



Audio Engineering Society Conference Paper 3

Presented at the AES International Conference on
Headphone Technology
2025 August 27–29, Espoo, Finland

This paper was peer-reviewed as a complete manuscript for presentation at this conference. This paper is available in the AES E-Library (<http://www.aes.org/e-lib>), all rights reserved. Reproduction of this paper, or any portion thereof, is not permitted without direct permission from the Journal of the Audio Engineering Society.

Evaluating the Effectiveness of Virtual Listening Tests for Balanced Armature Headphone Drivers

Jonas Hüsen¹, Markus von Berg¹, and Jochen Steffens¹

¹University of Applied Sciences Düsseldorf, Institute of Sound and Vibration Engineering (ISAVE)

Correspondence should be addressed to Jonas Hüsen (jonas.huesen@hs-duesseldorf.de)

ABSTRACT

In-ear headphones have become indispensable for performing musicians, sound engineers, and audiophiles; therefore, the precise assessment of their sound quality is crucial for the development and optimization of new models. In particular, the proliferation of Balanced Armature drivers in high-end headphones brings about new challenges, as the established Virtual Listening Tests designed for headphones with electrodynamic drivers may not sufficiently capture the complex acoustic characteristics of in-ear devices with balanced armature drivers. Therefore the study investigates whether standard virtual listening tests methodologies are adequate for balanced armature systems and which specific acoustic and technical parameters a reference in-ear headphone must possess to accurately virtualize balanced armature driver headphones. To this end, listening tests were conducted with an electrodynamic driver in-ear headphone and two balanced armature driver in-ear headphones, a one-way system, and a two-way system. Following the Virtual Listening Test paradigm, the headphones were reciprocally equalized to match each other's frequency response. Participants rated the headphones with and without equalization with regard to localizability, transient fidelity, and low frequency impulse fidelity. The results indicate audible differences across headphone drivers regarding localizability, transient fidelity, low-frequency reproduction and preference. Conclusively, certain rules should be considered when conducting Virtual Listening Tests with headphones containing balanced armature drivers or multi-way systems.

1 Introduction

Accurate sound reproduction in in-ear headphones is fundamental for musicians, sound engineers, and audiophiles, influencing both professional and consumer audio experiences. Virtual listening tests (VLTs) have become a widely used method for evaluating headphones. They provide a controlled and repeatable means of assessing sound quality by simulating the transmission characteristics of different headphones with a single reference device [1, 2]. While this approach has

proven effective for traditional electrodynamic (ED) driver headphones, the frequent use of balanced armature (BA) driver systems introduces new circumstances that remain insufficiently addressed. Unlike ED drivers, which operate on a moving-coil principle, BA drivers operate by means of an armature core that is balanced in a magnetic field and connected to a metal diaphragm by a conduction rod. This results in differences in nonlinearities, transient behavior, and impulse response that may not be fully captured using standard VLT procedures. Due to the fact that a single BA driver

has a limited bandwidth, multi-way BA systems are frequently used, that incorporate crossovers that alter phase and impulse response, further complicating their virtualization on a reference headphone. These fundamental differences raise crucial questions regarding the prerequisites a reference headphone must fulfill to ensure the validity of VLTs for BA-based in-ear monitors (IEMs).

Previous research explored the function and applicability of VLTs for over-ear headphones with ED drivers [3] and in in-ear headphones, again with ED drivers [4]. Both studies indicate applicability for VLTs, but do not consider the influence of different driver types. Notably, the influence of multi-way crossovers in BA systems on perceived localization, transient fidelity, and tonal balance remains unexplored in the context of VLT methodologies. The crossover networks in these multi-way systems can introduce additional challenges, as their frequency- and phase-dependent behavior may be difficult to simulate on a single-driver reference headphone. Previous research has shown that phase discrepancies in crossover networks can affect spatial perception and transient fidelity [5, 6], raising concerns about whether standard VLT methodologies can faithfully reproduce these effects. Furthermore, differences in transient response and distortion characteristics between BA and ED drivers may lead to perceptual artifacts, even when equalized to the same frequency response [7]. There is also a considerable difference in the mass moved in the drivers and therefore a difference in the inertia of the systems. This could result in a significant difference in the impulse fidelity. From previous work it is known that the phase and the non-linear distortions can be adjusted by binaural recordings of the headphones to be tested, i.e. added to the linear phases or the lowest possible distortions of the reference headphones [2, 4]. Since this does not apply to impulse fidelity which results from electro-mechanical properties and cannot be artificially reduced, the simulation would only be possible in one direction.

Given these challenges, the present study aims to systematically investigate the prerequisites for an effective reference headphone in VLTs involving BA driver IEMs. To this end, listening tests were conducted with an ED driver IEM and two BA driver IEMs, including both single-way and multi-way systems. By reciprocally equalizing these headphones and evaluating perceptual attributes such as localizability, transient fidelity, and impulse response, we sought to de-

termine the conditions under which BA driver IEMs can be accurately virtualized. The results of this study will contribute to refining VLT methodologies, ensuring their applicability to the next generation of high-performance in-ear monitoring systems.

2 Methods

To investigate the prerequisites for an effective reference headphone in VLTs involving BA and ED driver in-ear headphones, a listening experiment was conducted. The study aimed to examine the perceptual influence of driver type and multi-way crossover design by reciprocally equalizing different headphone models and evaluating their effect on spatial perception, transient response, and low frequency impulse fidelity as well as their liking.

The listening test was performed in a acoustically treated laboratory to minimize external influences on perception. Three in-ear headphones, given in Table 1, are tested: an ED driver IEM (ED1), a single-way Balanced Armature IEM (BA1), and a two-way Balanced Armature IEM (BA2) containing a crossover.

Table 1: Headphone selection

Des.	Brand	Driver	Crossovers
ED1	Sennheiser	1 Electrodynamic	0
BA1	RH Acoustics	1 Balanced Armature	0
BA2	RH Acoustics	2 Balanced Armature	1

To ensure consistent playback levels across all conditions, the headphones were adjusted to an equivalent sound pressure level of 80 dB(Z). Calibration and frequency response measurements were conducted using a G.R.A.S. measurement microphone in conjunction with an IEC 60318-4 standardized occluded-ear simulator. The resulting frequency responses are given in Figure 1.

2.1 Equalization and Virtualization Procedure

The core method of this study is based on the equalization of the three headphones with finite impulse response (FIR) filters. This process allows a specific investigation of the change in the perceptual characteristics of an IEM when its frequency response is transferred to a different model. The equalization conditions

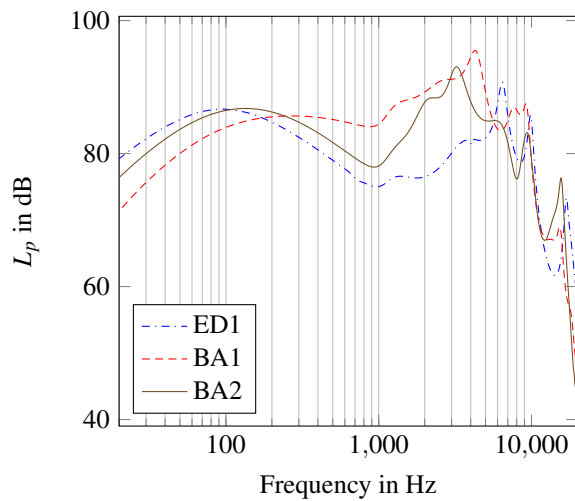


Fig. 1: Frequency response measurements of the three tested in-ear headphones (ED, BA1, BA2) before equalization (left and right side averaged). The variations highlight the differences in their native spectral balance, which the FIR filters aim to compensate for.

are listed in Table 2. In the first three reference conditions, the IEMs were presented without any equalization. Subsequently they were equalized to each other to assess the influence of the driver type and multi-way systems. The condition labels follow the format *source–target*, where the first element denotes the physical IEM worn by the listener, and the second the target of the equalization.

Table 2: Listening test conditions

Condition	Transformation	Research Focus
ED1	no equalization	Reference
BA1	no equalization	Reference
BA2	no equalization	Reference
ED1-BA1	ED1 equalized to BA1	Driver type
BA1-ED1	BA1 equalized to ED1	Driver type
BA1-BA2	BA1 equalized to BA2	Multi-way system
BA2-BA1	BA2 equalized to BA1	Multi-way system

In each condition, the frequency and phase response of the target headphone was measured, and a linearphase FIR filter was calculated to match the source headphone’s response. The filters are designed to correct only the magnitude of the frequency response, without

phase correction. In this experiment, this provides the advantage that the influences of physical variables for instance total harmonic distortion (THD), phase linearity and impulse rise time can be evaluated in detail. In addition, limiting the equalization to the magnitude response ensures the transferability of the results to the live context. The impulse response of the four filters used is given in Figure 2. The same filters are used for the left and right side due to the minor deviations.

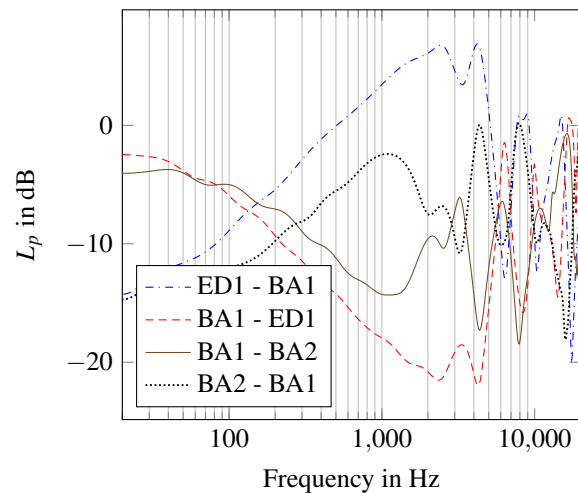


Fig. 2: Magnitude responses of the FIR filters used for equalizing the tested headphones. The frequency domain characteristics illustrate the adjustments applied to each transformation condition (ED1 to BA1, BA1 to ED1, etc.).

2.2 Participants and Pre-Test Survey

A total of 18 trained listeners aged between of 18 and 34 years participated in the study. All participants had prior experience in critical listening, either through professional audio work or academical background, ensuring a high level of perceptual sensitivity. For an in-detail assessment of individual expertise, participants completed a pre-test survey designed to assess their demographic background and the *Perceptual Abilities* subscale of the Gold-MSI questionnaire [8], along with three additional questions on previous musical experience [9] to ensure the expertise of the sample. The surveys were administered on paper immediately before the listening test.

2.3 Listening Test Procedure

The listening test was designed to evaluate perceptual differences across the equalized headphone conditions. In each condition, each participant listened to three 30-second audio excerpts given in Table 3, carefully selected to emphasize specific perceptual attributes relevant for the study. The headphone conditions were presented in a random order. The selected stimuli were chosen based on their acoustic properties and their ability to highlight differences in localizability, transient fidelity, and low frequency impulse fidelity. After each excerpt the participants answered three specific questions regarding one dependent variable.

Table 3: Songs used for the listening test and their underlying idea

Artist	Song	Dependent variable
Yosi Horikawa	Bubbles	Localizability
Rage Against the Machine	Vietnow	Transient fidelity
Massive Attack	Angel	LF Impulse fidelity

The first excerpt, was used to assess localizability. This track contains spatially distinct sound elements, including percussive effects and panned textures, which are distributed across the stereo field. The accurate reproduction of these spatial cues depends on the headphone's ability to maintain interaural level differences (ILDs) and interaural time differences (ITDs) without introducing spectral distortions or phase mismatches. Inaccurate localization may result in a perceived narrowing of the stereo image or unintended shifts in the spatial positioning of individual elements. By evaluating how well each equalization condition preserves spatial positioning, this test provides insight into whether different driver types and multi-way crossover designs affect the accuracy of binaural spatial reproduction.

The second excerpt, was chosen for its ability to test transient fidelity. The track features sharp snare drum hits, aggressive electric guitar attacks, and percussive elements that require precise transient reproduction. High transient fidelity ensures that fast attacks are rendered with sharp definition, without excessive smearing or softening of dynamic elements. ED drivers typically have a larger diaphragm with more inertia, potentially affecting transient sharpness, while BA drivers are often designed for fast transient behavior. This test eval-

uated whether these differences remain perceptually distinguishable after equalization, particularly in scenarios where transient details are critical for perceived clarity and articulation.

The third excerpt was used to evaluate the low frequency impulse fidelity. This track features deep, powerful bass hits and sudden dynamic bursts, which provide an ideal stimulus for assessing how well a headphone reproduces rapid changes in sound energy. Impulse fidelity is closely related to a system's ability to handle dynamic shifts without introducing unwanted artifacts such as ringing, overhang, or distortion. In multi-way BA systems with crossovers, phase interactions between drivers may alter the perceived impact of such impulses. By analyzing listener ratings in this condition, the test helps determine whether crossover-based multi-way systems introduce perceptible artifacts in dynamic accuracy, particularly when compared to single-way BA or ED systems. This variable is also crucial due to the differences in the size of the BA and ED drivers mean that the reproduction of low frequencies may work more effectively on the ED system.

Liking was also determined for each configuration by asking three further questions about whether the participants enjoyed listening with the headphones and if they would recommend them based on their sound reproduction.

2.4 Response Collection and Evaluation Metrics

After listening to each excerpt in each condition, participants performed their subjective rating for each of the three items that were subsequently averaged to one parameter. All responses are recorded via a digital survey interface where participants rated each condition using a continuous sliding scale from 1 to 5. The Cronbachs α values, given in Table 4 indicate a high inter-item correlation for the four dependent variables. Due to a ambiguity in one of the items regarding the preference, only two of them were considered for the following analyses.

In order to relate the subjective evaluations to physical measured variables, the phase response and the impulse rise time as well as the THD of the 2nd to 9th harmonic distortion were recorded in addition to the amplitude response. In order to express the phase response and THD as a ratio, they were averaged over the frequency. For the phase, this means that the root mean square of

Table 4: Reliability of the measured constructs

Attribute	Item	Cronbach's α
Localizability	1-3	0.838
Transient fidelity	4-6	0.786
LF Impulse fidelity	7-9	0.755
Preference	10-11*	0.899

*without item 12, otherwise $\alpha = 0.595$

the deviation from linearity was calculated and in the case of THD, this means that the mean value of the distortion was calculated as a percentage. The impulse rise time is assumed to be the time that elapses between a certain impulse signal and the actual reproduction.

3 Results

The absolute responses from the response sliders are given in Figure 3 and provide a direct evaluation of perceptual differences between the headphone conditions. The mean absolute ratings for each perceptual attribute—localizability, transient fidelity, low-frequency impulse fidelity, and preference—revealed systematic differences across conditions.

3.1 Evaluation of the Slider Responses

In terms of localizability, the systems in which the physical model contains a BA driver were rated higher than those with the ED driver. The localizability is significantly influenced by the impulse velocity which is significantly higher with BA drivers due to their design. This effect also applies to the transient fidelity. The evaluation of the low frequency impulse fidelity shows that the system with only one BA driver received a low rating. This might be due to the fact that the low frequencies could not be reproduced well because of the limited bandwidth of BA drivers. However, it can be observed that because of the minimal THD of the system, it is possible to improve the rating by equalizing the BA1 to ED1 or BA2, whose frequency responses are higher in the low frequency range. The equalized BA1 is even rated higher than the unequalized BA2 and ED1 systems, as the THD is significantly lower and the questionnaire also takes into account the clarity of the bass reproduction.

To assess the impact of different headphone architectures on the accuracy of VLTs, both objective acoustic

measurements and subjective evaluations were conducted. The subjective responses were analyzed using a Δ -score, which quantifies the perceptual deviation between the virtually equalized headphone and the corresponding target headphone as given in equation 1.

Here, $\overline{SR}$ represents the mean subjective rating under the virtual equalization condition, while $\overline{SR}_{ref}$ corresponds to the rating for the reference (unequalized) condition. The resulting Δ -score quantifies the deviation introduced by the virtualization process.

$$\Delta\text{-score} = \overline{SR} - \overline{SR}_{ref} \quad (1)$$

The Δ -score was computed for each of the four perceptual attributes: localizability, transient fidelity, low-frequency impulse fidelity, and overall preference. By comparing the Δ -scores given in Figure 4 across different equalization conditions, systematic differences between the tested headphones were identified. It can also be observed that the maximum deviation was around 10 %, which suggests that the virtualization was basically successful, but was still too inaccurate for the evaluation of high-end headphones.

3.2 Δ -Score Analysis Across Conditions

The statistical analysis, given in Table 7, revealed significant differences in Δ -scores across the tested conditions, indicating that some equalizations led to larger perceptual discrepancies than others.

In the ED1BA1 condition, where the ED1 was equalized to match the single BA headphone, the Δ -scores were particularly low for localizability (-0.56) and transient fidelity (-0.55), indicating a greater perceptual deviation between the virtualized and target response (see Figure 4). This finding suggests that equalizing an ED driver to a BA driver does not fully compensate for perceptual differences introduced by inherent differences in impulse response behavior and THD (see Table 5).

Conversely, in the BA1ED1 condition, where the single BA driver was equalized to match the ED headphone, the Δ -scores were generally higher. This finding indicates that the ED1 reproduced on the BA1 receives more positive rating than its original state. The most notable improvement was observed in low frequency impulse fidelity, likely due to the intrinsically faster impulse response of BA drivers. Localizability and

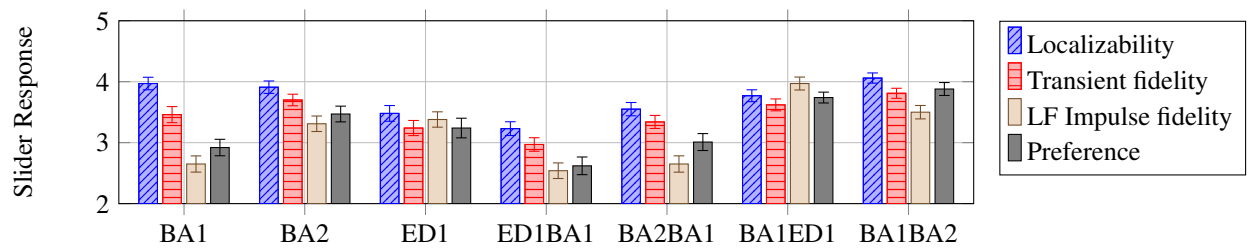


Fig. 3: Absolute slider response of the dependent variables with regard to the headphone system

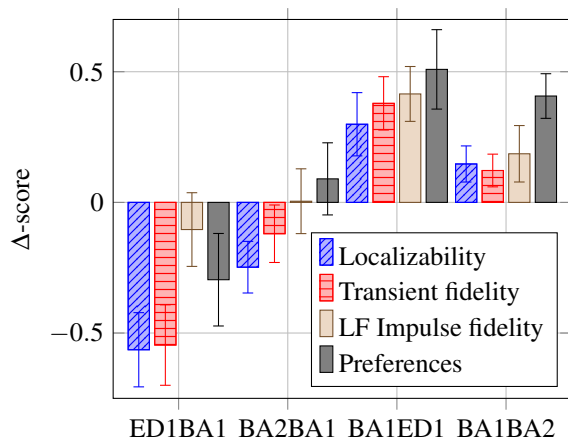


Fig. 4: Δ SR for the dependent variables over the equalization configurations

transient fidelity scores showed noticeable deviations, implying that other factors, such as phase response differences and THD, played a role in spatial perception mismatches.

Table 5: Measurement results of the ED1 and BA1

Headphone	$\overline{\Delta\text{Phase}_{\text{RMS}}}$	$\overline{\text{THD}}$	Rise time
ED1	124.1°	16 %	63.8 μs
BA1	87.6°	2.6 %	38.8 μs

The BA1BA2 condition, which examined the impact of a multi-way BA system by equalizing the single BA driver to the dual BA system, produced moderate Δ -scores across all attributes. While no extreme deviations were observed, the data suggest that the presence of a crossover in the BA2 system introduced additional perceptual differences. This effect was most prominent in preference (+0.41), indicating that multi-way BA configurations introduce subtle perceptual changes,

which are not fully compensated for by equalization alone. This can be attributable to the physical differences displayed in Table 6.

In the BA2BA1 condition, where the two-way BA system was equalized to match the single BA driver, the Δ -scores were also moderate. Notably, the localizability was rated lower (-0.25) suggesting that reducing a multi-way system with a higher phase shift due to the crossover to a single BA driver results in noticeable perceptual differences. This finding might be related to limitations in accurately replicating the phase response and transient characteristics of a system with separate drivers for different frequency ranges.

Table 6: Measurement results of the BA1 and BA2

Headphone	$\overline{\Delta\text{Phase}_{\text{RMS}}}$	$\overline{\text{THD}}$	Rise time
BA2	111.6°	13 %	31.4 μs
BA1	87.6°	2.6 %	38.8 μs

3.3 Statistical Evaluation of Condition Effects

To quantitatively assess the impact of headphone type on perceptual accuracy, a two-factorial repeated-measures ANOVA was conducted.

Table 7: Two-factorial RE-ANOVA

	SS	df	MS	F	p
Headphones	23.84	3	7.95	4.11	0.011
Residuals	98.62	51	1.934		
Attribute	3.40	3	1.134	2.18	0.101
Residuals	26.47	51	0.52		
Headphones *	1.79	9	0.20	0.51	0.867
Attribute					
Residuals	59.86	153	0.391		

The results show a significant main effect of headphone condition ($F(3,51) = 4.11$, $p = 0.011$), confirming that the choice of reference and target headphone had a systematic impact on the Δ -scores. However, there was no significant main effect of perceptual attribute ($F(3,51) = 2.184$, $p = 0.101$), though the result may indicate a statistical trend, suggesting that some perceptual attributes might have been affected more strongly than others across conditions. Additionally, no significant interaction between headphone condition and perceptual attribute could be observed ($F(9,153) = 0.509$, $p = 0.867$), suggesting that the overall perceptual mismatch is consistent across different listening dimensions rather than being tied to a specific attribute.

To further investigate the significant main effect of the equalization condition found in the repeated-measures ANOVA, a post-hoc analysis for the Δ -score was conducted using the Holm-Bonferroni correction to account for multiple comparisons. The results revealed that the BA1ED1 condition exhibited significantly higher Δ -scores compared to the ED1BA1 condition ($p_{HB} = 0.030$), indicating that equalizing the ED headphone to the single-way BA system did not fully compensate for perceptual differences.

Additionally, the BA1BA2 condition shows a nearly significant difference to the BA2BA1 condition ($p_{HB} = 0.052$), suggesting that phase interactions or crossover effects in the multi-way BA system may have influenced perception despite equalization. No other pairwise comparisons reaches statistical significance after correction, supporting the notion that driver topology plays a measurable role in the virtualization accuracy of BA-based IEMs. A detailed overview of the post hoc test results is provided in Table 8.

Table 8: Post-hoc test for Δ -scores across headphone models using Holm-Bonferroni correction

Modell	Modell	MD	SE	df	t	p_{HB}
ED1BA1	BA1ED1	-0.76	0.32	17	-2.36	0.030
	BA1BA2	-0.59	0.26	17	-2.28	0.036
	BA2BA1	-0.31	0.15	17	-2.08	0.053
BA1ED1	BA1BA2	0.17	0.22	17	0.77	0.454
	BA2BA1	0.45	0.21	17	2.10	0.052
BA1BA2	BA2BA1	0.28	0.19	17	1.45	0.164

These statistical findings reinforce the observation that the primary factor influencing virtual equalization accuracy is the structural difference between headphone

architectures rather than specific perceptual dimensions being disproportionately affected. The lack of an interaction effect implies that localizability, transient fidelity, and low frequency impulse fidelity all respond in a similar manner to discrepancies introduced by driver type and equalization direction.

3.4 Effects of Technical Parameters on Perceptual Scores

To better understand the underlying causes of these perceptual differences, the Δ -scores are correlated with key technical parameters extracted from the objective measurements. The most influential factors identified are the THD, rise time, and phase response consistency.

THD and Perceptual Deviation The THD is found to have a strong correlation with perceived transient fidelity. Headphones with higher THD values exhibited significantly lower Δ -scores for the localizability and transient fidelity, as given in Figure 5, suggesting that distortion interferes with the clarity of transient events. This effect was most pronounced when THD levels exceeded 16 %, at which point participants consistently reported reduced sharpness and precision.

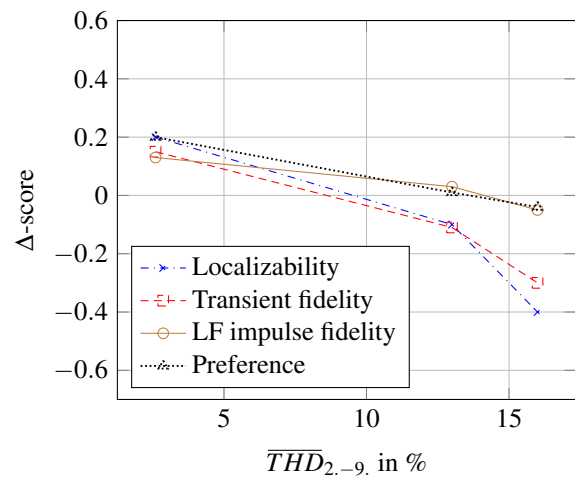


Fig. 5: The influence of the Δ -score on the dependent variables over the measured THD

This finding highlights the necessity of minimizing THD deviations to ensure that virtual equalization processes produce accurate perceptual matches. Even if the amplitude response is successfully equalized, excess distortion can introduce perceptual discrepancies that degrade the validity of VLTs.

Impulse Rise Time and Transient Accuracy The impulse response analysis given in Figure 6, revealed that headphones with faster impulse rise times were consistently rated higher regarding all four dependent variables. In detail, the ED driver exhibited a more extended impulse response compared to the balanced armature models, which was expected given the higher moving mass and greater inertia of a traditional voice coil. In contrast, the single BA driver demonstrated a significantly faster rise time, aligning with its perceptual advantage in all Δ -scores, especially regarding the preference. However, the dual BA system exhibited slightly slower impulse response times than the single BA system, possibly due to the fact that the BA2 contains a larger BA driver than the BA1 which could lead to a higher moment of inertia.

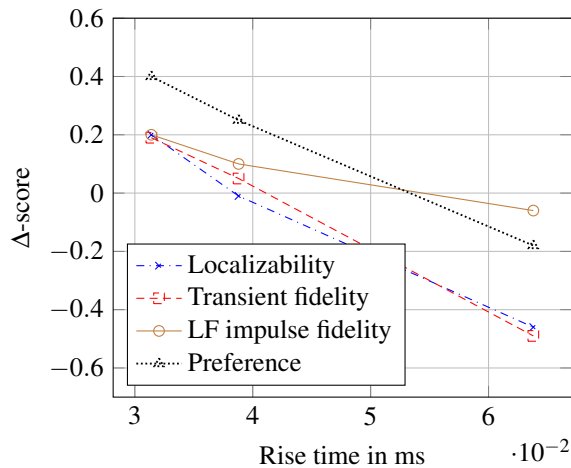


Fig. 6: The influence of the Δ -score on the dependent variables over the impulse rise time

The data in Figure 6 confirm that the impulse rise time is an important determinant of the accuracy in virtual listening scenarios. This has important implications for the design of reference headphones used in VLTs, as the impulse characteristics must be carefully controlled to maintain perceptual neutrality.

Phase Response and Localizability The phase response analysis (see Figure 7) indicated that phase nonlinearity had a measurable effect on the perceptual attributes, especially on localizability. In particular, BA1 and BA2 exhibited greater phase deviations at crossover frequencies, leading to small but noticeable differences in perceived spatial positioning. While

these differences were not as extreme as those caused by distortion or impulse response mismatches, they suggest that multi-way BA systems require precise phase alignment to avoid localization artifacts.

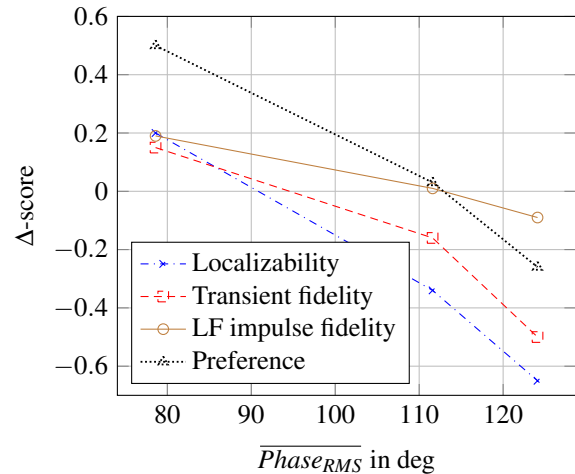


Fig. 7: The influence of the Δ -score on the dependent variables over the measured phase distortion

4 Discussion

The results demonstrate that, while reciprocal equalization can mitigate some differences between driver types, certain inherent acoustic characteristics of each driver topology remain perceptually relevant, leading to systematic deviations in the virtualized perception. The degree of deviation varies depending on the equalization direction, suggesting an asymmetry in how well certain headphone types can be transformed into others through equalization alone. Headphones with BA drivers were mostly rated higher in absolute slider response values than the ED headphones, which means that when ED drivers are mapped onto BA drivers, they were rated better than they actually are. Using ED drivers to simulate BA drivers led to the inverse effect (i.e. they were rated worse than the actual device).

4.1 Limitations of Reciprocal Equalization for Virtual Listening Tests

The observed Δ -score differences across conditions suggest that equalization of the frequency response amplitude alone cannot fully eliminate perceptual discrepancies between different driver types. This effect was particularly pronounced in the ED1BA1 and BA1ED1

conditions, where the largest deviations were observed. The difficulty in transforming an ED headphone into a BA-based system (or vice versa) is likely attributable to inherent differences in non-linear distortions, impulse rise time, and phase behavior. Unlike frequency response mismatches, which can be compensated for with FIR filters, impulse response discrepancies, particularly from ED to BA systems, cannot be corrected due to electro-mechanical constraints.

4.2 Influence of Multi-way BA Systems on Virtualization Accuracy

The results further highlight that multi-way BA systems introduce additional complexities in virtual listening tests. The BA1BA2 and BA2BA1 conditions exhibited moderate Δ -score differences, particularly in localizability and low-frequency impulse fidelity, indicating that a multi-driver configuration alters the spatial and temporal perception of sound in ways that are not easily replicated by a single-driver system. One likely explanation is that the crossover network in the BA2 system introduces frequency-dependent phase shifts, which, while perceptually subtle, could contribute to a mismatch in spatial imaging and transient response when equalized to a single BA driver.

This finding has important implications for the design of reference headphones in VLTs. That is, if a multi-way BA headphone cannot be reliably virtualized onto a single BA driver, then the choice of reference model becomes critical. Here the results suggest that a multi-way BA system may be better suited as a reference for other BA-based headphones. Future research should investigate whether alternative reference headphones with optimized phase coherence improve the virtualization accuracy of multi-way BA systems.

4.3 Perceptual Asymmetry in Equalization Directions

An other crucial observation is that the direction of equalization matters significantly. In detail, the BA1ED1 condition yielded better perceptual alignment than ED1BA1, suggesting that BA drivers can more accurately approximate the sound of ED drivers than vice versa. This asymmetry is likely due to the more precise transient response and reduced moving mass of BA drivers, which makes it easier to mimic the typically slower response of an ED system. In contrast, attempting to simulate the characteristics of a BA driver

with an ED transducer introduces fundamental limitations, as an ED driver may not be able to accurately reproduce the sharp transient characteristics of a BA system.

Similar asymmetries were observed in the BA1BA2 versus BA2BA1 conditions, where equalizing a multi-way BA system to a single BA driver resulted in more pronounced perceptual differences than the reverse transformation. This suggests that multi-way BA systems exhibit transmission characteristics that a single BA driver fails to replicate on a perceptually relevant level due to differences in cross-driver phase alignment, crossover implementation, and spatial radiation patterns.

This finding suggests that when designing VLT reference headphones, it may be beneficial to use a headphone with "faster" transient characteristics to maximize equalization accuracy. In practice, this means that a BA-based reference headphone may be more suitable for testing ED headphones than the other way around.

4.4 Implications for Future Virtual Listening Test Methodologies

Based on these findings, several considerations emerge for improving VLT methodologies for in-ear headphones.

For instance, the choice of the reference headphone was shown to significantly influence virtualization accuracy. To avoid imprecisions, it should contain the same type of driver than the headphones to be evaluated. Otherwise the perceptual subtleties can not be taken into account.

Moreover, while amplitude equalization alone provided reasonable results, phase mismatches, particularly in multi-way BA systems, remained a source of perceptual deviation. Future studies should evaluate whether minimum-phase equalization can be used in live evaluation settings. In case of comparing physical available headphones with prepared stimuli, one should consider binaural recordings.

Further, the data suggest that BA drivers are more effective at approximating ED headphones than vice versa when considering binaural recordings. This is due to the fact that one can decrease, but not increase the performance of the system. This suggests that, where possible, using BA-based reference headphones may improve equalization accuracy when testing different driver types.

5 Conclusion

This study investigated the suitability of VLTs for in-ear headphones with different driver topologies, focusing on ED and BA drivers, including both single-driver and multi-way BA systems. Using reciprocal equalization with FIR filters, three headphone models were perceptually compared across four key attributes: localizability, transient fidelity, low-frequency impulse fidelity, and overall preference. The goal was to determine whether a reference in-ear headphone could effectively reproduce the perceptual characteristics of different driver types.

The results revealed that, while frequency response equalization reduced spectral mismatches, residual differences in transient response, impulse fidelity, and phase behavior remained perceptually relevant. The asymmetry in equalization effectiveness suggests that BA drivers can better approximate ED headphones than the reverse, likely due to their faster transient response and lower moving mass. Moreover, multi-way BA systems introduced additional challenges, as the presence of crossover networks and phase shifts led to larger deviations in virtualization accuracy.

These findings highlight considerations for designing VLT methodologies for in-ear headphones, including the choice of a reference headphone and the role of phase and THD correction. The study concludes that, while VLTs are feasible for in-ear headphones, their effectiveness depends on minimizing driver-specific perceptual artifacts and optimizing equalization strategies (i.e. using binaural recordings from the headphones under test if available).

Future studies should investigate the combination of amplitude and phase equalization and employ additional stimuli that cover a broader range of acoustic properties. Furthermore, the development of a standardized training protocol for listeners could help reduce inter-subject variability in perceptual ratings.

References

- [1] Briolle, F. and Voinier, T., “Transfer function and subjective quality of headphones: Part 2, subjective quality evaluations,” in *Audio Engineering Society Conference: 11th International Conference: Test & Measurement*, Audio Engineering Society, 1992.
- [2] Hirvonen, T., Vaalgamaa, M., Backman, J., and Karjalainen, M., “Listening test methodology for headphone evaluation,” in *Audio Engineering Society Convention 114*, Audio Engineering Society, 2003.
- [3] Olive, S. E., Welti, T., and McMullin, E., “A virtual headphone listening test methodology,” in *Audio Engineering Society Conference: 51st International Conference: Loudspeakers and Headphones*, Audio Engineering Society, 2013.
- [4] Welti, T., Olive, S., and Khonsaripour, O., “Validation of a virtual in-ear headphone listening test method,” in *Audio Engineering Society Convention 141*, Audio Engineering Society, 2016.
- [5] Li, G., Zheng, C., Li, X., Yu, T., Bleeck, S., and Sang, J., “Evaluation of headphone phase equalization on sound reproduction,” *Applied Acoustics*, 156, pp. 208–216, 2019.
- [6] Suzuki, H., Morita, S., and Shindo, T., “On the perception of phase distortion,” *Journal of the Audio Engineering Society*, 28(9), pp. 570–574, 1980.
- [7] Fielder, L., “Perceptual Assessment of Headphone Distortion,” in *Audio Engineering Society Convention 143*, Audio Engineering Society, 2017.
- [8] Schaal, N. K., Bauer, A.-K. R., and Müllensiefen, D., “Der Gold-MSI: replikation und validierung eines fragebogeninstrumentes zur messung musikalischer erfahrung anhand einer deutschen stichprobe,” *Musicae Scientiae*, 18(4), pp. 423–447, 2014.
- [9] von Berg, M., Steffens, J., Weinzierl, S., and Müllensiefen, D., “Assessing room acoustic listening expertise,” *The Journal of the Acoustical Society of America*, 150(4), pp. 2539–2548, 2021.