regarding the perception of sound attributes among a more diverse population. Experienced listeners (who were assumed to have a shared understanding of the attributes under test) were used in order to maximize the likelihood of producing consistent and statistically significant results.

3.4 Future Work

While establishing the presence and general impact of attributes important for describing audio quality differences between reproduction systems using spatially dynamic audio program material, this study does not investigate whether the attributes have a positive or negative effect on BAQ, offering an opportunity for focused investigations in subsequent research. Additionally, the use of a single scene and spatial variation represents a preliminary step; future studies should include a broader range of program items with different types of spatially dynamic elements to enhance generalizability and validate findings. Such an approach will strengthen the robustness of the results.

In this study, experienced listeners were chosen to ensure consistency and statistical significance in the results, given their shared understanding of the attributes being tested and their potential to identify nuanced differences among a large set of attributes. It may be beneficial to repeat the experiment with naive listeners in order to compare the findings. This could provide insights into the generalizability of the findings across different levels of expertise.

Future research should also include an investigation of attribute ratings in conjunction with audio quality ratings for a new set of stimuli. In this way, the important attributes suggested here can be verified and further explored. Building upon the insights gained from the current study and recognizing the challenges common methodologies face when applied to spatially dynamic material (cf. [1]), it becomes crucial to explore the suitability of these established experimental methods. This exploration will contribute to refining experimental approaches in the context of spatially dynamic content, addressing potential limitations and enhancing the overall validity of future research in this domain.

4 ACKNOWLEDGMENT

Special thanks are extended to the assessors for their dedicated time and contributions. This work has received funding by the European Union's Horizon 2020 Research and Innovation Programme under Marie Skłodowska-Curie Grant 956369, Service-Oriented Ubiquitous Network-Driven Sound (SOUNDS).

5 REFERENCES

[1] P. N. Porysek Moreta, S. Bech, J. Francombe, J. Østergaard, S. Van de Par, and N. Kaplanis, "Sensory Evaluation of Spatially Dynamic Audiovisual Soundscenes: A Review," presented at the *152nd Convention of the Audio Engineering Society* (2022 May), paper 10594.

[2] U. Scuda, H. Stenzel, and D. Baxter, "Using Audio Objects and Spatial Audio in Sports Broadcasting," in *Proceedings of the 57th International AES Conference: The Future of Audio Entertainment Technology–Cinema, Television, and the Internet* (2015 Mar.), paper 5-2.

[3] N. Zacharov, *Sensory Evaluation of Sound*, 1st ed. (CRC Press, Boca Raton, LA, 2018). https://doi.org/10.1201/9780429429422.

[4] International Telecommunication Union, "Methods for the Subjective Assessment of Small Impairments in Audio Systems," *Recommendation ITU-R BS.1116-3* (2015 Feb.).

[5] International Telecommunication Union, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534* (2000 Oct.).

[6] M. Neuendorf and F. Nagel, "Exploratory Studies on Perceptual Stationarity in Listening Test - Part I: Real World Signals from Custom Listening Tests," presented at the *131st Convention of the Audio Engineering Society* (2011 Oct.), paper 8562.

[7] F. Rumsey, "Subjective Assessment of the Spatial Attributes of Reproduced Sound," in *Proceedings of the AES 15th International Conference on Audio, Acoustics, and Small Spaces* (1998 Oct.), paper 15-012.

[8] D. Quiroz and D. Martin, "Mapping Methods of 3D Input Devices and Immersive Audio Environments Using Polar Coordinates for Panning," presented at the *150th Convention of the Audio Engineering Society* (2021 May), paper 10486.

[9] J. D. Mathew, S. Huot, and B. F. G. Katz, "Survey and Implications for the Design of New 3D Audio Production and Authoring Tools," *J. Multimodal User Interfaces*, vol. 11, pp. 277–287 (2017 Apr.). https://doi.org/10.1007/s12193-017-0245-z.

[10] J. Herre, J. Hilpert, A. Kuntz, and J. Plogsties, "MPEG-H 3D Audio—The New Standard for Coding of Immersive Spatial Audio," *IEEE J. Sel. Top. Signal Process.*, vol. 9, no. 5, pp. 770–779 (2015 Aug.). https://doi.org/10.1109/JSTSP.2015.2411578.

[11] International Telecommunication Union, "Methods for Selecting and Describing Attributes and Terms in the Preparation of Subjective Tests," *Recommendation ITU-R BS.2399* (2017 Mar.).

[12] A. Lindau, V. Erbes, S. Lepa, H.-J. Maempel, F. Brinkmann, and S. Weinzierl, "A Spatial Audio Quality Inventory (SAQI)," *Acta Acust. united Acust.*, vol. 100, no. 5, pp. 984–994 (2014 Sep.). https://doi.org/10.3813/AAA.918778.

[13] J. Francombe, T. Brookes, and R. Mason, "Evaluation of Spatial Audio Reproduction Methods (Part 1): Elicitation of Perceptual Differences," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 198–211 (2017 Mar.). https://doi.org/10.17743/jaes.2016.0070.

[14] R. Mason, "How Important Is Accurate Localization in Reproduced Sound?" presented at the *142nd Convention of the Audio Engineering Society* (2017 May), paper 9759.

[15] N. Zacharov and K. Koivuniemi, "Unraveling the Perception of Spatial Sound Reproduction: Techniques and Experimental Design," in *Proceedings of the 19th AES International Conference: Surround Sound-Techniques, Technology, and Perception* (2001 Jun.), paper 1929.

[16] G. Lorho, "Perceptual Evaluation of Mobile Multimedia Loudspeakers," presented at the *122nd Convention of the Audio Engineering Society* (2007 May), paper 7050.

[17] L. S. R. Simon, N. Zacharov, and B. F. G. Katz, "Perceptual Attributes for the Comparison of Head-Related Transfer Functions," *J. Acoust. Soc. Am.*, vol. 140, no. 5, pp. 3623–3632 (2016 Nov.). https://doi.org/10.1121/1.4966115.

[18] C. Guastavino and B. F. G. Katz, "Perceptual Evaluation of Multi-Dimensional Spatial Audio Reproduction," *J. Acoust. Soc. Am.*, vol. 116, no. 2, pp. 1105–1115 (2004 Aug.). https://doi.org/10.1121/1.1763973.

[19] J. Gil, S. Bech, and F. Christensen, "Temporal Evaluation of Sound Quality Using TCATA," presented at the *152nd Convention of the Audio Engineering Society* (2022 May), paper 10602.

[20] J. C. Castura, L. Antúnez, A. Giménez, and G. Ares, "Temporal Check-All-That-Apply (TCATA): A Novel Dynamic Method for Characterizing Products," *Food Qual. Preference*, vol. 47, Part A, pp. 79–90 (2016 Jan.). https://doi.org/10.1016/j.foodqual.2015.06.017.

[21] T. Walton, M. Evans, D. Kirk, and F. Melchior, "A Subjective Comparison of Discrete Surround Sound and Soundbar Technology by Using Mixed Methods," presented at the *140th Convention of the Audio Engineering Society* (2016 Jun.), paper 9592.

[22] J. Francombe, T. Brookes, R. Mason, and J. Woodcock, "Evaluation of Spatial Audio Reproduction Methods (Part 2): Analysis of Listener Preference," *J. Audio Eng. Soc.*, vol. 65, no. 3, pp. 212–225 (2017 Mar.). https://doi.org/10.17743/jaes.2016.0071.

[23] L. Hicks, S. Moulin, and S. Bech, "Sensory Profiling of High-End Loudspeakers Using Rapid Methods—Part 3: Check-All-That-Apply With Naïve Assessors," *J. Audio Eng. Soc.*, vol. 66, no. 5, pp. 329–342 (2018 May). https://doi.org/10.17743/jaes.2018.0015.

[24] S. Bech and N. Zacharov, *Perceptual Audio Evaluation—Theory, Method, and Application* (John Wiley & Sons, Ltd, Chichester, UK, 2006). https://doi.org/10.1002/9780470869253.

[25] ITU-R, "Advanced Sound System for Programme Production," *Recommendation ITU-R BS.2051-3* (2022 May).

[26] N. Kaplanis, S. Bech, S. Tervo, *et al.*, "A Rapid Sensory Analysis Method for Perceptual Assessment of Automotive Audio," *J. Audio Eng. Soc.*, vol. 65, no. 1/2, pp. 130–146 (2017 Jan./Feb.). https://doi.org/10.17743/jaes.2016.0056.

[27] J. Woodcock, C. Pike, F. Melchior, P. Coleman, A. Franck, and A. Hilton, "Presenting the S3A Object-Based Audio Drama Dataset," presented at the *140th Convention of the Audio Engineering Society* (2016 May), paper 255.

[28] BBC R&D and IRT, "EAR Production Suite," https://ear-production-suite.ebu.io (accessed Nov. 15, 2023).

[29] International Telecommunication Union, "Audio Definition Model Renderer for Advanced Sound Systems," *Recommendation ITU-R BS.2127* (2019 Jun.).

[30] A. Gabrielsson, B. Hagerman, T. Bech-Kristensen, and G. Lundberg, "Perceived Sound Quality of Reproductions With Different Frequency Responses and Sound Levels," *J. Acoust. Soc. Am.*, vol. 88, no. 3, pp. 1359–1366 (1990 Sep.). https://doi.org/10.1121/1.399713.

[31] ISO, "Acoustics—Methods for Calculating Loudness—Part 1: Zwicker Method," *Standard 532-1:2017* (2017 May).

[32] J. Francombe, T. Brookes, R. Mason, and F. Melchior, "Loudness Matching Multichannel Audio Program Material With Listeners and Predictive Models," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9464.

[33] International Telecommunication Union, "General Methods for the Subjective Assessment of Sound Quality," *Recommendation ITU-R BS.1284-2* (2019 Jan.).

[34] E. Parizet, "Paired Comparison Listening Tests and Circular Error Rates," *Acta Acustica united Acustica*, vol. 88, no. 4, pp. 594–598 (2002 Jul./Aug.).

[35] C. Leys, C. Ley, O. Klein, P. Bernard, and L. Licata, "Detecting Outliers: Do Not Use Standard Deviation Around the Mean, Use Absolute Deviation Around the Median," *J. Exp. Social Psychol.*, vol. 49, no. 4, pp. 764–766 (2013 Jul.). https://doi.org/10.1016/j.jesp.2013.03.013.

[36] H. T. Lawless and H. Heymann, *Sensory Evaluation of Food: Principles and Practices*, 2nd ed. (Springer, New York, NY, 2010). https://doi.org/10.1007/978-1-4419-6488-5.

[37] A. Field, *Discovering Statistics Using SPSS*, 2nd ed. (Sage Publications Ltd, London, UK, 2005).

[38] S. S. Mangiafico, "Summary and Analysis of Extension Program Evaluation in R," Version 1.20.07 (2016). https://rcompanion.org/documents/RHandbookProgramEvaluation.pdf.

[39] M. Meyners, J. C. Castura, and B. T. Carr, "Existing and New Approaches for the Analysis of CATA Data," *Food Qual. Preference*, vol. 30, no. 2, pp. 309–319 (2013 Dec.). https://doi.org/10.1016/j.foodqual.2013.06.010.

[40] C. W. Pike, *Evaluating the Perceived Quality of Binaural Technology*, Ph.D. thesis, University of York, York, UK (2019 Jun.).

THE AUTHORS


Pia Nancy Porysek Moreta


Søren Bech


Jon Francombe


Jan Østergaard


Steven van de Par

Pia Nancy Porysek Moreta completed her Bachelor of Engineering in Hearing Technology and Audiology at Jade University of Applied Sciences, Oldenburg, Germany, in 2017. She then earned her Master of Science in Hearing Technology and Audiology from Carl von Ossietzky University, Oldenburg, Germany, in 2020. Currently, she is a Ph.D. research fellow at Bang & Olufsen in Denmark, with an affiliation to Aalborg University. Her research is centered on audio perception within the "Service-Oriented, Ubiquitous, Network-Driven Sound (SOUNDS)" project, which is part of the Marie Skłodowska-Curie Actions European Training Network.

•

Søren Bech is Director of Research at Bang & Olufsen and Professor of Audio Perception at Aalborg University's Section AI and Sound, The Technical Faculty of Information Technology and Design. He received a Master of Science and a Ph.D. from the Technical University of Denmark. He has authored over 52 peer-reviewed papers, more than 100 conference contributions, and led 19 funded international research projects. A Fellow of the Acoustical Society of America and the Audio Engineering Society (AES), he has served as Governor and Vice-President of the AES and currently acts as associate technical editor for the AES Journal. He has also been vice-chair of the International Telecommunication Union working group 10/3. His research interest includes psychoacoustics and in particular, human perception of reproduced sound in small and medium-sized rooms. Other interests include experimental procedures and statistical analysis of data from sensory analysis of audio and video quality.

•

Jon Francombe graduated with a Bachelor of Engineering in Music and Sound Recording (Tonmeister) from the University of Surrey, Guildford, United Kingdom, in 2010, and received a Ph.D. in perceptual audio quality evaluation from the same institution in 2014. He then worked as a research fellow on the Engineering and Physical Sciences Research Council-funded "S3A: Future Spatial Audio for an Immersive Listener Experience at Home" project, investigating perceptual attributes of spatial audio reproduction and new methods for immersive audio reproduction. This was followed by four years in the Audio Team at British Broadcasting Corporation Research & Development, where he co-led the team and continued development of technology for delivering spatial audio experiences over connected personal devices. He currently works as a senior specialist in audio perception at Bang & Olufsen, where he designs and conducts listening experiments and supervises various research projects.

•

Jan Østergaard received a Master of Science in Electrical Engineering from Aalborg University, Denmark, and a Ph.D. from Delft University of Technology, The Netherlands. He worked as a Research and Development Engineer at ETI A/S in Aalborg and ETI Incorporated in Virginia, USA. He also served as a Postdoctoral Researcher at The University of Newcastle in Australia. Currently, he is a Full Professor in Information Theory and Signal Processing, Head of the Section on Artificial Intelligence and Sound, and Head of the Centre for Acoustic Signal Processing Research at Aalborg University. He has received the Danish Independent Research Council's Young Researcher's Award, the Best Ph.D. Thesis award from the European Association for Signal Processing, and fellowships from the Danish Independent Research Council and the Villum Foundations Young Investigator Programme. He is an Associate Editor for Signal Processing and Source Coding for the IEEE Transactions on Information Theory.

•

Steven van de Par studied Physics at the Eindhoven University of Technology, the Netherlands, and completed a Doctor of Philosophy project on binaural detection followed by a postdoctoral project on auditory-visual synchrony perception. In 2000, he joined Philips Research to work on low-bitrate audio coding, music information retrieval, and computational auditory scene analysis. Since 2010, he has been a professor in acoustics at the Carl von Ossietzky University in Oldenburg, Germany. His current research includes spatial perception, applied psychoacoustics, computational auditory scene analysis, and virtual acoustics.