Technische Universität Berlin

Master Thesis

Development of a Headphone Audio Quality Inventory (HAQI)

Jonathan Rietz



audiokommunikation

Master Thesis

Development of a Headphone Audio Quality Inventory (HAQI)

Jonathan Rietz

Submitted by
Jonathan Rietz
Matr. Nr. 401144

For the academic degree
Master of Science (M.Sc.)

Submitted to
Technische Universität Berlin
Fakultät I – Geisteswissenschaften
Institut für Sprache und Kommunikation
Fachgebiet Audiokommunikation
`http://www.ak.tu-berlin.de`

Submission date
August 30th 2026

Supervision
Prof. Dr. Stefan Weinzierl
Prof. Dr.-Ing. Sebastian Möller

Date of the defence
September 29th 2026



1 Zusammenfassung

Die Klangqualität von Kopfhörern wird üblicherweise anhand physikalischer Messungen und globaler Präferenzurteile charakterisiert. Diese Ansätze liefern jedoch keine psychometrisch fundierte, mehrdimensionale Beschreibung der Kopfhörerwahrnehmung. In dieser Arbeit wurde das Headphone Audio Quality Inventory (HAQI) entwickelt, ein Instrument zur Erfassung wahrgenommener Aspekte des Kopfhörerklangs und des Tragekomforts. Zunächst erarbeitete eine Expertenfokusgruppe ein Konsensvokabular mit 68 Attributen, Skalenbezeichnungen und zugehörigen Beschreibungen. Dieses Vokabular wurde anschließend in einem institutionsübergreifenden Hörexperiment eingesetzt, in dem 104 Teilnehmende eine vielfältige Auswahl von 24 Kopfhörermodellen beurteilten. Mithilfe einer explorativen Faktorenanalyse wurden die latenten Dimensionen untersucht, die den Bewertungen der nach der Datenbereinigung verbleibenden 98 Teilnehmenden zugrunde lagen. Auf dieser Grundlage wurden mittels konfirmatorischer Faktorenanalyse hierarchisch verschachtelte Messmodelle mit fünf bzw. sechs Faktoren und 17 bzw. 20 Items konstruiert. Die ersten fünf Dimensionen wurden als Klarheit, Räumlichkeit, Hochton-Unbehagen, Bass und Tragekomfort interpretiert, das Sechs-Faktoren-Modell enthielt zusätzlich die Dimension Offenheit. Das Fünf-Faktoren-Modell wurde hinsichtlich Konstruktreliabilität, Konvergenz- und Diskriminanzvalidität sowie skalarer Messinvarianz über Audioinhalte, Grade akustischer Expertise und Sprachen des Fragebogens hinweg evaluiert. Die HAQI-Dimensionen erklärten 69,5% der Varianz im Gefallen und 76,9% der Varianz im Gesamteindruck des Klangs, wobei Klarheit der stärkste Prädiktor beider Urteile war. Darüber hinaus differenzierten die HAQI-Dimensionen die Wahrnehmungsprofile der untersuchten Kopfhörermodelle und -kategorien. Das HAQI stellt damit ein Instrument zur Charakterisierung der mehrdimensionalen Kopfhörerwahrnehmung bereit und kann physikalische Messungen bei der Bewertung und Entwicklung von Kopfhörern ergänzen. Eine unabhängige Evaluation des finalen Messmodells sowie die Untersuchung seiner zeitlichen Stabilität stehen jedoch noch aus.

Eidesstattliche Erklärung

Hiermit versichere ich, dass ich die vorliegende Arbeit eigenständig ohne Hilfe Dritter und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe. Alle Stellen, die den benutzten Quellen und Hilfsmitteln unverändert oder sinngemäß entnommen sind, habe ich als solche kenntlich gemacht.

Sofern generative KI-Tools verwendet wurden, habe ich Produktnamen, Hersteller, die jeweils verwendete Softwareversion und die jeweiligen Einsatzzwecke (z.B. sprachliche Überprüfung und Verbesserung der Texte, systematische Recherche) benannt. Ich verantworte die Auswahl, die Übernahme und sämtliche Ergebnisse des von mir verwendeten KI-generierten Outputs vollumfänglich selbst.

Die Grundsätze zur Sicherung guter wissenschaftlicher Praxis an der TU Berlin vom 30. Mai 2023 habe ich zur Kenntnis genommen.

Ich erkläre weiterhin, dass ich die Arbeit in gleicher oder ähnlicher Form noch keiner anderen Prüfungsbehörde vorgelegt habe.

Berlin, den 30.08.2026



……

Original-Unterschrift

Development of a Headphone Audio Quality Inventory (HAQI)

Jonathan Rietz

***Abstract*—Headphone sound quality is commonly characterized using physical measurements and global preference ratings, but these approaches do not provide a psychometrically grounded multidimensional description of headphone perception. This study developed the Headphone Audio Quality Inventory (HAQI), an instrument for assessing perceptual aspects of headphone sound and wearing comfort. First, an expert focus group established a consensus vocabulary comprising 68 attributes with scale labels and corresponding descriptions. This vocabulary was subsequently used in a multi-institutional listening experiment in which 104 participants evaluated a diverse set of 24 headphone models. Exploratory factor analysis was applied to identify the latent dimensions underlying the ratings of the 98 participants retained after data cleaning. Based on these results, confirmatory factor analysis was used to construct nested five- and six-factor measurement models comprising 17 and 20 items, respectively. The first five dimensions were interpreted as Clarity, Spaciousness, Treble Discomfort, Bass, and Wearing Comfort, while the six-factor model additionally included Openness. The five-factor model was evaluated in terms of construct reliability, convergent and discriminant validity, and scalar measurement invariance across audio contents, levels of acoustic expertise, and questionnaire languages. The HAQI dimensions explained 69.5% of the variance in liking and 76.9% of the variance in overall acoustic impression, with Clarity emerging as the strongest predictor of both judgments. Moreover, the HAQI dimensions differentiated the perceptual profiles of the headphone models and categories investigated. The HAQI thus provides a tool for characterizing multidimensional headphone perception and may complement physical measurements in headphone evaluation and development. Independent evaluation of the final measurement model and assessment of its temporal stability remain necessary.**

I. Introduction

HEADPHONES have become one of the most widely used audio reproduction systems. With the rise of smartphones and mobile internet in the last two decades, as well as convenient access to music through licensed streaming services, weekly music consumption has increased in recent years [1]. Especially for individual and mobile consumption of audio and audiovisual media, headphones play an important role. In the context of headphones, sound quality is considered a key attribute.

The acoustic behavior of headphones can be characterized using physical metrics, most importantly their frequency response. For loudspeakers in rooms, the acoustic behavior leading to the most preferred sound can be described as a flat anechoic on-axis response combined with smooth off-axis behavior, as well as smooth sound power and directivity [2]. For headphones, these characteristics cannot be directly applied as headphones do not interact with a room but are directly coupled to the ear. A headphone's frequency response, therefore, has to be measured on specific measurement fixtures that replicate the human ear's anatomical and acoustic properties, as described in [3]. The preferred frequency response at the ear-drum reference point (DRP) of such artificial ears, and its relationship to the response at the human eardrum, has been a topic of research for decades. Originally, headphones were calibrated to match the free-field response of a frontal sound source. This was challenged by [4] who proposed a diffuse-field calibration. [5] then, in turn, found that listeners preferred a diffuse-field calibration, but with a significantly reduced peak at 3 kHz. Later, a calibration based on a semi-reflective field (SRF), as found in a typical listening scenario in a room, was investigated in a series of studies by Olive and Welti, who proposed "Harman target curves" for both insert and circum-aural headphones, which were significantly preferred by listeners over earlier calibrations. However, it was discovered that there is variation in preference, with certain groups of people preferring more or less bass and treble [6]. In recent years, many more calibration targets have been proposed, see [2] for a review.

Although frequency response targets are highly informative for predicting tonal preference, they do not provide a multidimensional description of headphone perception. Such a description requires perceptual attributes and measurement scales that capture distinct aspects of the listening experience. Psychological measurement instruments can be used to evaluate the multidimensional aspects of perception in many domains. Usually, these consist of structured lists of attributes and corresponding scales. There are many ways to obtain these attributes, which can be divided into direct and indirect elicitation methods. The direct method is defined by presenting stimuli from the domain under investigation and identifying attributes that describe perception. Attributes can be developed in group discussions resulting in consensus vocabularies (CV), or individuals can elicit their own vocabulary, which must then be further analyzed. Indirect elicitation methods separate the presentation of stimuli from the verbalization of describing attributes [7].

Vocabularies exist for the general evaluation of sound quality; for example, [8], [9] describe guidelines for general sound quality assessment (of program material and/or the playback system) by experts, providing a list of descriptive terms organized into main attributes and sub-attributes. The *Sound Wheel for Reproduced Sound* [10] uses a hierarchical presentation of a lexicon for reproduced sound. A lexicon describes a "standardized vocabulary that objectively describes the sensory properties of consumer products" [10, p. 2]. Unlike

This work involved human participants. A self-assessment found that it complied with relevant laws and institutional guidelines, in accordance with the Ethics Commission of Faculty IV at TU Berlin.

a CV, this system is intended to be applicable beyond the panel or laboratory in which it was developed and to be usable in a broader field and in various scenarios. In the *Sound Wheel for Reproduced Sound*, various attributes are grouped into subcategories, which in turn form higher-level categories within the wheel. Together, they thus form a structured vocabulary that can be used to describe the sound characteristics of playback in various scenarios, such as playback through speakers or headphones.

While these vocabularies organize perceptual descriptors conceptually, they do not necessarily represent empirically established latent constructs. More recent domain-specific inventories about room-acoustic [11], stage-acoustic [12], or public address (PA) system-perception [13] have therefore combined attribute elicitation with psychometric analyzes. In these studies, after the word elicitation process, questionnaires were developed that were then used to investigate perceptual differences in the respective domains. Based on the questionnaire results, latent dimensions of room acoustics, stage acoustics, and PA perception were investigated using exploratory factor analysis (EFA). Measurement instruments for perception, described by different latent factors, underlying items, and respective scales, were constructed and validated using confirmatory factor analysis (CFA).

For headphone sound quality perception specifically, seminal early work was done by [14]. In one of the studies reported, they performed listening experiments with 20 participants, who listened to five mostly classical music programs on eight headphones using different transducer technologies and types of construction. The participants were presented with 30 adjectives describing headphone sound, mostly related to timbre, and had to evaluate the headphones on a 10-point scale. The perceptual structure was investigated using principal component analysis (PCA) on the means of the adjective ratings, resulting in five components interpreted as "Sharpness/Hardness-Softness", "Clearness/Distinctness", "Disturbing Sounds", "Brightness-Darkness", and "Feeling of Space". [15] covers the sensory analysis of sound quality of headphone playback. Two different sensory techniques were used, a consensus vocabulary (CV) and an individual vocabulary (IV). The consensus vocabulary describing headphone sound was made up of the three main categories "Localization", "Space", and "Timbre", and the corresponding attributes. The focus of this study, however, was the evaluation of spatial audio reproduction techniques on headphones, not the evaluation of different headphones themselves. The perceptual characteristics of active noise cancellation (ANC) headphones were investigated in [16]. Two consensus vocabularies for noise as well as music-in-noise were developed, consisting of ten and eight attributes, respectively. Based on a listening experiment with recordings of six headphones played back over a single model in a double-blind listening test, and analysis of the responses using hierarchical multiple factor analysis (HMFA), a four-dimensional solution for music reproduction was found. The four dimensions were described as "Timbre/Attenuation", "Dynamic & Spatial Attributes", "Precision/Stereo Space", and "Treble Attenuation". [17] sought to identify the dominating perceptual dimensions to differentiate headphones. For this, 21 different headphones playing two very short musical excerpts were recorded using a head-and-torso simulator (HATS) and reproduced on a single headphone in two listening experiments. The participants were not presented with attributes to evaluate headphone sound, instead, they rated the dissimilarity between different headphones. Multidimensional scaling (MDS) was then used to identify the underlying perceptual dimensions, suggesting a two-dimensional solution, which was interpreted in terms of the relative strength of bass and relative bass-midrange loudness.

As headphones are always worn in direct contact with the listener's head and (most of the time) ears, headphone usage is a multisensory experience. Despite the inherently multisensory nature of headphone use, potential interactions between auditory perception and wearing comfort have received little systematic attention. [18] mainly investigated the relationship between perceived tonal balance and listener preference, but comfort was also considered. However, no further analysis of the effect of comfort on acoustic preference took place; notably, the headphone considered to be best-sounding had the lowest comfort ratings. [19] analyzed the effects of different characteristics of a headphone, such as comfort, build quality, design, or sound quality, on user experience. Factor analysis was used to identify the dimensions of headphone user experience. It was found that attributes from the factor associated with "Comfort" had some of the highest correlations to the overall evaluation of the headphones.

In past research, there have been shortcomings in describing headphone sound quality in a psychometrically valid way. Research on target curves, such as those investigated by [6], only considered overall preference, which is deemed "neutral", but did not investigate the effects of different perceptual dimensions of headphone sound on preference. General vocabularies and dictionaries on sound quality, like [10], describe many perceptual aspects of (headphone) sound reproduction using attributes that are structurally organized, but there is no deeper research into whether these attributes form reliable and distinguishable latent constructs. While other psychometrically validated questionnaires on sound quality, such as [13], exist, they cannot be applied to the different domain of headphone sound reproduction.

For the works specifically aimed at analyzing the perceptual dimensions of headphone sound reproduction, there are multiple limitations to their applicability. The selection of headphone models in listening experiments is a limiting factor: [14] and [16] only used eight and six different models, respectively, but many headphones are needed to find all perceptual dimensions, as discussed in [17]. Moreover, in order to be applicable to all kinds of headphones available nowadays, many different types of headphones need to be investigated. While [14] already used headphones with different transducer and construction designs, these are hardly representative of headphones available today. [16] and [17] only used circumaural and supra-aural electrodynamic headphones, not covering relevant types such as insert or open-ear headphones or other transducer technologies, potentially resulting in the headphones not providing enough variance in all perceptual

dimensions. It is also unclear whether the attributes presented to evaluate headphones were sufficient to cover all perceptual dimensions of headphone sound. While [14] presented an extensive list of 30 adjectives, the vast majority were related to timbre, with no attributes about dynamics, and only one about the perception of space. [16] presented only eight attributes regarding music perception altogether. In neither of the studies was the interaction between comfort or other headphone-related characteristics and perceived sound quality investigated. Most importantly, in none of the headphone studies was the psychometric validity of the discovered structure investigated, namely convergent and discriminant validity, construct reliability, and measurement invariance. The statistical tools used in earlier research, namely PCA and the PCA-based HMFA, do not constitute latent-variable measurement models and do not explicitly separate common variance from item-specific variance and measurement error. Consequently, these analyzes alone do not permit a comprehensive psychometric evaluation of the resulting structure in terms of construct reliability, convergent and discriminant validity, and measurement invariance [20].

The goal of this study is the development of the *Headphone Audio Quality Inventory* (HAQI), which, to the author's knowledge, provides the first psychometrically grounded measurement instrument specifically for headphone sound reproduction, addressing the issues mentioned in earlier studies. This is done through an extensive process of creating a comprehensive preliminary list of potentially relevant perceptual characteristics, conducting a listening experiment with a large and diverse set of headphones, analyzing the latent dimensions of headphone perception, and constructing a measurement instrument evaluated in terms of validity, reliability, and measurement invariance.

II. Methods

The development of the *Headphone Audio Quality Inventory* (HAQI) followed a multistage procedure combining qualitative item development, perceptual evaluation, and psychometric modeling. An expert focus group [21] was initially conducted to establish a preliminary CV for describing headphone sound perception. This vocabulary subsequently formed the basis of a listening experiment in which different headphone models were evaluated. EFA was applied to the resulting ratings to identify the latent dimensions underlying headphone sound perception. These findings informed the specification of a measurement model, which was subsequently examined using CFA. A structural equation model (SEM) was then used to investigate the relationships between the latent HAQI dimensions and global evaluative judgments. Finally, a linear mixed-effects model (LMM) assessed whether the resulting factor scores differentiated systematically between headphone models and thus captured perceptually meaningful differences among them.

A. Expert Focus Group

As mentioned earlier, sensory vocabularies are often developed using a direct elicitation method, in which stimuli are presented during attribute elicitation and the resulting terms are discussed within a group. However, this procedure can be time-consuming [7]. Therefore, an expert focus group approach was employed [21]. Focus groups are a qualitative research method in which purposively selected participants discuss a defined topic through moderated interaction. Expert focus groups elicit domain-specific knowledge in a setting resembling experts' routine practice of developing and defending positions in committees and working groups. This interaction can provide a broader and more comprehensive account of expert knowledge than individual interviews [22]. In the present study, it facilitated the elicitation of attributes related to headphone sound and comfort perception without requiring the presentation of stimuli.

As a starting point for a discussion in the expert focus group, attributes and descriptions of sound-describing characteristics were collected from scientific literature [8]–[10], [15], [23], [24] as well as audio enthusiast forums [25]–[27]. All attributes and descriptions were put in a spreadsheet and further organized: Attributes that did not apply to headphone playback, such as those that describe room acoustics, were removed, as well as duplicates of nearly identical attributes, and attributes were grouped into categories such as timbre, transparency, or spaciousness. This resulted in 127 items[1].

German-speaking experts in headphone sound reproduction (PhD candidates, professors, researchers, acoustic engineers, and audio engineers) were invited to the focus group. The discussion was held in December 2025 in Berlin in a hybrid setting, with nine experts (aged 26 to 60 yrs, 7 male, 2 female) taking part. The focus group lasted about 4 hours including one break.

The vocabulary to be developed was aimed at being a list of auditory, haptic, as well as other perceptual qualities in the context of headphone sound reproduction. The intended use case was an effortless and comparable quality assessment of headphones, in the context of research, industry, or personal use. Headphones were defined in a broad sense including, but not limited to, conventional designs like circum-aural, supra-aural, or intra-concha and special designs like extra-aural, bone-conduction, or open-ear. Passive headphones as well as active headphones with digital signal processing (DSP), such as active noise cancellation (ANC), or wireless connectivity, were considered. The purpose of the focus group was formulated in a mission statement, which was *to create a German-language vocabulary for the perceptual, auditory, and haptic qualities of headphones, including related signal processing such as active noise cancellation.* The experts were instructed that the attributes should relate to the perceptual domain, not to physical measures. Furthermore, the vocabulary should be complete, but also be practically relevant at the same time.

The session was moderated by one of the professors and assisted by the author. The initial attribute list was displayed on a screen and was systematically reviewed item-by-item. For each attribute, the experts discussed its relevance in describing headphone (audio) quality, its perceptual interpretability, and its potential usefulness for subsequent listener evaluations.

[1]Digital Appendix, *Item_selection_from_literature.xlsx*

Attributes considered overly vague, redundant, or insufficiently discriminative were removed. In cases where a single attribute was found to encompass multiple perceptual aspects, it was subdivided into more specific descriptors. The panel was also encouraged to propose additional attributes that were not identified during the literature review. Throughout the discussion, attribute labels were refined where necessary to improve conceptual clarity and reduce ambiguity. Furthermore, opposing poles were defined for each retained item. Decisions regarding the inclusion, exclusion, or modification of attributes were made by majority vote among the participating experts. The resulting set of German items served as the basis for the questionnaire used in the subsequent listening experiment.

After the focus group session, German descriptions for all items were formulated. This preliminary HAQI vocabulary was then translated into English with the help of machine-learning-based translation tools and then went through a revision process by a bilingual and a native English-speaking expert, taking one hour.

B. Listening Experiment

1) Stimuli: The goal of the HAQI is to be a measurement instrument of audio and haptic quality for all kinds of headphones. Therefore, for the inventory's validation, a diverse collection of headphones has to be used to provide enough perceptual variance. For the listening experiment, 24 different models of headphones were selected. These ranged from very open constructions to closed designs, from insert earphones to extra-aural headphones. These models included historic as well as recent consumer, professional audio, and hi-fi products, ranging from just €5 to well over €1000 in price.[2] In addition to headphone selection, the audio content used in the listening experiment was important to cover all perceptual dimensions. Three contents were selected:

- Orchestral music (A. Bruckner: *Symphony No. 9*, 3rd movement) (1:09 min);
- Progressive rock music (Polyphia: *Playing God*) (1:01 min);
- Vocal-Blues (Dominique Fils-Aimé: *Birds*) (1:05 min)

All three selections are high-fidelity recordings with distinct acoustic profiles. The orchestral track exhibits a wide dynamic range, focusing on the timbre of strings, woodwinds, and brass. The progressive rock track features high-density transients from plucked strings and drums, combined with an electronic bass drum. Finally, the vocal-blues content provides a natural recording of a solo voice and an acoustic double bass.

Combining the 24 headphone models with the three audio contents, we have 72 different acoustical conditions in total. Additionally, three headphones that support ANC were tested with ANC turned on and off.

2) Procedure: To gain access to enough subjects and be able to investigate many different headphone types, the main research team decided to conduct the listening experiment in a distributed fashion, conducting it at nine different institutions, of which seven are in Germany, one in Sweden, and one in Finland. Each institution conducted the experiment with a subset of 7 out of the 24 headphones.[3] The distribution and presentation order of the headphones across institutions are provided in Appendix A.

The listening experiment was implemented as an online questionnaire using the survey platform LimeSurvey [28]. It consisted of a familiarization phase followed by the main evaluation task and a final demographic questionnaire. During the familiarization phase, participants listened to a 30 s music excerpt (Dream Lovers: *Brasil*) presented in a continuous loop while successively trying all seven headphones included in the experiment. For each headphone, participants were allowed to adjust the playback level to a personally comfortable listening level and were encouraged to listen until they had formed an initial impression before proceeding to the next device.

In the main evaluation, participants assessed each of the seven headphones individually, presented in a fixed order that depended on the participating institution. For every participant, one of the three audio excerpts described above was assigned according to a rotating schedule to ensure a balanced incomplete block design. For each headphone, the audio excerpt was presented continuously in a loop throughout the evaluation. Participants rated the perceived characteristics of the headphone using the preliminary HAQI vocabulary consisting of 68 items, which was subdivided into three sections *Sound Impression*, *Comfort and Suitability for Everyday Use* and *Overall Impression*. Within each section, item order was randomized separately for each participant. A short explanation of each item was available by hovering over the corresponding item label. Ratings were collected on continuous visual analog scales without numerical labels. Slider positions were internally recorded on a continuous scale ranging from 0 to 100, although no numerical values were displayed to participants. Participants were instructed to read the explanation of each item before providing their rating and to respond as intuitively as possible. They were informed that there were no correct or incorrect answers and encouraged to skip items they considered impossible to judge. Some items required active interaction with the headphones. For example, the evaluation of self-speech perception required participants to speak while wearing the headphones. Instructions for these procedures were provided both verbally by the experimenter and within the corresponding item explanations. After completing the ratings for one headphone, the next headphone was presented until all seven headphones had been evaluated. Participants were offered a five-minute break after completing the evaluation of the first four headphones.

Finally, participants completed a demographic questionnaire covering gender, age, educational background, language proficiency, and educational or professional experience in music and acoustics.

C. Statistical Analysis

1) Data Preprocessing: All raw data from *LimeSurvey* from July 2026 was exported into an SPSS-format file. In the first step, only completed questionnaires were considered for

[2]Digital Appendix, *Headphone_pool.xlsx*

[3]The only exception is one institution with a subset of eight headphones.

further analysis. For every item, the mean, median, standard deviation (SD), skewness, kurtosis, and missing values were computed, and responses were investigated in histograms to check for unusual response patterns. For every participant, the mean, SD, range, skewness, kurtosis, and number of missing ratings were calculated. Furthermore, to compare each participant's responses with those of others at the same institution, the mean absolute Z-Score, maximum absolute Z-score, the number of rating values outside of the interquartile range (IQR) were computed, as well as the median Pearson correlation between the ratings of a participant and other participants of the same institution, and the Pearson correlation between the institution's mean ratings and a participant in a leave-one-out fashion. The comparison with other participants at the same institution is particularly relevant since they were using the same headphones, and ratings can therefore be more easily compared. If multiple metrics were abnormal for participants, histograms of their response patterns as well as ratings of multiple low-level items were investigated to determine if they should be excluded from further analysis.

2) Exploratory Factor Analysis: To estimate the latent factors of headphone audio perception and their corresponding items, an EFA using the `psych` [29] package inside the statistical software *Jamovi* [30] was performed. The EFA drew on the information from the correlation matrix of all item ratings, except for those describing more global judgments, i.e., *liking*, *overall acoustic impression*, *naturalness*, *transparency of the playback system* and *perceived value*. The factor extraction method of minimum residuals (MINRES) combined with the oblique Oblimin rotation was used to account for the non-normality of the data and allow for correlated factors. The sampling adequacy was investigated with the Kaiser-Meyer-Olkin (KMO) criterion. It is a measure to check the suitability of the data for factor analysis by analyzing the correlation structure of the variables. Generally, values below 0.5 are not acceptable, and values above 0.9 are considered excellent [31]. To estimate the optimal number of latent factors, the scree test (keeping all factors before a significant drop in eigenvalue), the Kaiser-criterion (using all factors with eigenvalues greater than one), and parallel analysis (comparing eigenvalues of the empirical data with those of randomly-generated data) were used. Since the first two criteria suggested a 4-factor and a 5-factor solution, and the latter suggested a 7-factor solution, the following steps were performed for models with 4, 5, 6, and 7 factors.

3) Confirmatory Factor Analysis: Based on the factor structure identified in the EFA, an initial confirmatory measurement model was specified in which each item was assigned to a single factor ("simple structure"). Before fitting the CFA, items with a primary EFA loading below $.40$ were excluded. Items were also considered insufficiently distinct and therefore excluded when the difference between their two largest loadings was below $.10$.

The measurement model was estimated in `lavaan` [32] using robust maximum likelihood estimation (MLR), which is robust to violations of multivariate normality. Missing item values were handled using full information maximum likelihood (FIML), which uses all available observations under a missing-at-random assumption. Because each participant evaluated multiple headphones, participant ID was specified as the clustering variable. Consequently, standard errors and the model test statistic were adjusted for the non-independence of ratings obtained from the same participant.

Following estimation of the initial model, modification indices, standardized factor loadings, and item content were examined to identify sources of model misfit and items were removed accordingly. At least three indicators were retained for each factor. Because the model was further adjusted on basis of these results, the CFA loses its confirmatory character and should therefore be validated on an independent sample [20].

Model fit was evaluated using the scaled χ^2 test and several fit indices. The Comparative Fit Index (CFI) and Tucker-Lewis Index (TLI) quantify the improvement in fit relative to an independence model in which the covariances among the observed variables are constrained to zero. The Root Mean Square Error of Approximation (RMSEA) assesses approximate model misfit while accounting for model complexity, whereas the Standardized Root Mean Square Residual (SRMR) summarizes the discrepancy between the observed and model-implied standardized covariance matrices. As the χ^2 test is sensitive to sample size, model evaluation was based on the overall pattern of results rather than on the test alone. Following Hu and Bentler [33], CFI and TLI values around or above $.95$, RMSEA values around or below $.06$, and SRMR values around or below $.08$ were used as guidelines for good fit. These thresholds were treated as heuristic reference values rather than strict decision rules.

For each factor, McDonald's ω was calculated as a model-based estimate of factor reliability [34]. Values of approximately $.70$ or higher were interpreted as indicating acceptable reliability. Convergent validity was assessed using the Average Variance Extracted (AVE), which represents the average proportion of item variance accounted for by the corresponding latent factor. An AVE of at least $.50$ was used as a conventional reference value [35]. Discriminant validity was examined using the heterotrait-monotrait ratio of correlations (HTMT). Pairwise HTMT values below $.85$ were interpreted as evidence that conceptually distinct factors were sufficiently differentiated [36].

Measurement invariance across musical content was examined using multiple-group CFA, with audio content as the grouping variable. Configural invariance required the same factor structure and item assignments across the three content conditions. Metric invariance additionally constrained the factor loadings to equality, while scalar invariance also imposed equality constraints on the item intercepts. The increasingly constrained models were compared using changes in CFI and RMSEA. A decrease in CFI exceeding $.01$, supplemented by an increase in RMSEA exceeding $.015$, was considered indicative of non-invariance [37], [38]. Measurement invariance across acoustic expertise and questionnaire language was examined analogously. For the former, participants were grouped according to whether they had at least two years of professional or educational experience in acoustics; for the latter, they were grouped according to whether they completed

the German- or English-language version of the questionnaire.

4) Validity Analysis: To investigate the predictive validity of the HAQI, the previously specified CFA measurement model was extended into a structural equation model (SEM) using `lavaan`. Multiple global judgments, namely *liking*, *overall acoustic impression*, *transparency of the playback system*, *naturalness* and *perceived value* were predicted as correlated dependent variables by the latent dimensions of the HAQI. This model estimates the relationships between the dimensions and global judgments while considering the measurement errors of the underlying items. The relationships are described by the standardized regression coefficient β, the explained variance R^2 describes to which extent the HAQI-dimensions explain each global judgment, such as *liking*.

To investigate differences in the perceived sensory profiles of the headphones, a linear mixed-effects model (LMM) was fitted using the `GAMLj` module [39] in *jamovi*. The HAQI factor scores served as the dependent variable. Headphone model, HAQI dimension, and institution were included as categorical fixed effects, while the presentation position of each headphone was included as a continuous covariate. The model further included interactions between HAQI dimension and headphone model, between HAQI dimension and presentation position, and between HAQI dimension and institution. These interactions were used to examine whether the headphones exhibited different dimension-specific quality profiles, whether potential linear position effects varied across HAQI dimensions, and whether dimension-specific ratings differed between institutions, respectively. Because each participant rated multiple headphones, participant ID was included as a random intercept to account for the non-independence of repeated observations. To ensure a consistent semantic direction across dimensions and permit the interpretation of effects averaged across dimensions, the negatively oriented factor *Treble Discomfort* was inverted by multiplying its factor scores by -1, resulting in the factor *Treble Comfort*. Fixed effects were evaluated using F-tests with Satterthwaite-approximated degrees of freedom. Pairwise comparisons of the estimated marginal means (EMMs) for the interaction between HAQI dimension and headphone model were conducted, with the resulting p values adjusted for multiple testing using the Holm procedure. In an extension of the original model, the dimension-specific factor score of the immediately preceding headphone was included as a lagged contextual predictor, together with its interaction with HAQI dimension, to investigate possible carryover effects previously observed in sensory evaluation experiments [40]. Observations at the first presentation position were excluded from this analysis because no preceding headphone was available. Because headphone order was fixed within institutions, preceding headphone, institution, and current headphone condition were not fully separable. Consequently, the model was not regarded as providing a definitive correction for carryover effects, but rather as an exploratory analysis of potential sequential effects. To analyze perceptual differences between different types of headphones, e.g. circum-aural closed back or supra-aural open back headphones, another LMM was fitted with an identical configuration, but this time grouping headphones of the same type together.

III. Results

A. Expert Focus Group

The vocabulary resulting from the expert focus group consists of 68 items structured into qualities related to timbre, spaciousness, clarity, dynamics, distortions, haptics, sound isolation, artifacts, and overall impression. Each item is accompanied by bipolar scale labels and a short description. The attribute set comprises both lower-level perceptual characteristics related to the physical properties of the reproduced sound, e.g., *treble emphasis*, *sub bass*, or *distortion*, and higher-level constructs that summarize broader psychological aspects of auditory perception and the overall listening experience, such as *relaxedness*, *source imaging*, or *resolution*.[4]

While most of the resulting items can be found in the existing literature, items relating to haptics (*weight distribution*), isolation (*perception of own speech*), playback-unrelated artifacts (*cable noise*), and overall impression (*liking*) were newly suggested by the expert focus group and included in the vocabulary.[5]

B. Listening Experiment: Descriptive Statistics

Overall, 104 people participated in the study across 8 different institutions in July 2026[6]. On average, the participants needed a median of $80\,\mathrm{min}$ (interquartile range (IQR): $65\,\mathrm{min}$ to $94\,\mathrm{min}$) to complete the experiment. $75\,\%$ of the participants were male, and $23\,\%$ were female. The median age was 28 years (IQR: 25-31 years). $74\,\%$ held an undergraduate degree or higher. $77\,\%$ completed the questionnaire in German, with the remaining $23\,\%$ using English. $72\,\%$ reported having an acoustic-related profession, training, or educational background. $32\,\%$ had a music-related profession, training, or educational background, $70\,\%$ answered that they learned an instrument or regularly perform music in a group. Participants reported spending, on average, 2.5 hours a day actively listening to music or sound. Moreover, $64\,\%$ had previously taken part in scientific listening experiments.

The responses of six participants were excluded from further analysis because they exhibited unusual response patterns, including very high or very low standard deviations, very low correlations with those of other participants at the same institution, or a disproportionately large concentration of responses at the scale minimum, midpoint, or maximum.[7]

C. Factor Analysis

1) Exploratory Factor Analysis: Parallel analysis suggested a 7-factor solution, the Kaiser-Guttmann criterion suggested a 5-factor solution and the scree test a four-factor solution

[4]For differentiation, HAQI items are written in *lowercase italics*, while the factors are written in *Uppercase Italics*.

[5]Digital Appendix, *Consensus_vocabulary.xlsx*

[6]No data was available from the 9th institution during this period, more data is still being collected at the point of this submission, which will be analyzed in the future.

[7]Further details are provided in the Digital Appendix, *Data_cleaning_candidates.xlsx*.

as seen in Fig. 1. The different solutions suggested similar factors that could be interpreted as *Clarity*, *Treble Discomfort*, *Spaciousness*, *Wearing Comfort*, *Bass*, *Openness* and *Wearing Perception* for the 7-factor suggestion, the 5-factor suggestion by Kaiser criterion missing *Openness* and *Spaciousness* compared to the 7-factor suggestion, and the 4-factor suggestion resulting from the scree test proposing the factors interpretable as *Clarity*, *Treble Discomfort*, *Bass + Isolation* and *Wearing Comfort*. Therefore confirmatory factor analyses were performed for all three variants, as well as a 6-factor variant. The overall KMO reports excellent adequacy (0.956) of the samples, with no single variable being below 0.7. The EFA results of the 7-factor solution are provided in Appendix B.

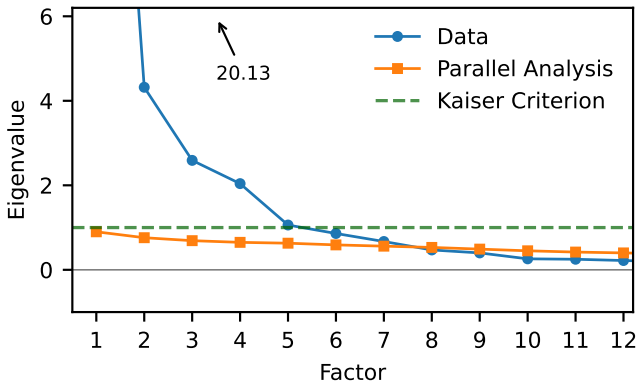


Fig. 1. Scree plot of the initial EFA results. Parallel analysis proposes seven factors, the Kaiser criterion five factors, the Scree test four factors.

2) Confirmatory Factor Analysis: Fit statistics of all evaluated CFA solutions are found in Table I. Employing a 7-factor CFA with the factors mentioned above and using the procedure described in Section II-C3, the resulting measurement model slightly missed a good fit, with AVE of the *Wear Perception* factor being only 0.396, suggesting that this factor does not describe the underlying items well.[8] Since only the three selected items *clamp force*, *isolation* and *weight* loaded substantially on this factor in the initial EFA, no further improvement was possible and the 7-factor solution including *Wear Perception* was therefore rejected. A 6-factor solution without *Wear Perception* resulted in an excellent fit with convergent validity described by AVE between 0.565 and 0.754 for the factors. However, for *Openness*, factor reliability measured by McDonald's ω was not acceptable with $\omega = .362$ suggesting that the underlying items *airiness*, *spatial openness* and *veiledness* do not describe this factor well. For the other factors however, ω was between 0.810 and 0.926. The discriminant validity described by the heterotrait-monotrait (HTMT) ratios of correlations was sufficient between all factors, with .845 between *Clarity* and *Spaciousness* being the highest. A 5-factor solution also removing *Openness* also resulted in an excellent fit, this time reporting the best factor reliability and convergent validity with $\omega \geq 0.81$ and AVE≥ 0.59 for all factors, as well as the best discriminant validity with the highest HTMT ratio being .836 between *Clarity* and *Spaciousness* again, see Table II. A 4-factor solution without *Spaciousness* resulted in slightly reduced model fit, therefore the 5-factor solution is considered as the best-fitting solution. The alternative 5-factor solution proposed by the Kaiser-criterion, drawing on *Wear Perception* as a factor instead of *Spaciousness*, reported inadequate fit and was rejected. Any further analysis steps in Section III-C3 were performed for the 5-factor solution drawing on *Clarity*, *Spaciousness*, *Treble Discomfort*, *Bass* and *Wearing Comfort*.

Since the 6-factor solution also reported a great fit and describes the additional perceptual dimension of *Openness*, albeit with low factor reliability, a nested 5- and 6-factor HAQI model is proposed. Additionally to the just discussed factors, three items for global judgments, namely *liking*, *transparency of the playback system* and *perceived value*, were added to the questionnaire as single items. *liking* was included as a direct measure of general preference. *transparency of the playback system* was retained because only approximately 50% of its variance was explained by the HAQI dimensions[9], suggesting that it captures aspects not fully represented by the latent factor structure. *perceived value* was retained because it provides a complementary value-oriented perspective and was likewise not fully explained by the HAQI dimensions ($R^2 = .56$). The final HAQI questionnaire is presented in Fig. 2.

Scalar invariance was found for the test of measurement invariance across audio contents with only minor drops in CFI, retaining an excellent fit for all constraint levels. Scalar invariance was also found for the test of measurement invariance across participants with different expertise in acoustics. For the language of the questionnaire, scalar invariance was also found. Detailed fit statistics between constraint levels can be found in Appendix D. Therefore, the factor structure, loadings, as well as intercepts are invariant to the audio content used, as well as the expertise of the participant and the language of the questionnaire. This means that neither a difference in musical content nor acoustical expertise or questionnaire language challenges the validity of the measurement instrument within the present sample.

3) Validity Analysis: The relationships between the HAQI dimensions of the 5-factor solution and the global judgment items are presented in Table III. All reported coefficients are standardized path coefficients and thus indicate the unique statistical relationship between each HAQI dimension and the respective outcome when controlling for the remaining dimensions. For *liking*, *Clarity* was the strongest predictor, followed by *Wearing Comfort*. Smaller coefficients were found for *Bass* and *Treble Discomfort*, with the latter being negatively related to *liking*. In contrast, *Spaciousness* did not significantly predict *liking*. Together, the five HAQI dimensions explained 69.5% of its variance. For *overall acoustic impression*, *Clarity* again displayed the largest standardized coefficient. Smaller positive relationships emerged for *Bass* and notably *Wearing Comfort*, whereas *Treble Discomfort* was negatively related to the evaluation. No significant relationship was found for *Spaciousness*. Overall, the five dimensions accounted for 76.9% of the variance in *overall acoustic impression*. For

[8]AVE and McDonald's ω for all factor solutions are provided in Appendix C.

[9]More on this in Section III-C3.

TABLE I
FIT STATISTICS OF THE CFA MODELS. MLR ESTIMATION WAS USED, THEREFORE ROBUST FIT INDICES ARE REPORTED.

Model	# of items	χ^2	df	p	CFI	TLI	SRMR	RMSEA
7 factors	23	816	209	$< .001$	.940	.927	.071	.062
6 factors	20	432	155	$< .001$	.972	.966	.056	.047
5 factors	17	340	109	$< .001$	.973	.966	.057	.052
5 factors (with *Wear Perception)*	17	561	109	$< .001$	.932	.915	.071	.075
4 factors	14	294	71	$< .001$	.963	.953	.059	.056

TABLE II
HETEROTRAIT–MONOTRAIT (HTMT) RATIOS FOR THE FIVE-FACTOR HAQI MODEL (ROUNDED TO TWO DECIMALS).

	Clar.	Spac.	Trebl.	Bass	Wear.
Clarity	1.00	–	–	–	–
Spaciousness	0.84	1.00	–	–	–
Treble Discomfort	0.32	0.35	1.00	–	–
Bass	0.25	0.32	0.25	1.00	–
Wearing Comfort	0.40	0.42	0.18	0.09	1.00

transparency of the playback system, *Clarity* was the only significant predictor . The model explained 50.3% of the variance in this evaluation. A similar pattern emerged for *naturalness*. *Clarity* showed the strongest positive relationship with this outcome, while *Treble Discomfort* was negatively related to it. The coefficients of the remaining dimensions did not reach statistical significance. Collectively, the HAQI dimensions accounted for 71.7% of the variance in *naturalness*. Finally, the results for *perceived value* closely resembled those for *liking*. *Clarity* and *Wearing Comfort* were the strongest predictors. Smaller coefficients were observed for *Bass* and *Treble Discomfort*, whereas *Spaciousness* did not significantly predict the outcome. Together, the five dimensions explained 56.2% of the variance in *perceived value*. The SEM model showed an excellent fit ($\chi^2 = 443$, df $= 169$, $p < .001$, CFI $= 0.980$, TLI $= 0.973$, SRMR $= 0.053$, RMSEA $= 0.044$).

The linear mixed-effects model (LMM) omnibus test results are shown in Table IV. The omnibus test revealed a significant omnibus effect of headphone model on factor scores ($p < .001$). Furthermore, the interaction between headphone model and HAQI dimension was significant ($p < .001$), indicating that differences between headphone models depended on the perceptual dimension considered. Presentation position also showed a significant positive association with factor scores ($p < .001$). The estimated effect was 0.03, meaning that factor scores increased by an average of 0.03 standard deviations for each subsequent position in the presentation sequence. The institution at which the experiment was conducted was significantly associated with mean factor scores ($p = .013$), and the interaction between institution and HAQI dimension was also significant ($p < .001$). Thus, differences between institutions varied across the latent perceptual dimensions. Under the applied deviation-coding scheme, the parameter estimates indicated that one institution had adjusted factor scores that were, on average, 0.429 standard deviations above the grand mean ($p = .005$). However, this institution had recruited relatively few participants at the time of analysis, which may have contributed to the comparatively large estimate. Dimension-specific estimates suggested that the differences between institutions were most pronounced for *Spaciousness*, *Wearing Comfort*, and *Treble Comfort*. These institutional differences should be interpreted cautiously because the institutions differed in their participant samples and did not evaluate identical sets of headphones. The extended model using the factor score of the preceding headphone as a contextual predictor revealed no statistically significant overall association with the current score ($p = .069$). However, this association differed significantly across HAQI dimensions ($p = .013$). The strongest dimension-specific association was observed for *Treble Comfort*: a one-unit increase in the preceding headphone's factor score was associated with a 0.130-unit increase in the current headphone's factor score ($p = .004$), indicating a small positive assimilation effect.

The LMM in which headphone models were grouped by headphone type produced a broadly similar pattern of results. Most importantly, the interaction between headphone category and HAQI dimension was significant ($p < 0.001$), indicating that the perceptual profiles differed across headphone categories. The EMMs and their confidence intervals are presented in Fig. 3. Descriptively, extra-aural and the various circum-aural open-back headphone categories showed some of the highest estimated scores for *Clarity* and *Spaciousness*. The planar-magnetic open-back and circum-aural closed-back headphones showed comparatively high *Bass* scores. Supra-aural, insert, and bone-conduction headphones showed comparatively low *Wearing Comfort* scores. The supra-aural closed-back headphones also showed low estimated scores for *Clarity* and *Spaciousness*, whereas their *Treble Comfort* and *Bass* scores were closer to the overall average. The investigated insert headphones generally showed relatively low factor scores, with *Bass* as an exception, for which they showed a higher-than-average score. Open-ear headphones likewise showed relatively low factor scores, particularly for *Bass*, but positive scores for *Wearing Comfort*. Finally, the investigated bone-conduction headphones showed particularly low estimated scores for *Clarity*, *Spaciousness*, and *Bass*. Their *Treble Comfort* scores were comparatively higher, although still below the overall average.

IV. DISCUSSION

In this thesis, a psychometric measurement instrument for headphone perception, the *Headphone Audio Quality Inventory* (HAQI), was developed. It addresses the lack of a psychometrically grounded and validated measurement tool for modern headphones. First, a CV comprising attributes that

Model		Factors	Items	β	I	Poles	Explanation
6-factor HAQI	5-factor HAQI	**Clarity**	Transparency	0.89	58	blurred – transparent	Describes how easily individual sounds can be distinguished within the overall mix; a transparent sound is open and well-structured, a blurred one is dense and unclear.
			Clarity	0.88	62	unclear – clear	Clarity of the reproduction of voices and instruments.
			Presence	0.83	61	not present – present	Describes how clearly and tangibly voices and instruments appear in the sound image.
			Impulse fidelity	0.82	57	muddy – punchy	Indicates how precisely transient sound events (e.g. plucked strings, percussion) are perceived.
			Brilliance	0.81	55	not brilliant – brilliant	Describes how clear, lively, and finely resolved the sound appears in the treble; high brilliance makes details sparkle.
		Spaciousness	Localizability	0.89	54	imprecise – precise	Describes how well the direction and distance of sound sources can be defined.
			Sense of direction	0.88	56	imprecise – precise	Describes how well the direction of sound source(s) can be defined.
			Sense of distance	0.84	50	imprecise – precise	Describes how well the distance between sound source(s) and listener can be defined.
		Treble discomfort	Shrillness	0.84	32	not shrill – shrill	Describes an overemphasized, shrill sound impression that feels intrusive and fatiguing.
			Piercing sound	0.78	36	not piercing – piercing	Describes a sound that stands out unpleasantly in the treble; it feels piercing and can be fatiguing.
			Hissing	0.76	30	not hissy – hissy	Excessive hissing sound in the treble, making sibilants like cymbals sound piercing.
		Bass	Sub-bass	0.87	47	not perceptible – strongly perceptible	Indicates how strongly very low bass frequencies (e.g. bass drums in electronic music) are reproduced.
			Punch	0.73	48	weak – powerful	Indicates how strongly impacts (e.g. drums, bass) are perceived, almost as if they can be physically felt.
			Bass emphasis	0.70	40	not emphasized – overemphasized	Low frequencies are more pronounced relative to others.
		Wearing comfort	Ear comfort	0.87	62	unpleasant – pleasant	The comfort directly at the outer ear.
			Material comfort	0.86	63	unpleasant – pleasant	The comfort of the material in contact with the head, especially around the ears.
			Clamp force distribution	0.72	64	unpleasant – pleasant	Describes how the clamping force is distributed on the head.
		Openness	Airiness	0.82	53	not airy – airy	The sound conveys a sense of openness and spaciousness.
			Veiledness	-0.72	45	open – veiled	The sound conveys a sense of openness versus veiling.
			Spatial Openness	0.71	54	closed – open	Describes how open or closed the acoustic space is perceived.
Single items			Liking			I don't like it – I like it	Overall assessment for the headphones.
			Transparency of the playback system			low – high	The headphones are not perceived as a sound-altering system.
			Perceived value			low – high	Describes how valuable the headphones appear.

Fig. 2. The complete nested HAQI of the 5- and 6-factor solution, including normalized factor loadings β and intercepts I taken from the 6-factor solution, as well as the used poles and explanation of each item. The intercepts are rounded to integers and are based on a 100-point scale. Additionally, three global judgment items have been added as single items.

extensively describe headphone sound perception and haptic aspects was developed through an expert focus group, resulting in 68 attributes accompanied by corresponding scales and descriptions. Based on this vocabulary, a listening experiment was conducted with 104 participants at eight institutions, employing a diverse pool of 24 headphone models. Based on the participants' responses, latent factors of headphone perception were identified through EFA, and 5- and 6-factor measurement models were constructed and evaluated using CFA. The factors were interpreted as *Clarity*, *Spaciousness*, *Treble Discomfort*, *Bass*, and *Wearing Comfort*, with the 6-factor solution additionally including *Openness*.

Clarity describes a detailed and clear sound in which individual sound sources can be easily distinguished. It is by far the strongest predictor of global judgments such as *liking*, *transparency of the playback system*, and *perceived value*. It resembles the "Clearness/Distinctness" dimension identified in [14], which included the items "clear" and "feeling of presence", similar to the HAQI items *clarity* and *presence*. The PAQI's "Transparency" factor [13] also appears to measure the same construct using almost identical, albeit fewer, items. [14, p. 1029] reports that "systems with broad frequency range, fairly flat frequency response, and low nonlinear distortions are [...] rated favorably", whereas "narrow-band systems, systems with marked resonance peak(s), and systems with more distortion get poorer ratings". This suggests that *Clarity* depends primarily on the faithful reproduction of the signal, a conclusion also reached in [13]. Notably, items describing

TABLE III
STANDARDIZED REGRESSION COEFFICIENTS OF THE HAQI DIMENSIONS ON GLOBAL JUDGMENT ITEMS INSIDE THE SEM.

	Dependent variable				
HAQI dimension	*liking*	*overall acoustic impression*	*transparency of the playback system*	*naturalness*	*perceived value*
Clarity	0.51^{***}	0.67^{***}	0.70^{***}	0.71^{***}	0.51^{***}
Spaciousness	0.02	0.03	-0.05	0.04	-0.05
Bass	0.12^{***}	0.14^{***}	0.06	0.02	0.11^{***}
Treble discomfort	-0.11^{**}	-0.15^{***}	-0.04	-0.15^{***}	-0.12^{**}
Wearing comfort	0.35^{***}	0.11^{***}	0.04	0.06	0.31^{***}
R^2	0.695	0.769	0.503	0.717	0.562

Note. R^2 denotes the proportion of variance explained in each dependent variable. $^{*}p \leq .05$, $^{**}p \leq .01$, $^{***}p \leq .001$.

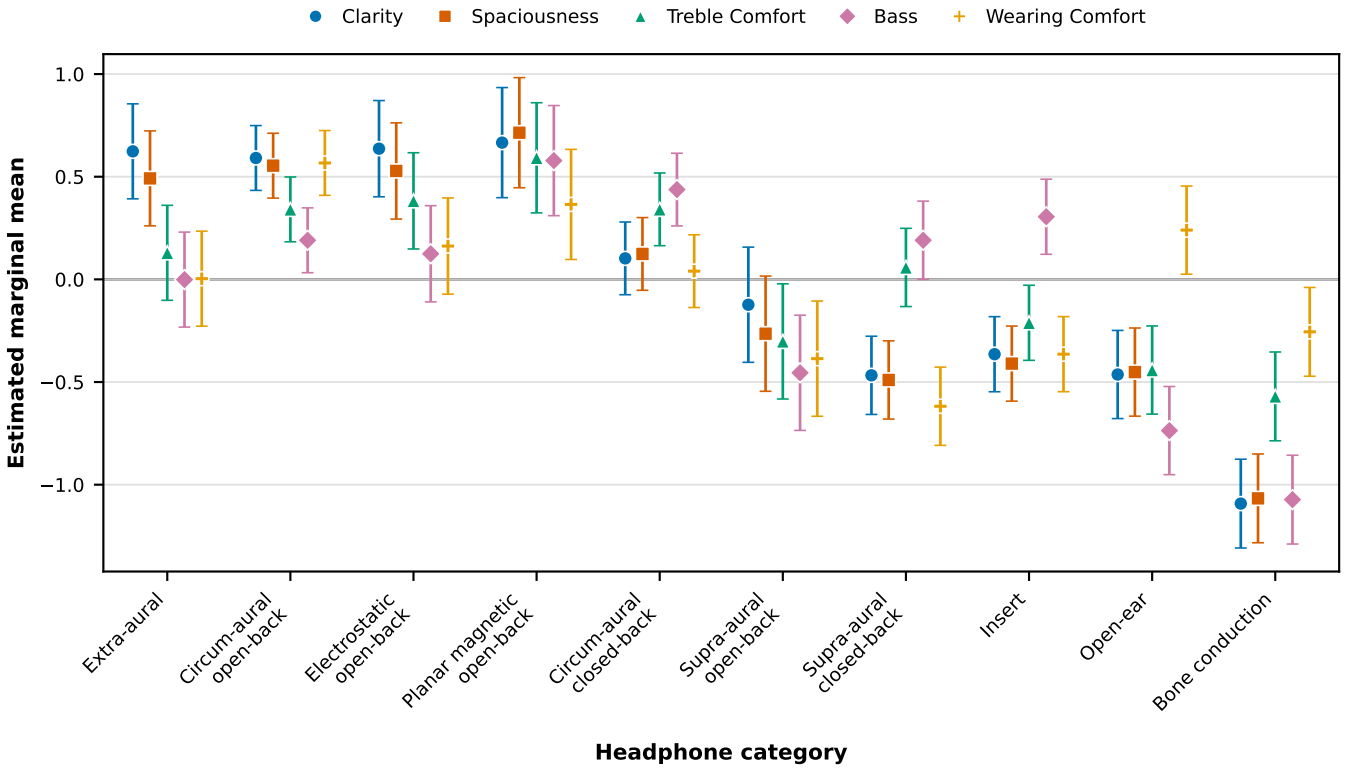


Fig. 3. Estimated marginal means (EMMs) including 95 % confidence intervals of each HAQI dimension by headphone category.

TABLE IV
FIXED EFFECTS OMNIBUS TESTS FOR THE LMM WITH HEADPHONES SEPARATED BY MODEL

	F	df	df (res)	p
Presentation position	13.73	1	3305	$< .001$
HAQI dimension	0.28	4	3300	.889
Institution	2.69	7	115	.013
Headphone model	59.99	23	3303	$< .001$
HAQI dim. × Presentation position	1.69	4	3300	.150
HAQI dim. × Institution	2.69	28	3300	$< .001$
HAQI dim. × Headphone model	5.81	92	3300	$< .001$

Note. Degrees of freedom of the residuals rounded to the nearest integer.

midrange colorations, such as *boxiness* and *nasality*, primarily loaded negatively on this factor in the initial EFA but were removed during the CFA, as their removal resulted in a better explanation of the factor. [2, p. 324] summarizes that "preference" and attributes such as "neutrality", "transparency", and "resolution" all describe the same concept: the absence of audible problems.

Spaciousness describes the precise localization and spatial perception of sound sources within the sound image. When *Clarity* is held constant, *Spaciousness* has no significant direct relationship with *liking* or any of the other global judgments. *Spaciousness* is therefore possibly determined to some extent by a clear and transparent sound. The PAQI [13] also includes a latent factor termed "Spaciousness"; however, its underlying items differ from those of the HAQI. They describe the perception of a wide and deep sound image rather than the precision with which sound sources can be localized along these dimensions. One possible explanation is that, in PA sound reinforcement scenarios, localization strongly depends on the listener's position within the audience. With headphones, by

contrast, the listener is in a "sweet spot" for a balanced stereo image, with timing and level differences between the left and right channels kept to a minimum, potentially enabling a more precise perception of the sound image.

Bass describes the impact or punch of low frequencies rather than merely their emphasis. The perception of sub-bass contributes substantially to this dimension. *Bass* shows a significant but small positive association with *liking*, *overall acoustic impression*, and *perceived value*, but notably not with *transparency of the playback system* or *naturalness*. This pattern might indicate that *Bass* contributes to preference without affecting evaluations of transparent or natural sound to the same extent. [2] found that bass extension, meaning perceptible sub-bass, and smoothness accounted for approximately 30% of the perceived sound quality of loudspeakers in rooms. Judging from the present findings, *Bass* also has a positive but likely smaller effect on perceived headphone sound quality.

Treble Discomfort describes discomfort associated with perceived excessive or disturbing high-frequency content. It has a small but significant negative relationship with most global judgments when the perception of all other latent dimensions, such as *Clarity*, is held constant. It is therefore somewhat distinct from the perception of *Clarity*. [14] describes a dimension termed "Disturbing Sounds", which includes attributes such as "hissing" and generally characterizes systems with an elevated high-frequency response. The emergence of *Treble Discomfort* as a distinct perceptual dimension might be particularly relevant to headphone reproduction. The sound pressure at the eardrum depends on the acoustic coupling between the headphone and the listener's outer ear, including the geometry and acoustic impedance of the pinna and ear canal. Consequently, the headphone-to-eardrum transfer function can vary substantially across listeners and headphone models, potentially producing listener-specific resonances [41]. These variations are most pronounced at high frequencies, where narrow-band peaks and dips occur [42].

Wearing Comfort is largely self-explanatory. It is primarily characterized by comfort at the pinna and the rest of the head, resulting from the materials used and the force exerted by the headphones. A substantial direct effect of *Wearing Comfort* on *liking* and the *perceived value* of the headphones was observed. [19] similarly reported a strong correlation between headphone comfort and overall evaluation. This finding emphasizes that headphone design should devote considerable attention not only to sound but also to construction and comfortable wear. Notably, a significant but small relationship between *Wearing Comfort* and *overall acoustic impression* was also observed. One possible explanation is an interaction between the somatosensory and auditory systems during headphone use. However, because the *Wearing Comfort* of a headphone was always confounded with its sound signature, only limited conclusions can be drawn. A further listening experiment is therefore proposed in which a subset of headphones differing in comfort is used and comfort and sound are varied independently through headphone transfer function (HpTF) equalization.

Openness describes the extent to which the sound image is perceived as open. [14] identified a dimension termed 'feeling of space", which describes a very similar perception. This dimension is particularly relevant in the context of headphones, for which in-head localization and, conversely, *externalization* are prevalent perceptual phenomena. The item *externalization* loaded primarily on this factor in the initial EFA but was subsequently removed from the final model, as the remaining items provided a better explanation of *Openness*.

Further perceptual dimensions of headphone sound may exist but may not have been identified in the present study. For example, the initial EFA suggested a factor interpreted as *Wear Perception*, but this factor was found to be unstable. The number and diversity of acoustic conditions may have limited the independent variation of additional perceptual qualities. Although more than 72 acoustic conditions[10] represent a comparatively broad stimulus set, they may not have covered sufficiently independent combinations of attributes to reveal further stable dimensions. A listening experiment employing a substantially larger and more systematically varied stimulus set might therefore reveal an additional dimension, although such an experiment would be difficult to realize within a single study.

Regarding the psychometric quality of the HAQI, the 5-factor solution demonstrates excellent construct reliability and good convergent validity. Discriminant validity was also generally supported, although the HTMT ratio between *Clarity* and *Spaciousness* was only marginally below the applied threshold, indicating that the empirical distinction between these two factors should be interpreted with some caution. Scalar measurement invariance was supported across audio content, acoustic expertise, and questionnaire language. The more fundamental requirement of metric invariance therefore permits covariance-based analyzes and comparisons of structural relationships, such as correlations and regression coefficients, across these groups. Scalar invariance additionally permits comparisons of latent factor means within a common measurement model. Nevertheless, these findings are subject to several limitations. Acoustic expertise was operationalized as the number of years of professional or educational experience in acoustics, which does not necessarily reflect extensive experience with headphone acoustics in particular. A more targeted measure would assess the duration and nature of participants' professional or educational experience with headphones and, ideally, complement this information with a direct measure of relevant listening expertise. Furthermore, the measurement invariance established across audio content is limited to the musical content included in the experiment, namely orchestral, rock, and blues music. Invariance across substantially different musical genres or other types of audio content, such as speech, film sound, or video-game audio, therefore cannot be assumed without further investigation. Finally, the findings regarding questionnaire language should be interpreted with caution because language was partly confounded with institution and, consequently, with the headphone models evaluated. The institution located outside Germany conducted its experiments in

[10]Resulting from all combinations of headphone, audio content, and ANC mode.

English, whereas the vast majority of experiments in Germany were conducted in German. Thus, although the results support measurement invariance across the German and English questionnaire versions within the present sample, they do not provide a fully independent test of language effects.

Two approaches are proposed for applying the HAQI in headphone evaluations. The first is a simplified scoring approach in which all items belonging to the factors of interest are administered, and loading-weighted composite scores are calculated. For each factor, the item responses can be multiplied by the corresponding normalized factor loadings reported in Fig. 2, summed, and divided by the sum of the applied weights. Optionally, the model-implied intercept of each item can first be subtracted from its response, yielding scores that are approximately centered relative to the reference point of the original calibration model. These composite scores may facilitate descriptive comparisons between headphones but should not be interpreted as model-based factor scores on the exact latent scale. Moreover, this simplified scoring procedure neither accounts for measurement error nor provides a test of measurement invariance. The second approach is to specify the complete HAQI measurement model[11] in SEM software such as *lavaan*. This requires the factor structure and the estimated item loadings, intercepts, residual variances and factor covariances. When the original parameter estimates and factor scaling are fixed, the fit of the calibrated HAQI model to a new dataset can be assessed and model-based factor scores can be estimated.

The LMM revealed significant differences in the multidimensional perceptual profiles described by the HAQI both between individual headphone models and between the headphone categories represented in the experiment. The comparatively high *Clarity* and *Spaciousness* scores of the tested open-back headphones may partly reflect the particular selection of professional and high-fidelity models in this category, which were designed to provide a relatively neutral and uncolored presentation. Their high *Spaciousness* scores might additionally be related to their open-back construction. However, because construction type, model selection, and intended application were not varied independently, their respective contributions cannot be distinguished based on the present results. The relatively high *Bass* scores observed for the tested closed-back and insert headphones may be related to their acoustic coupling to the ear. A sufficiently sealed coupling facilitates the maintenance of sound pressure at low frequencies, whereas leakage can substantially attenuate the low-frequency response of headphones [43]. Nevertheless, the amount and character of the reproduced bass remain dependent on the tuning and fit of the individual model and cannot be inferred from construction type alone. The high *Bass* score of the planar-magnetic headphone probably results from a low frequency response extending well into the sub-bass. The relatively low *Wearing Comfort* scores of the tested supra-aural and insert headphones may have resulted from direct contact with the pinna and ear canal entrance, respectively. The low comfort rating of the bone-conduction headphone might partly be associated with a tingling sensation at high signal amplitudes that was frequently reported by participants. Conversely, the tested open-ear headphones received comparatively favorable comfort ratings, possibly because they made only light contact with the pinna. Their low *Bass* scores may be related to the absence of sealed acoustic coupling to the ear canal, which makes the efficient reproduction of low-frequency sound pressure more challenging. Finally, the bone-conduction headphone models included in the present experiment received low scores across most HAQI dimensions. Within the range of devices examined here, their advantages in maintaining environmental awareness therefore appeared to be accompanied by substantial limitations in the reproduction of a clear and spacious sound with impactful bass.

Several limitations should be considered when using the EMMs to compare the perceptual profiles of headphones across institutions. The LMM adjusts the estimates for the effects explicitly represented in the model, including institution, presentation position, headphone model, and their dimension-specific relationships. However, such adjustment does not guarantee the complete removal of systematic differences. Each institution evaluated a particular subset of headphones in a fixed order, resulting in partial confounding between headphone model, institution, presentation position, and preceding headphone. This confounding was reduced by presenting individual models at multiple institutions and varying their positions across institutional sequences. Nevertheless, because not every headphone was evaluated at every institution and presentation position, the model cannot fully separate these influences without assuming that the estimated institution and position effects apply similarly to headphone models for which the corresponding combinations were not observed. Differences between institutional stimulus sets may also have affected how participants used the rating scales. Stimulus-spacing and range-equalizing biases can cause perceptual differences to be expanded or compressed depending on the distribution and range of the stimuli presented within a particular context [44]. Although each institution was assigned a perceptually diverse set of headphones, practical constraints prevented the stimulus ranges from being fully standardized. Common low and high anchor headphones could provide more consistent references across institutions in future experiments. Furthermore, the linear effect of presentation position included in the LMM represents an overall trend across the experimental sequence but cannot account for predecessor-specific carryover effects. The exploratory analysis of sequential effects indicated that such effects may have occurred for some HAQI dimensions. More extensive sequence randomization or counterbalancing would be required to separate carryover effects from headphone and institution effects more reliably. An additional familiarization or recalibration phase after the mid-session break might also help maintain a common perceptual reference. Finally, the experiment was not visually blinded because participants had to handle and put on the headphones. Although model names and manufacturer logos were concealed, distinctive or familiar models may still have been recognized, allowing prior expectations or brand-related attitudes to influence the

[11]The complete 5- and 6-factor models are documented in the Digital Appendix, *HAQI_fully_defined_models.xlsx*

ratings [44]. The EMMs therefore provide useful model-adjusted comparisons within the present experiment, and their confidence intervals quantify uncertainty under the specified model. They should not, however, be interpreted as bias-free estimates of absolute differences between headphone models or categories, because their confidence intervals do not capture additional systematic uncertainty resulting from differences in the headphone sets presented at each institution, predecessor-specific carryover effects, or visual recognition.

Several potential improvements and directions for future research have already been identified throughout this discussion. In interpreting the EMMs, possible relationships between the HAQI dimensions and physical properties of the headphones were proposed. However, these explanations remain speculative. Future studies should therefore systematically investigate whether and to what extent HAQI factor scores can be predicted from acoustic parameters. The most important next step is an independent evaluation of the final HAQI measurement model. As discussed in Section II-C3, modification indices were used to adjust the CFA model. Consequently, the final model was developed and evaluated using the same data, reducing the strictly confirmatory character of the analysis and potentially resulting in an optimistic assessment of model fit. A further listening experiment should therefore evaluate the fit of the current model in an independent sample without introducing additional data-driven modifications. Ideally, this study would include both new participants and at least some previously untested headphone models, allowing the generalizability of the measurement structure across listeners and stimuli to be examined. If additional participants instead complete the present experiment, their data should be retained as an independent validation sample. Finally, the temporal stability of the HAQI should be examined in a retest study. Previous participants could repeat the evaluation of the same headphone conditions after an appropriate interval, allowing the assessment of item-level retest reliability, latent factor stability, and measurement invariance across measurement occasions. Together, independent model evaluation and temporal reliability assessment would provide a substantially stronger psychometric validation of the HAQI.

V. Conclusion

In this thesis, the *Headphone Audio Quality Inventory* (HAQI), a psychometrically grounded instrument for assessing headphone sound and comfort perception, was developed. First, a consensus vocabulary was established through an expert focus group. Based on this vocabulary, a multi-institutional listening experiment was conducted using a diverse set of headphone models. Exploratory and confirmatory factor analyzes resulted in a five-factor solution comprising 17 items and an alternative six-factor solution comprising 20 items, plus 3 single items describing global judgments. The five core dimensions were interpreted as *Clarity*, *Spaciousness*, *Treble Discomfort*, *Bass*, and *Wearing Comfort*, while the six-factor solution additionally included *Openness*. The resulting measurement models demonstrated high construct reliability and generally good convergent and discriminant validity, as well as scalar measurement invariance across the examined audio content, levels of acoustic expertise, and questionnaire languages. The HAQI dimensions differentiated the perceptual profiles of the headphone models and categories examined in the listening experiment. Furthermore, the five-factor structural model explained roughly 70% of the variance in *liking* and 77% of the variance in *overall acoustic impression*, demonstrating that the identified dimensions capture substantial aspects of global headphone evaluations. The HAQI questionnaire and its associated measurement model can therefore complement objective acoustic measures, such as frequency response, by providing a multidimensional account of perceived headphone quality. Establishing relationships between HAQI dimensions and physical measurements could ultimately translate perceptual qualities into measurable engineering objectives, thereby enabling perceptually informed decisions throughout the headphone development process. The most important next step is to evaluate the final measurement model in an independent sample, followed by an assessment of its temporal reliability and stability in a retest study. Further research should systematically investigate which physical properties predict the individual HAQI dimensions and which perceptual profiles contribute to favorable global evaluations. Such knowledge could help define technical design targets for achieving desired perceptual qualities and thus support the development and optimization of future headphones. Finally, the observed relationship between wearing comfort and global sound-quality judgments warrants further investigation under conditions in which physical comfort and acoustic reproduction can be varied independently.

Acknowledgments

The author would like to thank the experts in headphone sound reproduction who participated in the focus group (listed in alphabetical order): Konstantin Davy, Vera Erbes, Axel Grell, Dmitry Grigoriev, Julia Havenstein, Jonas Hüsen, Thomas Karcher, Alexander Lindau, and Stefan Weinzierl.

The author would also like to express his sincere gratitude to everyone involved in organizing and conducting the listening experiments at the following institutions: the Audio Technology Group, Chalmers University of Technology; the Chair of Audio Engineering, Mittweida University of Applied Sciences; the Department of Technology and Health for Humans, Jade University of Applied Sciences; the Institute for Hearing Technology and Acoustics, RWTH Aachen University; the Institute of Communications Technology, Leibniz University Hannover; the Institute of Sound and Vibration Engineering, Düsseldorf University of Applied Sciences; the Research Group Signal Processing and Virtual Acoustics, University of Rostock; and the Technical Psychoacoustics Research Group, Aalto University.

References

[1] International Federation of the Phonographic Industry, "Engaging with music 2023," London, 2023. [Online]. Available: https://www.ifpi.org/wp-content/uploads/2023/12/IFPI-Engaging-With-Music-2023_full-report.pdf

[2] F. E. Toole, S. Olive, and T. Welti, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers, Rooms and Headphones*, 4th ed. Focal Press, 2025. [Online]. Available: https://doi.org/10.4324/9781003477495

[3] International Telecommunication Union, "Artificial ears," International Telecommunication Union, Geneva, Switzerland, ITU-T Recommendation P.57, Jun. 2021. [Online]. Available: https://www.itu.int/rec/T-REC-P.57-202106-I/en

[4] G. Theile, "On the Standardization of the Frequency Response of High-Quality Studio Headphones," *Journal of the Audio Engineering Society*, vol. 34, pp. Paper 956–969, December 1986.

[5] G. Lorho, "Subjective evaluation of headphone target frequency responses," *Journal of the Audio Engineering Society*, no. 7770, may 2009.

[6] S. E. Olive, "The perception and measurement of headphone sound quality: What do listeners prefer?" *Acoustics Today*, vol. 18, no. 1, pp. 58–67, 2022. [Online]. Available: https://doi.org/10.1121/AT.2022.18.1.58

[7] S. Bech and N. Zacharov, *Quantification of Impression*. John Wiley & Sons, Ltd, 2006, ch. 4, pp. 39–96. [Online]. Available: https://onlinelibrary.wiley.com/doi/abs/10.1002/9780470869253.ch4

[8] EBU Tech. 3286, *Assessment methods for the subjective evaluation of the quality of sound programme material – Music*. Geneva, Switzerland: European Broadcasting Union, 1997. [Online]. Available: https://tech.ebu.ch/publications/tech3286

[9] ITU-R BS.1284-2, *General methods for the subjective assessment of sound quality*. Geneva, Switzerland: International Telecommunication Union, 2019. [Online]. Available: https://www.itu.int/rec/R-REC-BS.1284-2-201901-I/en

[10] T. Holm Pedersen and N. Zacharov, "The development of a sound wheel for reproduced sound," *138th Audio Engineering Society Convention 2015*, vol. 2, pp. 765–777, 01 2015.

[11] S. Weinzierl, S. Lepa, and D. Ackermann, "A measuring instrument for the auditory perception of rooms: The room acoustical quality inventory (RAQI)," *The Journal of the Acoustical Society of America*, vol. 144, no. 3, pp. 1245–1257, 09 2018. [Online]. Available: https://doi.org/10.1121/1.5051453

[12] Z. Schärer Kalkandjiev and S. Weinzierl, "An instrument for measuring the perception of room acoustics from the perspective of musicians: The stage acoustic quality inventory (STAQI)," *Fortschritte der Akustik - DAGA 2018 : 44. Jahrestagung für Akustik*, pp. 1747–1750, 03 2018.

[13] S. Büchner, S. Lepa, A. Lindau, and S. Weinzierl, "A public address system quality inventory (PAQI)," *Journal of the Audio Engineering Society*, 2026, in press.

[14] A. Gabrielsson and H. Sjögren, "Perceived sound quality of sound-reproducing systems," *The Journal of the Acoustical Society of America*, vol. 65, no. 4, pp. 1019–1033, 04 1979. [Online]. Available: https://doi.org/10.1121/1.382579

[15] G. Lorho, "Perceived quality evaluation: An application to sound reproduction over headphones," Ph.D. dissertation, Aalto University, Finland, 2010.

[16] N. Zacharov, J. Ramsgaard, G. Le Ray, and C. V. Jørgensen, "The multidimensional characterization of active noise cancelation headphone perception," in *2010 Second International Workshop on Quality of Multimedia Experience (QoMEX)*, 2010, pp. 130–135.

[17] C. P. Volk, M. Lavandier, S. Bech, and F. Christensen, "Identifying the dominating perceptual differences in headphone reproduction," *The Journal of the Acoustical Society of America*, vol. 140, no. 5, pp. 3664–3674, 11 2016. [Online]. Available: https://doi.org/10.1121/1.4967225

[18] S. Olive and T. Welti, "The Relationship between Perception and Measurement of Headphone Sound Quality," in *AES Convention 133*, AES. Journal of the Audio Engineering Society, October 2012.

[19] R. Jensen, N. Lauridsen, A. Poulsen, C. Tofte, and F. Christensen, "Analysis of Subjective Evaluation of User Experience with Headphones," in *AES Conference: 2016 AES International Conference on Headphone Technology*, AES. Journal of the Audio Engineering Society, August 2016.

[20] J. C. Gäde, K. Schermelleh-Engel, and H. Brandt, *Konfirmatorische Faktorenanalyse (CFA)*. Berlin, Heidelberg: Springer Berlin Heidelberg, 2020, pp. 615–659. [Online]. Available: https://doi.org/10.1007/978-3-662-61532-4_24

[21] D. W. C. Stewart and P. N. Shamdasani, "Focus groups : theory and practice," Newbury Park u.a., 1990.

[22] A. Bogner and M. Leuthold, *Was ich dazu noch sagen wollte…"*. Wiesbaden: VS Verlag für Sozialwissenschaften, 2002, pp. 155–172. [Online]. Available: https://doi.org/10.1007/978-3-322-93270-9_7

[23] W. Hoeg, F. Camerer, and I. Mielke, *19 Qualitätssicherung*. Berlin, Boston: De Gruyter Saur, 2023, pp. 1511–1590. [Online]. Available: https://doi.org/10.1515/9783110759921-019

[24] R. Schulte, "Public Address System Quality Inventory (PAQI) – Statistische Erhebung zur Entwicklung eines psychologisch fundierten Messinstruments für die Qualität von Beschallungsanlagen auf Veranstaltungen," Master's thesis, Berliner Hochschule für Technik, 06 2023.

[25] Headphone.shop. (2024, Dec.) How to describe sound: The audiophile's guide to sonic vocabulary. Headphone.shop. [Online]. Available: https://headphone.shop/how-to-describe-sound/

[26] V. McCuaig. (2023, Jun.) Sound quality glossary. RTINGS.com. [Online]. Available: https://www.rtings.com/headphones/learn/sound-quality-glossary

[27] Headphonesty. [glossary] an audiophile guide to describing sound. Headphonesty. [Online]. Available: https://www.headphonesty.com/sound-description-glossary/

[28] LimeSurvey GmbH, "Limesurvey: An open source survey tool," https://www.limesurvey.org/, Hamburg, accessed: 2026-07-20.

[29] W. Revelle, *psych: Procedures for Psychological, Psychometric, and Personality Research*, 2025, r package. [Online]. Available: https://cran.r-project.org/package=psych

[30] The jamovi project, *jamovi*, 2025, computer software. [Online]. Available: https://www.jamovi.org

[31] H. F. Kaiser, "An index of factorial simplicity," *Psychometrika*, vol. 39, no. 1, pp. 31–36, Mar. 1974. [Online]. Available: https://doi.org/10.1007/BF02291575

[32] Y. Rosseel, "lavaan: An r package for structural equation modeling," *Journal of Statistical Software*, vol. 48, no. 2, pp. 1–36, 2012. [Online]. Available: https://doi.org/10.18637/jss.v048.i02

[33] L. Hu and P. M. Bentler, "Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 6, no. 1, pp. 1–55, 1999. [Online]. Available: https://doi.org/10.1080/10705519909540118

[34] R. P. McDonald, "Test theory: a unified treatment," Mahwah, NJ u.a., 1999.

[35] C. Fornell and D. F. Larcker, "Evaluating structural equation models with unobservable variables and measurement error," *Journal of Marketing Research*, vol. 18, no. 1, pp. 39–50, 1981. [Online]. Available: http://www.jstor.org/stable/3151312

[36] J. Henseler, C. M. Ringle, and M. Sarstedt, "A new criterion for assessing discriminant validity in variance-based structural equation modeling," *Journal of the Academy of Marketing Science*, vol. 43, no. 1, pp. 115–135, 2015. [Online]. Available: https://doi.org/10.1007/s11747-014-0403-8

[37] G. W. Cheung and R. B. Rensvold, "Evaluating goodness-of-fit indexes for testing measurement invariance," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 9, no. 2, pp. 233–255, 2002. [Online]. Available: https://doi.org/10.1207/S15328007SEM0902_5

[38] F. F. Chen, "Sensitivity of goodness of fit indexes to lack of measurement invariance," *Structural Equation Modeling: A Multidisciplinary Journal*, vol. 14, no. 3, pp. 464–504, 2007. [Online]. Available: https://doi.org/10.1080/10705510701301834

[39] M. Gallucci, *GAMLj: General Analyses for Linear Models*, 2019, jamovi module. [Online]. Available: https://gamlj.github.io/

[40] D. D. Torrico, W. Jirangrat, J. Wang, P. Chompreeda, S. Sriwattana, and W. Prinyawiwatkul, "Novel modelling approaches to characterize and quantify carryover effects on sensory acceptability," *Foods*, vol. 7, no. 11, p. 186, 2018.

[41] H. Deng and J. Yang, "Modeling and estimating acoustic transfer functions of external ears with or without headphones," *The Journal of the Acoustical Society of America*, vol. 138, no. 2, pp. 694–707, 08 2015. [Online]. Available: https://doi.org/10.1121/1.4926560

[42] K. I. McAnally and R. L. Martin, "Variability in the headphone-to-ear-canal transfer function," *Journal of the Audio Engineering Society*, vol. 50, no. 4, pp. 263–266, 2002.

[43] M. R. Avis and L. J. Kelly, "Principles of Headphone Design—A Tutorial Review," in *AES Conference: UK 21st Conference: Audio at Home*, AES. Journal of the Audio Engineering Society, April 2006.

[44] S. K. Zieliński, F. Rumsey, and S. Bech, "On some biases encountered in modern audio quality listening tests-a review," *Journal of The Audio Engineering Society*, vol. 56, pp. 427–451, 2008. [Online]. Available: https://api.semanticscholar.org/CorpusID:55377153

[45] T. D. Jorgensen, S. Pornprasertmanit, A. M. Schoemann, Y. Rosseel, P. Miller, C. Quick, M. Garnier-Villarreal, J. Selig, A. Boulton, K. Preacher, D. Coffman, M. Rhemtulla, A. Robitzsch, C. Enders, R. Arslan, B. Clinton, P. Panko, E. Merkle, S. Chesnut, J. Byrnes, J. D.

Rights, Y. Longo, M. Mansolf, M. S. Ben-Shachar, and M. Rönkkö, *semTools: Useful Tools for Structural Equation Modeling*, 2019, r package. [Online]. Available: https://CRAN.R-project.org/package=semTools

APPENDIX A
HEADPHONE ALLOCATION

TABLE V
HEADPHONE CONDITIONS AND PRESENTATION ORDER ACROSS INSTITUTIONS.

Order	Institution								
	1	2	3	4	5	6	7	8	9
1	Extra-aural 1	Circum-aural, closed-back 3	Circum-aural, open-back 4	Circum-aural, open-back 1	Circum-aural, open-back, electrostatic	Insert 2	Circum-aural, open-back, planar magnetic	Supra-aural, open-back 1	Insert 4 (ANC on)
2	Open-ear 1	Open-ear 2	Insert 3	Insert 1	Supra-aural, closed-back 2	Supra-aural, open-back 1	Supra-aural, open-back 1	Circum-aural, closed-back 4	Extra-aural 2
3	Supra-aural, open-back 1	Supra-aural, closed-back 1 (ANC off)	Circum-aural, open-back, planar magnetic	Circum-aural, open-back 2	Bone conduction 1	Circum-aural, closed-back 2	Insert 3	Supra-aural, closed-back 2	Bone conduction 2
4	Circum-aural, closed-back 1 (ANC off)	Insert 4 (ANC off)	Supra-aural, closed-back 2	Insert 2	Circum-aural, closed-back 4	Open-ear 1	Open-ear 2	Bone conduction 2	Circum-aural, closed-back 3
5	Bone conduction 1	Circum-aural, open-back, planar magnetic	Circum-aural, closed-back 4	Extra-aural 1	Extra-aural 2	Supra-aural, closed-back 2	Circum-aural, open-back, electrostatic	Circum-aural, open-back 2	Insert 2
6	Supra-aural, closed-back 1 (ANC on)	Circum-aural, closed-back 2	Supra-aural, open-back 2	Open-ear 1	Circum-aural, open-back 1	Extra-aural 1	Insert 1	Circum-aural, open-back 1	Supra-aural, open-back 2
7	Circum-aural, open-back, electrostatic	Circum-aural, closed-back 1 (ANC on)	Circum-aural, open-back 2	Supra-aural, open-back 1	Insert 1	Circum-aural, open-back 4	Circum-aural, open-back 1	Insert 3	Circum-aural, closed-back 4
8	Bone conduction 2	–	–	–	–	–	–	–	–

Note. Data was not yet available for Institution 9 at the time of analysis.

APPENDIX B
EFA RESULTS

TABLE VI
EFA FACTOR LOADINGS FOR THE 7-FACTOR SOLUTION, WITH UNIQUENESS AND MSA VALUES FOR EACH VARIABLE. FACTOR LOADINGS BELOW 0.4 WERE EXCLUDED FROM THIS VISUALIZATION.

	Factor								
	1	2	3	4	5	6	7	Uniqueness	MSA
25 *dullness*	-0.78							0.36	0.97
16 *presence*	0.72							0.29	0.98
21 *brilliance*	0.71							0.31	0.97
38 *clarity*	0.70							0.23	0.98
15 *directness*	0.69							0.44	0.98
10 *boxiness*	-0.63							0.47	0.97
37 *resolution*	0.63							0.20	0.98
40 *transparency*	0.62							0.27	0.98
39 *coherence*	0.59							0.36	0.98
42 *impulse fidelity*	0.59							0.29	0.98
14 *nasality*	-0.55							0.47	0.97
1 *neutrality*	0.53							0.39	0.97
11 *hollowness*	-0.53							0.49	0.97
2 *balance*	0.52							0.30	0.97
43 *attack*	0.51							0.44	0.96
44 *bass precision*	0.50							0.44	0.97
41 *dynamic range*	0.49							0.31	0.98
33 *source imaging*	0.49							0.23	0.98
4 *midrange emphasis*								0.92	0.82
31 *sense of direction*		0.58						0.28	0.97
47 *extraneous noise*		-0.52						0.65	0.92
30 *sense of distance*		0.52						0.32	0.98
32 *localizability*		0.49						0.31	0.97
28 *sound image width*		0.48						0.25	0.98
45 *distortion*		-0.41						0.43	0.96
26 *left right balance*								0.67	0.95
29 *spatial depth*								0.33	0.98
62 *occlusion noise*								0.71	0.86
27 *directional stability*								0.59	0.96
61 *cable noise*								0.85	0.78
22 *piercing sound*			0.82					0.36	0.90
23 *shrillness*			0.79					0.31	0.95
46 *hissing*			0.66					0.46	0.93
3 *treble emphasis*			0.64					0.58	0.82
13 *tinniness*			0.55					0.45	0.96
18 *aggressiveness*			0.51					0.65	0.75
24 *sharpness*			0.43					0.68	0.90
6 *warmth*			-0.43					0.61	0.95
17 *relaxedness*								0.38	0.98
53 *material comfort*				0.85				0.27	0.90
52 *clamp force distribution*				0.76				0.33	0.88
54 *ear comfort*				0.68				0.41	0.92
55 *wearing stability*				0.63				0.58	0.83
50 *weight distribution*				0.57				0.51	0.82
66 *mechanical stability*				0.55				0.51	0.91
56 *head shape adjustability*								0.61	0.95
60 *mechanical resonances*								0.90	0.71
5 *bass emphasis*					0.73			0.46	0.80
8 *punch*					0.69			0.41	0.93
7 *sub bass*					0.68			0.40	0.93
9 *boominess*	-0.48				0.51			0.49	0.94
12 *thinness*					-0.44			0.33	0.96
48 *loudness*								0.77	0.82
34 *envelopment*								0.57	0.97
20 *airiness*						0.53		0.40	0.97
36 *spatial openness*						0.50		0.51	0.96
35 *externalization*						0.46		0.76	0.88
19 *veiledness*						-0.46		0.49	0.97
59 *hear through quality*								0.51	0.84
49 *weight*							0.70	0.43	0.88
51 *clamp force*							0.69	0.45	0.79
57 *isolation*							0.48	0.42	0.83
58 *perception of own speech*								0.81	0.78

APPENDIX C
FACTOR RELIABILITY AND CONVERGENT VALIDITY OF CFA SOLUTIONS

TABLE VII
FACTOR RELIABILITY AND CONVERGENT VALIDITY ACROSS THE CFA MODELS.

	7 factors		6 factors		5 factors		5 factors (with *Wear Perception*)		4 factors	
Factor	ω_2	AVE	ω_2	AVE	ω_2	AVE	ω_2	AVE	ω_2	AVE
Clarity	.93	.71	.93	.71	.93	.71	.93	.71	.93	.71
Spaciousness	.90	.75	.90	.75	.90	.75	—	—	—	—
Treble Discomfort	.83	.62	.83	.62	.83	.62	.83	.62	.83	.62
Bass	.81	.59	.81	.59	.81	.59	.81	.59	.81	.59
Wearing Comfort	.86	.67	.86	.67	.86	.67	.86	.67	.86	.67
Openness	**.36**	.57	**.36**	.57	—	—	—	—	—	—
Wear Perception	**.66**	**.40**	—	—	—	—	**.10**	.50	—	—

Note. ω_2 denotes McDonald's omega as calculated by `semTools` [45]; AVE denotes average variance extracted. Values below the respective thresholds of $\omega_2 = .70$ and $\text{AVE} = .50$ are shown in bold. Dashes indicate that the factor was not included in the respective model.

APPENDIX D
MEASUREMENT INVARIANCE

TABLE VIII
MEASUREMENT INVARIANCE ACROSS CONTENT GROUPS

Model	χ^2_{robust}	*df*	p	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
Configural	550.5	327	1.276×10^{-13}	0.9672	0.05703	0.06307	—	—	—
Metric	599.2	351	3.109×10^{-15}	0.9625	0.05888	0.07188	-0.004724	0.001850	0.008806
Scalar	628.0	375	4.774×10^{-15}	0.9613	0.05790	0.07314	-0.001229	-0.000989	0.001258

Note. Content 1 (Orchestra) comprised 30 participants and 218 individual ratings, Content 2 (Progressive rock) comprised 34 participants and 247 individual ratings, and Content 3 (Vocal-Blues) comprised 34 participants and 245 individual ratings. Changes in fit indices refer to the immediately preceding model.

TABLE IX
MEASUREMENT INVARIANCE ACROSS LEVELS OF ACOUSTIC EXPERTISE

Model	χ^2_{robust}	*df*	p	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
Configural	395.8	218	1.943×10^{-12}	0.9724	0.05199	0.06117	—	—	—
Metric	401.3	230	1.943×10^{-11}	0.9730	0.05007	0.06296	0.000593	-0.001922	0.001798
Scalar	404.3	242	2.799×10^{-10}	0.9739	0.04804	0.06336	0.000852	-0.002034	0.000395

Note. The group without professional acoustic expertise comprised 43 participants and 313 individual ratings. The group with professional acoustic expertise comprised 55 participants and 397 individual ratings. Changes in fit indices refer to the immediately preceding model.

TABLE X
MEASUREMENT INVARIANCE ACROSS QUESTIONNAIRE LANGUAGES

Model	χ^2_{robust}	*df*	p	CFI	RMSEA	SRMR	ΔCFI	ΔRMSEA	ΔSRMR
Configural	389.4	218	8.308×10^{-12}	0.9738	0.05108	0.06226	—	—	—
Metric	433.1	230	1.321×10^{-14}	0.9689	0.05410	0.06757	-0.004820	0.003023	0.005308
Scalar	462.4	242	5.551×10^{-16}	0.9655	0.05560	0.06877	-0.003456	0.001498	0.001200

Note. The German-language group comprised 76 participants and 560 individual ratings. The English-language group comprised 22 participants and 150 individual ratings. Changes in fit indices refer to the immediately preceding model.