OPEN ACCESS Freely available online

P. N. Porysek Moreta, S. Bech, J. Francombe, J. Østergaard, and S. van de Par, "Continuous and Overall Evaluation of Spatial Audio Reproduction Systems With Spatially Dynamic Content," *J. Audio Eng. Soc.*, vol. 74, no. 4, pp. 199–211 (2026 Apr.). https://doi.org/10.17743/jaes.2022.0258.

Continuous and Overall Evaluation of Spatial Audio Reproduction Systems With Spatially Dynamic Content

PIA NANCY PORYSEK MORETA,[1, 2, *] (p.porysek@gmail.com) **SØREN BECH,**[2] *AES Fellow* (sbe@es.aau.dk), **JON FRANCOMBE,**[1] (JOFR@bang-olufsen.dk)

JAN ØSTERGAARD,[2] (jo@es.aau.dk) **AND STEVEN VAN DE PAR**[3] (steven.van.de.par@uol.de)

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

This study addresses potential challenges in evaluating reproduction systems with complex, spatially dynamic audio material because current standardized methods may lead to biased or unreliable results. To investigate this, a listening experiment was conducted comparing two assessment methods—continuous and overall evaluation—applied to two attributes, basic audio quality and surrounding, using spatially dynamic content reproduced on two reproduction systems (a stereo and a 3D surround configuration). To enable comparison with overall evaluations, continuous ratings were summarized using different strategies. Although ratings were mainly influenced by the reproduction system, spatial variation, and program item, the choice of evaluation method also had a significant effect on the scores. This effect varied depending on the attribute and the metric used to summarize the continuous data. In cases where differences were observed, continuous evaluations consistently produced higher scores than overall ratings, regardless of the summary metric. These findings indicate that continuous evaluation can capture perceptual variations over time that are lost in overall ratings, suggesting it can be a useful approach when assessing attributes influenced by spatially dynamic changes.

Keywords: listening experiments, perceptual audio quality, spatial audio evaluation

1 INTRODUCTION

State-of-the-art spatial audio technology enables more complex audio content, such as spatially dynamic program material. The term "spatially dynamic program material" describes content characterized by noticeable spatial changes over time, particularly object position shifts (i.e., movement of sound sources) [1]. There is evidence suggesting that standardized methods of perceptual audio quality evaluation may encounter challenges when applied with these stimuli, resulting in unreliable outcomes influenced by various biases [2]. This paper explores the validation of methodologies that use single-value or overall rating methods, which are commonly used in the standardized audio experiment paradigms, in comparison to the less commonly used instantaneous or continuous evaluation (e.g., single-stimulus continuous quality evaluation [3], in which the stimulus is measured continuously over time). The comparison is applied to reproduction systems using spatially dynamic program material.

To minimize biases and rating errors in standardized methodologies [4–6], literature suggests selecting program material that is both low in complexity [7] and short in duration [6]. The complexity of the stimulus has been described as a possible cause of confusion, making the listener unable to clearly follow a given task [7]. A low stimulus complexity is best achieved if the stimulus does not exceed a certain duration. According to Zielinski et al. [8] a longer stimulus is more prone to timbral and spatial changes and therefore makes it harder for the listener to "average" the experience over time.

Auditory memory, which is crucial for distinguishing perceptual differences, has a maximum span of a few seconds [9]. As a result, listeners are prone to primacy and recency effects [2]. These effects describe biases where assessors are influenced by the beginning [10] or the end

*To whom correspondence should be addressed, email: p.porysek@gmail.com.

of a recording [8], driven by strong initial impressions or recent events. Ratings can swing positively or negatively. Assessors may overlook occasional degradations if the final impression remains strongly positive ("forgiveness effect" [11]), whereas impairments at the end can disproportionately affect the overall evaluation [12]. Such temporal biases align with ecological perspectives on listening, which emphasize that perception involves continuous adaptation to the unfolding auditory environment [13, p. 4–5].

Allowing assessors to isolate specific loops of the content is common for direct comparisons [8]. However, this can introduce errors, especially with complex or longer stimuli. For instance, Schinkel-Bielefeld [14] studied the impact of loop selection on the evaluation of music in 5.1 surround sound at bit rates between 80 and 256 kbps using the International Telecommunication Union (ITU) recommendation ITU-R BS.1534 (MUSHRA) method [6]. The study found that assessors (a) selected different loops for evaluation, and (b) gave significantly different ratings for the same stimulus, depending on the loop selected. This suggests that isolated loops of the same signal can vary greatly in spectral and temporal properties. Because there is no guarantee that all assessors select the same loop, this can lead to biased evaluations.

By avoiding complex stimuli, it is assumed that the investigated aspect of sound quality can be considered as stationary enough to be an accurate representation of the whole sound excerpt under evaluation [15]. Stimuli are chosen based on perceptual constancy, and dynamic aspects (e.g., spatially dynamic content) are avoided due to being considered a distraction [2].

One type of more complex and spatially dynamic stimulus is spatial reproduction involving any kind of movement. Although spatial audio and movement in sound (e.g., panning) are not new concepts in audio reproduction, technical development of production, transmission and reproduction technologies for 3D audio (or "next generation" spatial audio) led to an increasing interest both on the commercial and consumer side [2]. Examples of these developments are the standardization of the MPEG-H 3D audio format [16], which supports multiple immersive sound signal formats, the rising popularity of Dolby Atmos outside of cinemas and in combination with music, and the development of Sony's 360 Reality Audio format. With the growing availability of tools and research to support the development of new spatially dynamic content [17], the need for effective evaluation of reproduction systems with this increasingly complex material becomes even more essential.

The approaches to addressing this need, discussed in more detail by Porysek Moreta et al. [2], include continuous evaluation [18, 19]. This method is mentioned in the literature as a way to reduce biases in current methodologies, especially when dealing with complex audio content [8]. However, it has rarely been applied to audio stimuli.

Continuous evaluation, involving the tracking of changes over time, was first used in food evaluation as time-intensity methods [20, p. 179 ff.] and later adapted for television [3] and speech quality assessment [21]. It has been shown to significantly reduce the recency effect [19]. However, to date, continuous evaluation has mostly been used for measuring "overall" speech quality in audio stimuli.

The objective of this work is to investigate the impact of spatially dynamic program material on two attributes of audio quality: *basic audio quality* (BAQ), selected as a baseline measure, and *surrounding*, chosen for its relevance to spatial perception [1], as discussed in SEC. 2.4. This study focuses on reproduced sound. Specifically, it will address the following research question: Do ratings of audio quality attributes for reproduction systems with spatially dynamic content differ when made using continuous evaluation versus overall rating methods?

To answer this research question, a listening experiment was designed. The paper is structured as follows: first, the methodology is presented (SEC. 2), including the physical setup, stimuli creation, attribute selection, experimental procedure, and assessor information. The results section (SEC. 3) follows, detailing data preprocessing and analysis. The discussion section (SEC. 4) then explores the findings and their implications. Finally, the conclusion (SEC. 5) provides a brief summary of the work, outlines its limitations, and highlights directions for future research.

2 METHODOLOGY

Two evaluation methods were applied in the experiment, each assessing the audio quality attributes *BAQ* and *surrounding* across two reproduction systems with varying spatial content. Ratings were collected following the approach of Gros and Chateau [19]. The first method involved overall ratings, where assessors could adjust their rating during or after listening to the entire audio sample, finalizing their score at the end [6]. The second method involved continuous ratings, following the single-stimulus continuous quality evaluation approach [3], where assessors provided real-time judgments of the audio stimulus.

To introduce a range of perceivable spatial variation in the audio material, two program items were manipulated, as described by Porysek Moreta et al. [1]. Two versions of each item were created: one with stationary spatial characteristics and one with dynamic characteristics. These items were reproduced over two reproduction systems that differed in the number of loudspeakers. The following sections provide more details about the materials and methodology of the experiment.

2.1 Physical Setup

The experimental setup closely followed that used by Porysek Moreta et al. [1]. To minimize the influence of the room's acoustics on the evaluated sound field, an anechoic chamber (6.07 × 4.95 × 2.91 m) was used [15, p. 246]. The loudspeakers used in the experiment were arranged on a loudspeaker sphere with a radius of 1.35 m from the listener's position.

Reproduction system layouts aiming to replicate the setups commonly found in modern domestic spatial audio systems were selected. Based on Porysek Moreta et al.'s findings [1], which revealed no significant difference in

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

Loudspeaker layout	Channel	Azimuth	Elevation
0 + 2 + 0	Left/right	±30	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

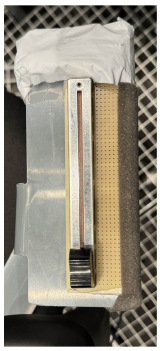


Fig. 1. Analog slider fixture.

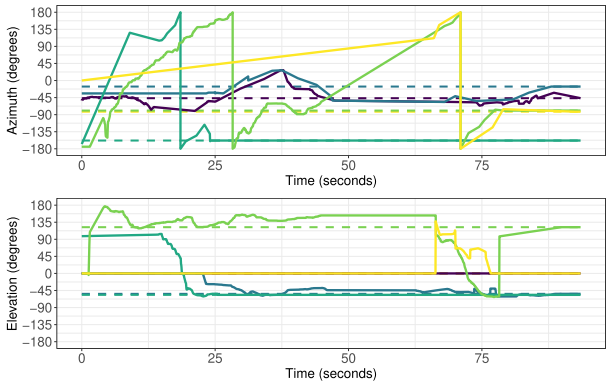


Fig. 2. The movement paths (azimuth and elevation) of four audio objects showing the most prominent changes in object positions in the dynamic version (solid lines) and their stationary equivalents (represented by dotted lines) of the audio excerpt from the scene titled *Family*.

perceived *BAQ* between surround and surround with height setups, two specific loudspeaker layouts adhering to ITU-R BS.2051 [22] were selected for this study: a 3D surround configuration (*4 + 7 + 0*) and a standard stereo configuration (*0 + 2 + 0*). The loudspeakers used were Genelec 8320A and 8331A. The azimuth and elevation coordinates for the configurations are detailed in Table 1. The loudspeaker array calibration followed the methodology outlined by Kaplanis et al. [23].

Preprocessed and rendered audio stimuli were assessed using an interface developed in Cycling 74's Max software, running on a desktop computer (OMEN X 900; HP) located in the control room. This setup was connected to the loudspeaker sphere in the anechoic chamber via an audio interface (RME M-32 DA, M-16 DA).

Assessors were seated on a fixed chair positioned such that their heads were located at the center of the sphere. The interface was mirrored from the control room's monitor to an iPad that the assessors used on a small lap desk. To record assessors' ratings, an analog slider (see Fig. 1) was mounted on the right side of the chair. The slider was built with a Teensy 4.0 microcontroller and a 100-mm travel fader, all assembled on a prototyping board. This analog slider was connected to a computer in the control room, enabling assessors to adjust the visual slider on the interface.

2.2 Stimulus Creation

Two 90-s excerpts of two different audio drama scenes from the S3A object-based audio drama dataset [24] were selected for stimulus creation. The two scenes were chosen because they represent two contrasting use cases: one centered on indoor dialogue with character movement, and one focused on outdoor ambient motion, offering complementary examples for evaluating dynamic spatial reproduction. The scenes will further be referred to as the *Family* and the *Forest* scenes.

The *Family* scene features a conversation among a family of four spread across two floors of a house. The characters engage in dialogue, run downstairs, and move between rooms, accompanied by the ambient sounds typical of a household. The *Forest* scene tells the story of two children playing in a forest. It includes a narrator's voice, the sounds of the children's calls and footsteps as they run and play in the forest, and general forest ambiance, music, and the ominous approach of a monster.

Using the method from Porysek Moreta et al. [1], two versions of each program item were created: a dynamic version with amplified spatial trajectories and/or elevated sound levels for tracks exhibiting the most movement and a stationary version with fixed audio objects and no movement patterns. The end position of each audio object was chosen as the stationary point because they aligned more logically with the positions of the speakers in the conversation, making it easier for the listener to understand the scene. The audio objects with the most pronounced movement patterns (azimuth and elevation) within the scene are visualized in Figs. 2 and 3.

Both scenes featured distinct horizontal and vertical movement patterns, demonstrating different potential use cases for spatial audio. The patterns were edited using metadata with the audio workstation REAPER and the virtual

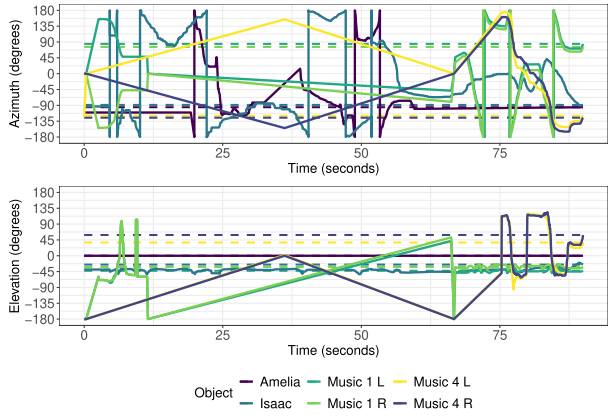


Fig. 3. The movement paths (azimuth and elevation) of four audio objects showing the most prominent changes in object positions in the dynamic version (solid lines) and their stationary equivalents (represented by dotted lines) of the audio excerpt from the scene titled *Forest*.

studio technology plug-in EAR Production Suite [25]. The two scenes (dynamic/stationary) were then rendered for both selected reproduction system setups at a sampling frequency of 48 kHz. For offline rendering, the ITU Audio Definition Model renderer [26] was used. This resulted in six stimuli (two spatial variations of the two program items for each of two reproduction systems).

2.3 Loudness Matching

Loudness matching was performed to eliminate potential biases in the results due to variations in loudness among the material [27] and reproduction systems [28]. Following Porysek Moreta et al. [1], loudness was estimated from initial playback recordings of the unadjusted stimuli using the Zwicker model (International Organization for Standardization Standard ISO 532-1) for time-varying sounds [29]. The maximum predicted loudness (in Sone) was converted to Phon and compared with a target of 93 Phon, which was selected in preliminary tests to balance audibility of the complex scenes with listener comfort. The resulting gain factor was then applied to the corresponding preprocessed digital stimuli.

After loudness matching, the measured equivalent A-weighted sound pressure levels ($L_{A,eq}$) at the listening position ranged from 65.4 to 67.0 dB for the *Forest* and 69.3 to 71.4 dB for the *Family*. This range aligns with the typical preferred or comfortable listening levels for music, which generally range from 65.0 to 85.0 dB [15].

2.4 Attribute Selection

The attribute *BAQ* was preselected because it provides a broad evaluation of the overall audio experience and serves as a baseline for comparison against other attributes. The definition of *BAQ* was as follows: "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, reverberance, echoes, harmonic distortions, quantization noise, pops, clicks and background noise. Scale: Bad–Excellent" [30].

To complement *BAQ* with an additional descriptive attribute, the six attributes identified by Porysek Moreta et al. [1] as important for assessing audio quality with spatially dynamic content were considered. A pre-experiment was conducted with two expert listeners to identify the most suitable descriptive attribute for this purpose.

Assessors listened to all stimuli and were asked to 1) identify changing attributes within and between stimuli, and 2) rank their importance. All attributes were selected at least once, showing their relevance in describing stimulus differences. Attributes chosen by both assessors, particularly those related to changes within a stimulus, were considered further. *Envelopment* was selected for the *Forest* scene, and *position of sound* and *surrounding* were chosen for the *Family* scene.

The second task, ranking attributes by importance, identified *surrounding* as the most important. Expert discussions revealed challenges in ranking attributes, such as *envelopment*, *surrounding*, and *horizontal width*, due to their conceptual similarities. Given these overlaps, *surrounding* was chosen for the experiment because its clear definition made it suitable for nonexpert listeners. The definition of *surrounding* was "the spread of sound around the space. Scale: unidirectional (coming from one direction)–surrounding" [31].

2.5 Experiment Procedure

The experimental procedure spanned two sessions conducted on separate days, each utilizing a distinct evaluation method: session A (continuous evaluation) and session B (overall evaluation). To mitigate order bias, the group of assessors was divided, with half starting with session A and the other half with session B.

Each session was divided into two blocks, with each block focusing on the evaluation of a specific attribute. The first block included stimulus familiarization, method familiarization, and the rating phase. For the second block, method familiarization was omitted. Both blocks included a complete repeat, and the trial order was randomized within each attribute. Each session ended with a short questionnaire. The total duration of one session was approximately 1.5 h, resulting in a total experimental duration of 3 h per assessor. The following sections provide a more detailed explanation of the four steps (familiarization with the stimuli, familiarization with the interface and method, rating phase, and postsession questionnaire) involved in each session.

2.5.1 Stimulus Familiarization

In the stimulus familiarization phase, assessors were provided the opportunity to listen to the specific audio excerpts they would later evaluate, focusing on the attribute under test (Fig. 4). Assessors could freely switch between stimuli and were asked to "use the slider on the right-hand side of your chair [...] to find a position on the scale that represents your perception of the attribute in this stimulus." This step was important for helping assessors familiarize themselves

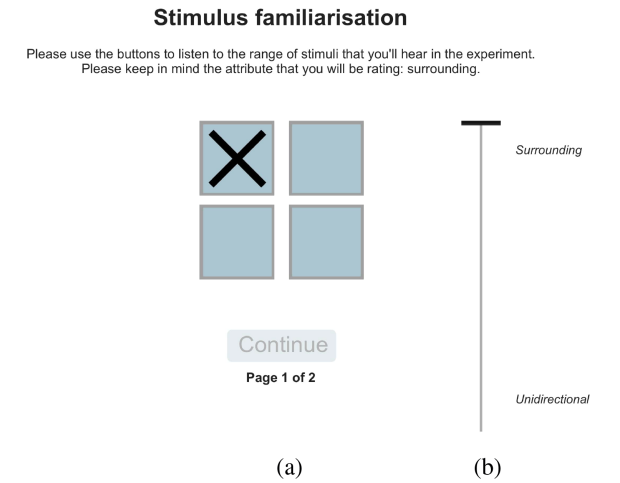


Fig. 4. First of two familiarization pages. (a) Center: Buttons for audio playback. First stimulus is currently playing (button is activated marked by the black x). (b) Right: Rating scale for attribute *surrounding* with slider marker at the top. The "Continue" button is deactivated during audio playback.

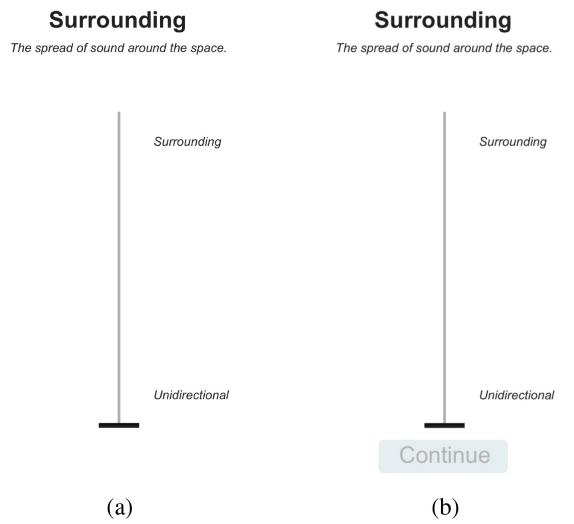


Fig. 5. Comparison of interface sliders for both methods. (a) Session A—continuous method. (b) Session B—overall method. The "Continue" button is deactivated during audio playback.

with both the content and range of the attribute's intensity within the excerpts, enabling them to effectively use the attributes and scale in the main experiment.

The stimulus familiarization process was repeated twice per session at the start of each block for each of the two attributes separately. This phase included two pages, each with four buttons, one page for each scene and its variations. Each stimulus had to be played for at least 15 s before the assessor was able to proceed.

2.5.2 Method Familiarization

After the stimulus familiarization, assessors were introduced to the experimental interface, which differed slightly depending on the session (session A: continuous evaluation; session B: overall evaluation). Each session included one training run equivalent in duration to a full repeat. The familiarization phase was crucial in acquainting the assessors with the tools and methods they would use during the main experiment; more specifically, it acquainted them with how to use the slider for each method.

The interface for both sessions included a scale with a movable marker, controlled by a physical slider for the ratings, and a definition of the attribute under test. Both methods used a dynamic scale with labels placed toward the top and bottom of the scale. The scale was for *BAQ* was labeled "excellent" at the top and "bad" at the bottom. For *surrounding*, the labels were "surrounding" and "unidirectional".

The task for session A (continuous method) was as follows: "Before each trial you will see a countdown and then hear an audio excerpt. Please use the slider to rate the attribute you are asked about. Leave your hand on the slider and move the slider while the excerpt is playing, so that it always reflects your current perception of the attribute." Each trial began with the slider at its lowest position, the defined start position. A 2-s countdown preceded the audio playback and rating phase. If the slider was moved from its starting position before the countdown finished, the trial was restarted. The rating phase lasted for the duration of the stimulus, during which assessors could provide ratings [Fig. 5(a)]. As soon as playback ended, a new page was displayed, asking assessors to reset the slider to the start position, after which the countdown began again.

The task for session B (overall method) was as follows: "In the rating phase each stimulus will be presented individually. Your task is to provide a single rating for the attribute under test for each of the played stimuli." As in session A, trials began with the slider at the defined start position, and a 2-s countdown preceded playback. Assessors could adjust their rating at any point during playback or afterward. The next trial was initiated by confirming the rating by pressing "Continue" and returning the slider to the start position [Fig. 5(b)].

2.5.3 Rating Phase

In the main rating phase, assessors applied the rating method that had been demonstrated during the familiarization phase. Each assessor completed 128 judgments, 64 in each session.

2.5.4 Postsession Questionnaire

After completing each session, assessors filled out a brief questionnaire. The questionnaire collected basic demographic information and included items assessing their judgments of task difficulty and level of focus. In addition, assessors were given an open-ended text field where they could provide general impressions and comments about the session.

3 RESULTS

Twenty assessors (aged 25 to 63 years, mean age 38) participated in the study, all of whom self-reported nor-

mal hearing. The group consisted of 17 males and three females, with varying levels of critical listening experience as indicated by their professional background in audio and acoustics and prior participation in audio experiments. Each assessor completed two experimental sessions on separate days, with each session lasting between 60 and 90 minutes.

During the experiment, the slider was operated only with the right hand. The group of assessors included three left-handed assessors, 16 right-handed assessors, and one ambidextrous individual. No performance differences were observed that could be attributed to handedness. A comparison of the results with and without the three left-handed assessors confirmed that their inclusion did not affect the statistical outcomes, and their questionnaire responses indicated no increased task difficulty. This aligns with research showing that left-handed and ambidextrous individuals often exhibit strong bilateral motor control [32]. For these reasons, handedness was not further considered in the analysis.

Before analysis, the results were preprocessed (SEC. 3.1), and the data series for the continuous results for each trial were summarized for comparison with the single values of the overall method (SEC. 3.2). Based on the summary metric analysis, the results for both methods were compared using analysis of variance (ANOVA) (SEC. 3.3 and SEC. 3.4).

3.1 Preprocessing of Continuous Data

Answers collected through the physical slider were recorded using MIDI continuous control messages with a resolution of 128 steps (ranging from 0 to 127) corresponding to a step size of ∼0.787 mm on the 10-cm slider movement range. New data points were generated each time the slider was moved, capturing the position and time stamp of the movement. Due to the nature of how results were recorded, depending on how frequently the slider was moved in a trial, the length of the result vectors for each trial varied.

The time stamps were logged with a resolution of 1 ms. However, in rare cases, two or three values were logged at the same time stamp. To maintain a consistent sampling rate of 1000 samples per second across all trials, only one value was assigned per time stamp by taking the mean of the values recorded within the same millisecond.

Subsequently, forward filling was applied to bring all result vectors to the same length. Forward filling is a method where missing values are replaced by the most recent recorded value. This approach is based on the instructions given to assessors: if no changes are recorded, it is assumed that there are no perceived changes ("the slider position always reflects your current perception of the attribute"). Thus, data are not added artificially. Instead, by "filling the gaps" parts of the signal that were not recorded due to the technical construction of the system were recovered.

A reduction of noisy high-frequency fluctuations in the signals was achieved by applying a linear-phase low-pass filter, designed using the windowed-sinc method at a sampling rate of 1 kHz. This included eliminating fluctuations introduced by the physical slider, which occasionally produced artifacts at frequencies of 5 Hz and above due to rapid transitions between neighboring values while the assessor was not intentionally moving the slider. The filter was designed with a cutoff frequency of 1 Hz because plotting the magnitude responses suggested that most of the signals' energy was concentrated below this threshold. A Hann window was applied to smooth the filter's response. The filter window length was set to 500 samples, balancing frequency selectivity and temporal smoothness. Due to the linear-phase nature of the finite impulse response filter, a group delay of 250 ms (corresponding to half the 500-sample window length) was introduced; this delay was compensated for by discarding the first 250 samples of the filtered signal.

Because assessors required different amounts of time to adjust to the initial sound impressions from the starting position, the first 30 s of data were removed. This decision was based on findings by Gros and Chateau [19], which suggest that the time needed to adapt to a negative or positive change ranges between 10 and 30 s. After compensating for the 250-ms filter delay and discarding the first 30 s, the resulting filtered signals had a duration of 59.75 s.

3.2 Selection of Summary Metric

To compare overall ratings with continuous results, several metrics were used to summarize the preprocessed continuous data into a single comparable value for each trial. A selection of relevant metrics was drawn from the literature, covering some known occurring biases in standard audio reproduction system evaluation.

Overall ratings after listening to stimuli are usually assumed to reflect the mean experience of the testing condition. Therefore, the first metric used to summarize the continuous data was the mean of all data points.

Because standard listening experiments typically use stimuli lasting at least 10 s (ITU-R BS.1534-1 [6]), one summary metric was defined as the average over the final 10 s of each trial, capturing potential recency effects [8]. Additionally, the last recorded value in each result vector was included as a separate metric. This value may be particularly relevant because it reflects the final position at which the slider was left—potentially representing a deliberate evaluative choice, and aligning more closely with traditional end-of-trial rating methods. Further, the metrics included extreme values, such as the maximum and minimum, because these may influence final decisions due to particularly strong sensations [19].

To assess whether the different summary statistics capture different aspects of the data, their correlations with the *mean* metric were calculated. For both *BAQ* and *surrounding* (Fig. 6), all metrics exhibit a strong positive correlation with the *mean*, ranging from 0.802 to 0.905 for *BAQ* and from 0.896 to 0.971 for *surrounding*.

In addition to reporting correlation coefficients, scatter plots were inspected to examine the distribution of values and ensure that the high correlations were meaningful.

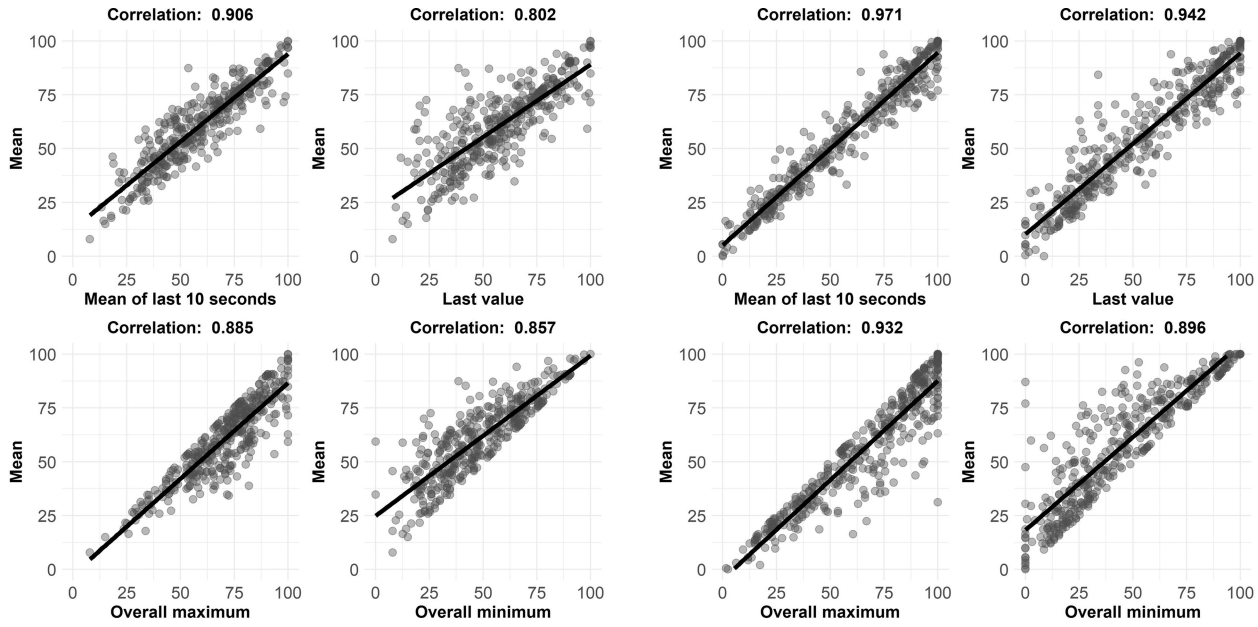


Fig. 6. Correlation plots of *mean* vs. different statistical metrics to summarize the continuous data. (a) Correlation plots for attribute *BAQ*. (b) Correlation plots for attribute *surrounding*.

A truncated pattern was observed for the *maximum* and *minimum* metrics: maximum values rarely fell below the mean, especially when the mean was low, and minimum values rarely exceeded the mean, especially when the mean was high. This effect is expected because the maximum and minimum are by definition constrained relative to the mean.

3.3 Results: Basic Audio Quality

An ANOVA was conducted to examine the effects of the independent variables (*method*, *spatial variation*, *program item*, and *reproduction system*) on the dependent variable *BAQ*, with a particular focus on the employed method.

Initially, the ANOVA was applied to all tested summary metrics. As expected from the high intermetric correlations (see SEC. 3.2), the *mean*, *final 10-s mean*, and *last recorded value* produced almost the same set of significant and nonsignificant effects. In contrast, the *maximum* and *minimum* metrics showed slight deviations. The following section, therefore, focuses on the *mean* metric and includes selected deviations observed for the *maximum* metric for comparison.

The analysis was conducted using a linear mixed-effects model that included the fixed effects of *method*, *spatial variation*, *program item*, and *reproduction system*, along with all two-way interactions between these factors. Assessors were modeled as a random effect, with random intercepts for each assessor and for each assessor-by-factor combination. No interaction terms were included in the random effects.

This model was selected after an initial full model, which also contained all possible three-way and four-way interactions, showed that higher-order interactions were either nonsignificant or negligible, with F-values at least seven times smaller than those of the main effects and two-way interactions. The reduced model resulted in a slight decrease in model fit (measured by marginal and conditional R^2, representing variance explained by fixed effects and by both fixed and random effects, respectively) for the *mean* metric compared with the complex model (cf., Table 2). However, these differences were marginal ($\Delta R^2 < 0.01$), and the simplified model was considered nearly equivalent in explanatory power.

Table 2. Comparison of marginal and conditional R^2 values for complex and reduced ANOVA models for the attribute *BAQ*.

Metric	ANOVA model	R_m^2	R_c^2
Mean	Complex	0.190	0.502
	Reduced	0.183	0.496
Maximum	Complex	0.220	0.513
	Reduced	0.213	0.506

Residuals from the reduced model were tested for normality using the Shapiro-Wilk test. For both metrics, deviations were observed, but Q-Q plots revealed largely linear patterns. Although minor deviations at the distribution tails were present, there was no strong evidence of skewness or heavy tailedness, and the normality assumption was deemed sufficiently satisfied.

ANOVA results for both metrics (Table 3) show significant main effects ($p < 0.05$) of *spatial variation*, *program item*, and *reproduction system* on both *mean* and *maximum* metric *BAQ* scores. Both metrics have significant in-

Table 3. ANOVA table of the fixed effects of the linear mixed-effects model for *BAQ*.

	Mean			Maximum		
Effect	DF	F	p	DF	F	p
Method	(1, 19.69)	1.64	0.215	(1, 20.78)	27.13	**0.000**
Spatial variation	(1, 22.91)	5.81	**0.024**	(1, 22.32)	5.16	**0.033**
Program item	(1, 20.83)	21.18	**0.000**	(1, 20.67)	19.25	**0.000**
Reproduction system	(1, 25.47)	16.47	**0.000**	(1, 24.86)	19.13	**0.000**
Method × spatial variation	(1, 533.25)	2.85	0.092	(1, 533.21)	4.94	**0.027**
Method × program item	(1, 533.25)	2.70	0.101	(1, 533.21)	0.94	0.332
Method × reproduction system	(1, 533.25)	0.06	0.812	(1, 533.21)	0.02	0.883
Spatial variation × program item	(1, 533.25)	41.38	**0.000**	(1, 533.21)	37.95	**0.000**
Spatial variation × loudspeaker setup	(1, 533.25)	2.38	0.123	(1, 533.21)	2.30	0.130
Program item × reproduction system	(1, 533.25)	4.94	**0.027**	(1, 533.21)	4.30	**0.039**

Note. Values with $p < 0.05$ are emphasized in bold.
DF = degrees of freedom; DenDF = denominator DF (residual/error); NumDF = numerator DF (effect).

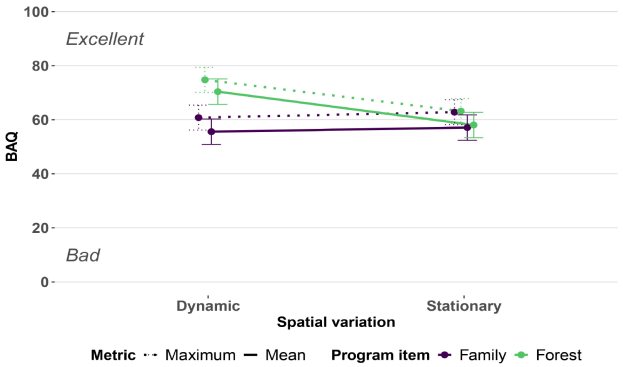


Fig. 7. Interaction effect between *spatial variation* and *program item* for *maximum* and *mean* metrics for the attribute *BAQ*.

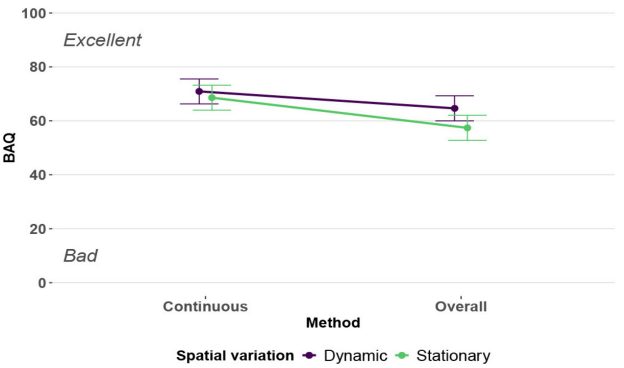


Fig. 8. Interaction effect between *method* and *spatial variation* for the maximum metric for the attribute *BAQ*.

teractions for *program item* × *reproduction* system and the strongest interaction for *spatial variation* × *program item*.[1]

Post hoc analysis (Fig. 7) revealed that for the *Forest* program item the *dynamic* condition produces significantly higher *mean* and *maximum* metric scores compared with the *stationary* condition ($p = 0.000$). Additionally, within the *dynamic* condition, *Forest* was rated significantly higher than *Family* ($p = 0.000$), showing that spatial variation has a stronger effect depending on the content.

For the *maximum* metric, *method* yields a significant main effect and additionally shows an interaction with *spatial variation* (Fig. 8). Post hoc comparisons show that in the case of the *maximum* metric the *continuous* method resulted in significantly higher *BAQ* scores compared with the *overall* method for both spatial variation conditions ($p = 0.003$ for *dynamic* and $p = 0.000$ for *stationary*). A significant difference between *dynamic* and *stationary* spatial variation is observed only within the *overall* method ($p = 0.000$), with the *dynamic* condition rated higher.

[1] Note that half of the dataset remains unchanged between the two metric ANOVAs because only the summary method of the continuous evaluation differs, whereas the underlying data contributing to the overall results remain the same.

Table 4. Comparison of marginal and conditional R^2 values for complex and reduced ANOVA models for the attribute *surrounding*.

Metric	Model	R^2_m	R^2_c
Mean	Complex	0.657	0.881
	Reduced	0.657	0.881
Maximum	Complex	0.643	0.875
	Reduced	0.641	0.874

3.4 Results: Surrounding

The same modeling procedure as in SEC. 3.3 was applied to the attribute *surrounding*, using the identical fixed and random effects structure. As with *BAQ*, a comparison between the complex and simplified ANOVA models revealed minimal differences in model fit (see Table 4), with $\Delta R^2 < 0.01$ in all cases. Thus, the reduced model was again deemed sufficient in capturing the key effects.

Residuals from the reduced model were tested for normality using the Shapiro-Wilk test. Residuals from the *mean* metric did not significantly deviate from normality, whereas those from the *maximum* metric did. However, Q-Q plots again showed largely linear patterns, and the normality assumption was considered sufficiently met.

Table 5. ANOVA table of the fixed effects of the linear mixed-effects model for *surrounding*.

	Mean			Maximum		
Effect	DF	F	p	DF	F	p
Method	(1, 21.97)	2.97	0.099	(1, 23.41)	29.35	**0.000**
Spatial variation	(1, 22.03)	34.32	**0.000**	(1, 21.68)	30.29	**0.000**
Program item	(1, 21.99)	17.50	**0.000**	(1, 21.40)	17.64	**0.000**
Reproduction system	(1, 26.88)	207.55	**0.000**	(1, 27.07)	183.59	**0.000**
Method × spatial variation	(1, 532.72)	4.76	**0.030**	(1, 532.61)	11.22	**0.001**
Method × program item	(1, 532.72)	0.01	0.902	(1, 532.61)	0.38	0.536
Method × reproduction system	(1, 532.72)	10.09	**0.002**	(1, 532.61)	15.33	**0.000**
Spatial variation × program item	(1, 532.72)	7.79	**0.005**	(1, 532.61)	3.25	0.072
Spatial variation × reproduction system	(1, 532.72)	21.11	**0.000**	(1, 532.61)	9.28	**0.002**
Program item × reproduction system	(1, 532.72)	0.31	0.578	(1, 532.61)	0.22	0.640

Note. Values with $p < 0.05$ are emphasized in bold.
DF = degrees of freedom; DenDF = denominator DF (residual/error); NumDF = numerator DF (effect).

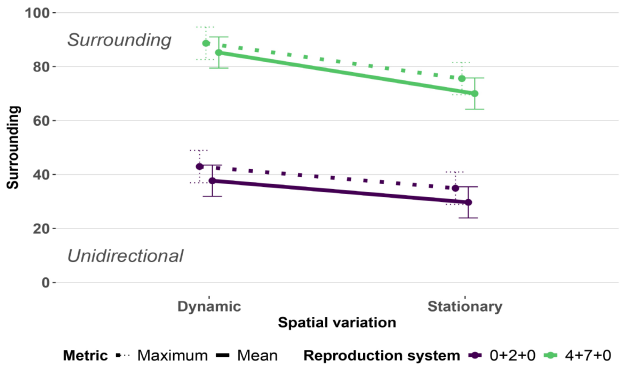


Fig. 9. Interaction effect between *spatial variation* and *reproduction system* for *maximum* and *mean* metrics for the attribute *surrounding*.

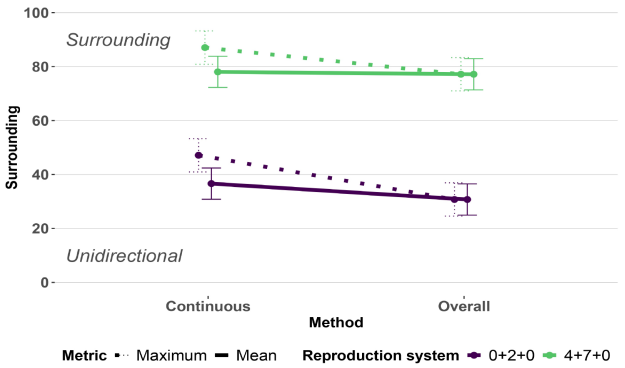


Fig. 10. Interaction effect between *method* and *reproduction system* for *maximum* and *mean* metrics for the attribute *surrounding*.

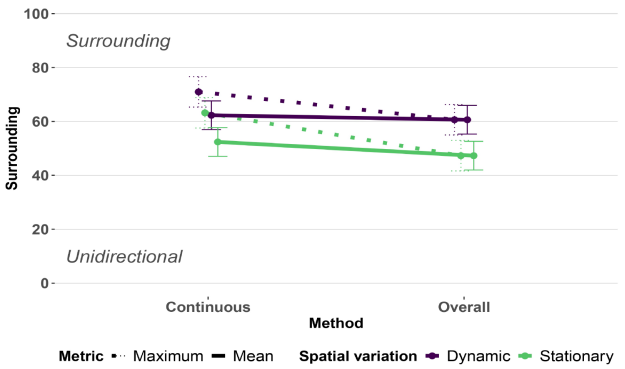


Fig. 11. Interaction effect between *method* and *spatial variation* for *maximum* and *mean* metrics for the attribute *surrounding*.

The ANOVA results (Table 5) show significant main effects of *spatial variation*, *program item*, and *reproduction system* on both metrics, whereas *method* was a significant main effect only for the *maximum* metric. Significant interactions were observed for *method* × *spatial variation*, *method* × *reproduction system*, and *spatial variation* × *reproduction system* on both metrics and for *spatial variation* × *program item* on the *mean* metric.

Post hoc tests showed the strongest interaction ($p = 0.000$) between *spatial variation* and *reproduction system* (Fig. 9). *Dynamic* was consistently rated higher than *stationary*. The surround-with-height setup (*4 + 7 + 0*) was rated more *surrounding* than *stereo* (*0 + 2 + 0*) across both metrics. All comparisons were significant ($p = 0.000$).

For *method* × *reproduction system* (Fig. 10), the continuous method yielded higher scores than the overall method for the *maximum* metric ($p = 0.000$). The continuous method also produced higher scores for the *mean* metric in the *0 + 2 + 0* setup ($p = 0.010$). *Stereo* was consistently rated lowest across all conditions ($p = 0.000$).

The interaction *method* × *spatial variation* (Fig. 11) showed higher ratings for the continuous method. Specifically, the continuous method received significantly higher scores for the *maximum* metric ($p = 0.000$) and for the *mean* metric under *stationary* spatial variation ($p = 0.023$). *Dynamic* spatial variation was rated as more surrounding across both metrics and methods ($p \leq 0.001$).

Finally, a significant interaction for the *mean* metric only was found between *spatial variation* and *program item* (Fig. 12). *Dynamic* was perceived more surrounding *stationary* for both program items ($p = 0.000$) and *Forest* was perceived as more surrounding than *Family*, regardless of the spatial variation condition ($p \leq 0.008$).

3.5 Questionnaire

The questionnaire was mainly used as supplementary information and should therefore be treated carefully and only as an addition to the main results. Part of the questionnaire was to answer two questions regarding the assessors' rating

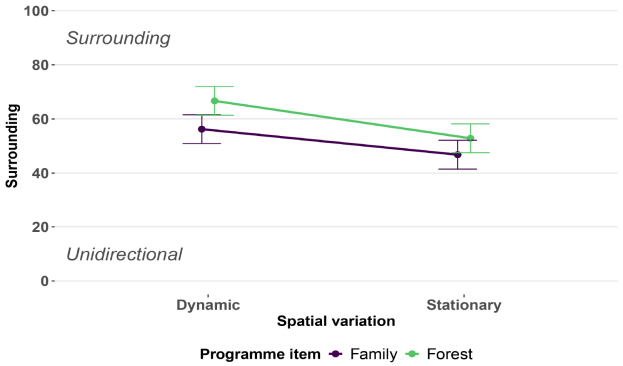


Fig. 12. Interaction effect between *spatial variation* and *program item* for the *mean* metric for the attribute *surrounding*.

experience: "How easy was it for you to understand and use the rating method?" and "How easy was it for you to stay focused on the listening task?" The response options were *very easy* (1), *easy* (2), *moderate* (3), *difficult* (4), and *very difficult* (5). A Wilcoxon signed-rank test revealed no statistically significant differences between the overall and continuous methods in terms of perceived understanding or focus.

Additional comments suggested that some assessors found it difficult to evaluate *BAQ* due to changing stimuli and uncertainty about how to weigh changing spatial aspects, although continuous rating generally seemed to help track such changes more effectively. Furthermore, a lack of reference, the test length, and fatigue were mentioned as concerns that may have influenced scoring.

4 DISCUSSION

The presented study examined how the method of rating (continuous or overall) influences the evaluation of spatial audio attributes alongside *reproduction system*, *spatial variation*, and *program item*. Although the latter factors were the dominant influences, the results demonstrate that the choice of method nevertheless introduces significant differences in listener responses.

4.1 Primary Influences on Ratings

The main influences on perceived *BAQ* and *surrounding* were the physical and content-related factors: *reproduction system*, *spatial variation*, and *program item*. The reproduction system with greater spatial resolution (*4+7+0*) consistently showed higher ratings for both *BAQ* and *surrounding*, confirming findings from the literature [1, 33] showing higher-quality related ratings for reproduction systems with more loudspeakers.

Excerpts with moving sources or dynamic spatial patterns had higher ratings compared with stationary excerpts for both *surrounding* and *BAQ* ratings. These findings extend earlier observations by Porysek Moreta et al. [1], who reported only slight differences in attribute selection between stationary and dynamic content. Although their results suggested that motion could play a role in spatial perception, the present study provides clear evidence that it affects ratings of the selected attributes. However, the magnitude of this effect varied depending on the program material.

Both program items, *Forest* and *Family*, were designed to include spatial variation. However, the stronger effect of spatial variation observed for the *Forest* scene suggests that its spatial cues benefited more from dynamic movement, resulting in higher ratings. This may be due to more pronounced spatial contrasts or a wider spatial distribution of sounds in *Forest*, making dynamic changes more perceptually noticeable than in the speech-based, room-bound *Family* scene. Differences may also relate to expectation bias: the room setting of *Family* might be easier to compare with an internal reference, so inconsistencies in spatialization might be judged more critically than in the more open and naturalistic *Forest* environment.

4.2 Impact of Rating Method

Across both attributes, *method* affected the results. For *BAQ*, effects were observed only for the *maximum* metric (main effect and interaction with *spatial variation*), whereas for *surrounding*, effects appeared for both the *maximum* metric (main effect and interactions with *spatial variation* and *reproduction system*) and the *mean* metric (interactions with *spatial variation* and *reproduction system*). This highlights that method choice matters when evaluating reproduction systems with spatially dynamic content because overall scores may overlook subtle differences in attribute perception.

Analyzing the results by metric highlights how the two methods differ. When summarizing the continuous ratings using the *maximum* metric, higher scores than the overall ratings are expected: the maximum value captures the strongest perceptual moment within an excerpt and is by definition always equal to or greater than the mean perception across the excerpt.

For *surrounding*, a similar effect was observed for the *mean* metric under specific conditions: for stationary excerpts (within *method* × *spatial variation*) and for the stereo setup (within *method* × *reproduction system*). These cases occur at the lowest levels of spatialization (stereo reproduction and stationary scenes). This suggests that continuous evaluation increases sensitivity to brief peak moments of surroundness, which in turn makes these short fluctuations more likely to appear in the continuous rating signal and thus raise the resulting mean score.

An explanation for this is that in limited spatial reproduction (*0+2+0*) assessors may have been more affected by short fluctuations when rating continuously, resulting in higher scores. In contrast, with the *4+7+0* setup, the experience may have been enveloping enough to minimize audible differences.

5 CONCLUSION

The study presented in this work investigated the impact of different methods for audio quality assessment (continu-

ous and overall) on the evaluation of the attributes *BAQ* and *surrounding* across two reproduction systems with spatially dynamic audio content. The research question formulated at the start was: Do ratings of audio quality attributes for reproduction systems with spatially dynamic content differ when made using continuous evaluation versus overall rating methods?

Ratings differ between continuous and overall evaluation methods, with the extent of these differences largely depending on how continuous evaluation data are summarized (e.g., *mean* versus *maximum*) and on the specific attribute assessed. When continuous ratings are summarized using the *maximum* value, they tend to produce higher scores than overall evaluations. This suggests that there is a benefit to using continuous methodologies because some information is lost when listeners are required to aggregate their perception over the full length of a spatially dynamic stimulus.

For the attribute *surrounding*, differences also emerged in *mean*-based summaries under certain conditions, particularly in less spatial scenarios. In contrast, for *BAQ*, *mean* summaries yielded consistent ratings across both methods. Notably, assessor self-reported engagement and perceived task workload did not differ between evaluation approaches. These findings suggest that continuous evaluation methods may be particularly useful when assessing attributes that vary noticeably over time, such as *surrounding*, or when peak experiences are of special interest.

5.1 Limitations

Despite careful study design, several limitations must be acknowledged. One potential limitation is the risk of listener fatigue or bias due to repeated exposure to the same stimuli because the same material was used for both sessions of the experiment. Because assessors listened to the same program items multiple times, prior experience from the first session might have influenced their judgments in the second session. Another difficulty was that the way the task was approached may have been shaped by previous experience because assessors could have carried over expectations or evaluation strategies from the initial session into the subsequent one. Both factors were mitigated by counterbalancing the starting session across assessors (continuous method first versus overall method first), which helped distribute potential bias equally. Preliminary inspection suggested that main trends were consistent between starting groups, although some effects differed in magnitude or significance between groups, indicating minor task-order variability that could be examined in future work.

Additionally, the study lacked an absolute reference for the attributes used in the comparison (and for slider use). This limitation was addressed by providing time for stimulus familiarization, where assessors were exposed to the actual material used in the tasks. The aim was to help establish an internal reference, enabling assessors to become familiar with the possible extremes of the attributes they would encounter in the main experiment and thus use the slider more effectively.

A final limitation concerns the gender distribution of the assessor group. Although balanced representation is desirable and gender-related differences have been observed in some auditory domains (e.g., loudness sensitivity or localization), evidence for effects in quality or attribute-based evaluations is limited. Potential influences cannot be entirely excluded, but the available sample does not permit drawing reliable conclusions.

5.2 Future Work

Future work could explore several aspects that were only partially addressed in this study. One area of interest could be the investigation of additional attributes because the results have shown that differences and potentially advantages of using one method over another depend on the specific attributes being considered. A major influence in this study was the program item, so testing with a larger variety of content would likely be beneficial to see if similar or different results emerge.

Another important consideration is that this study focused on a small selection of metrics to summarize the continuous data, and exploring additional metrics could offer valuable insights. Additionally, the metrics used consolidate a lot of information into a single value, which may limit the potential to uncover certain details. For example, analyzing time-varying trajectories of the response curves could reveal interesting correlations with physical properties of the stimuli (e.g., movement patterns of audio objects), which future work will examine in more detail.

6 ACKNOWLEDGMENT

Special thanks go to the assessors for their time and contributions, to Pete Websdell for help building the slider, and to Thomas Ottesen for support with the experimental setup. This work was funded by the European Union's Horizon 2020 program (Grant No. 956369).

6 REFERENCES

[1] 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.

[2] 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 Conv. Audio Engineering Society* (2022 May), paper 10594. https://aes.org/publications/elibrary-page/?id=21707.

[3] ITU-R, "Methodologies for the Subjective Assessment of the Quality of Television Images," *Recommendation ITU-R BT.500-15* (2023 May).

[4] ITU-T, "Methods for Subjective Determination of Transmission Quality," *Recommendation ITU-T P.800* (1996 Aug.).

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

[6] ITU-R, "Method for the Subjective Assessment of Intermediate Quality Level of Audio Systems," *Recommendation ITU-R BS.1534-3* (2015 Oct.).

[7] F. Rumsey, "Subjective Assessment of the Spatial Attributes of Reproduced Sound," in Proc. *15th AES Intl. Conf. Audio, Acoustics, & Small Spaces* (1998 Oct.), paper 15-012. https://aes.org/publications/elibrary-page/?id=8096.

[8] S. Zielinski, F. Rumsey, and S. Bech, "On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," *J. Audio Eng. Soc.*, vol. 56, no. 6, pp. 427–451 (2008 Jun.). https://aes.org/publications/elibrary-page/?id=14393.

[9] N. Cowan, "On Short and Long Auditory Stores," *Psychol. Bull.*, vol. 96, no. 2, pp. 341–370 (1984 Sep.). https://doi.org/10.1037/0033-2909.96.2.341.

[10] D. Oberfeld, J. Hots, and J. L. Verhey, "Temporal Weights in the Perception of Sound Intensity: Effects of Sound Duration and Number of Temporal Segments," *J. Acoust. Soc. Am.*, vol. 143, no. 2, pp. 943–953 (2018 Feb.). https://doi.org/10.1121/1.5023686.

[11] V. Seferidis, M. Ghanbari, and D. E. Pearson, "Forgiveness Effect in Subjective Assessment of Packet Video," *Electron. Lett.*, vol. 28, no. 21, pp. 2013–2014 (1992 Oct.). https://doi.org/10.1049/el:19921290.

[12] D. S. Hands and S. E. Avons, "Recency and Duration Neglect in Subjective Assessment of Television Picture Quality," *Appl. Cogn. Psychol.*, vol. 15, no. 6, pp. 639–657 (2001 Nov.). https://doi.org/10.1002/acp.731.

[13] E. F. Clarke, *Ways of Listening: An Ecological Approach to the Perception of Musical Meaning* (Oxford University Press, New York, USA, 2005). https://doi.org/10.1093/acprof:oso/9780195151947.001.0001.

[14] N. Schinkel-Bielefeld, "Audio Quality Evaluation in MUSHRA Tests—Influences Between Loop Setting and a Listeners' Ratings," presented *142nd Conv. Audio Engineering Society* (2017 May), paper 9779. https://aes.org/publications/elibrary-page/?id=18655.

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

[16] 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.

[17] 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 In.*, vol. 11, pp. 277–287 (2017 Apr.). https://doi.org/10.1007/s12193-017-0245-z.

[18] M. Hansen and B. Kollmeier, "Continuous Assessment of Time-Varying Speech Quality," *J. Acoust. Soc. Am.*, vol. 106, no. 5, pp. 2888–2899 (1999 Nov.). https://doi.org/10.1121/1.428136.

[19] L. Gros and N. Chateau, "Instantaneous and Overall Judgements for Time-Varying Speech Quality: Assessments and Relationships," *Acta Acust. United Ac.*, vol. 87, no. 3, pp. 367–377 (2001 May).

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

[21] ITU-T, "Continuous Evaluation of Time-Varying Speech Quality," *Recommendation ITU-T P.880* (2004 May).

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

[23] N. Kaplanis, S. Bech, T. Sakari, 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.

[24] 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 Conv. Audio Engineering Society* (2016 May), paper 255. https://aes.org/publications/elibrary-page/?id=18159.

[25] BBC R&D and IRT, "EAR Production Suite," https://ear-production-suite.ebu.io (accessed Aug. 27, 2025).

[26] ITU-R, "Audio Definition Model Renderer for Advanced Sound Systems," *Recommendation ITU-R BS.2127-0* (2019 Jun.).

[27] 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.

[28] J. Francombe, T. Brookes, R. Mason, and F. Melchior, "Loudness Matching Multichannel Audio Program Material With Listeners and Predictive Models," presented at the *139th Conv. Audio Engineering Society* (2015 Oct.), paper 6464. https://aes.org/publications/elibrary-page/?id=18020.

[29] ISO, "Acoustics – Methods for Calculating Loudness – Part 1: Zwicker Method," *Standard 532-1* (2017 Jun.).

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

[31] 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.

[32] J. Judge and J. Stirling, "Fine Motor Skill Performance in Left- and Right-Handers: Evidence of an Advantage for Left-Handers," *Lat-*

erality, vol. 8, no. 4, pp. 297–306 (2003 Nov.). https://doi.org/10.1080/13576500342000022a.

[33] 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.

THE AUTHORS


Pia N. 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 Systematic and professor of audio perception at Aalborg University's section AI and Sound. He retired in 2024 as Director of Research at Bang & Olufsen. 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. 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 and 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 and served as a postdoctoral researcher at University of Newcastle in Australia. Currently, he is a professor in information theory and signal processing, head of the section on artificial intelligence and sound, and head of the Center 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 program. He is an associate editor for signal Processing and source coding for the Institute of Electrical and Electronics Engineers (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.