Audio Engineering Society

Convention Paper 10193

Presented at the AES 157th Convention

2024 October 8-10, New York, NY, USA

This paper was peer-reviewed as a complete manuscript for presentation at this convention. 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.

A comparison of in-ear headphone target curves for the Brüel & Kjær Head & Torso Simulator Type 5128

Sean E. Olive

Harman International Industries

8500 Balboa Blvd. Northridge, California, USA

Correspondence should be addressed to Sean Olive (sean.olive@harman.com)

ABSTRACT

Controlled listening tests were conducted on five different in-ear (IE) headphone target curves measured on the latest ITU-T Type 4.3 ear simulator (e.g. Bruel & Kjaer Head & Torso Simulator Type 5128). A total of 32 listeners rated each target on a 100-point scale based on preference for three different music programs with two observations each. When averaged across all listeners, two target curves were found to be equally preferred over the other choices. Agglomerative hierarchical clustering analysis further revealed two classes of listeners based on dissimilarities in their preferred target curves. Class One (72% of listeners) preferred the top two rated targets. Class two (28% of listeners) preferred targets with 2 dB less bass and 2 dB more treble than the target curves preferred by Class 1. Among the demographic factors examined, age was the best predictor of membership in each class.

1 Introduction

The ideal preferred frequency response or target curve of a headphone based on its perceived sound quality has been the focus of numerous research studies. Headphone target curves have been proposed for around-ear (AE) and on-ear (OE) headphones [1-5], and more recently for in-ear (IE) types [6-8]. In all cases, the proposed target was preferred to the diffuse-field (DF) target curve recommended in current in international standards [9-11].

The preferred AE/OE target curve discussed in [4] approximates the steady-state response of a neutral loudspeaker (i.e. anechoically flat on-axis with smooth sound power) measured in a semi-reflective. Popularly known as the "Harman Target Curve", it has received broad acceptance among headphone enthusiasts, manufacturers and testing organizations as a benchmark when testing headphones.

There has been less research focus and consensus on what the ideal target curve should be for IE headphones, and many research questions remain unanswered. The current IE target response proposed in [6,7] has about 3.5 dB more bass and slightly more treble relative to the AE/OE target curve based on listeners' adjustments in controlled tests. Recent evidence suggests these bass and treble difference between IE and AE headphones may be related to interactions between the acoustic impedances of the ear simulator and the acoustic load presented by the

headphone [12]. Miller and Downey focused on the IE target curve above 10 kHz and found listeners prefer additional gain that varied depending on their hearing loss [8]. Silver argues that the DF response of current headphone test fixtures contain more high frequency than typically measured on humans; therefore, this should be accounted for in the target responses [13]. Finally, there is evidence that a single target IE target curve cannot satisfy everyone due to individual differences related to age, hearing loss, listening experience and gender that affect our preferred levels of bass and treble [14]. Also, individual differences in the size and geometry of our ear canals affect the magnitude and phase response at the ear drum, which if compensated for can potentially improve the sound quality [15,16].

Finally, there are no published studies of a preferred IE headphone target curve based on the latest ITU-T Type 4.3 headphone test fixture like the B&K 5128 HATS. It seems unlikely that the current target curves specified on Type 3.3 (popularly known as "711") ear simulators can be applied to headphones measured on the Type 4.3 versions due to differences in their acoustic impedances [12].

This paper addresses some of the above questions. Sec. 2 reports differences in measurements of IE headphones made on Type 3.3 and 4.3 test fixtures and a method for translating the current Harman IE target curve in [6,7] to the Type 4.3. Sec. 3 describes the methodology for a listening experiment where this translated target curve is evaluated against four other targets. The results are reported and discussed in Secs. 4. and 5 with conclusions reported in Sec. 6.

2 Measurements on Type 3.3 Versus Type 4.3 Ear Simulators

Twenty different models of IE headphones were measured on both Type 3.3 (GRAS RA045) and Type 4.3 (B&K Type 5128-B) test fixtures. They included 18 models with either single or dual dynamic drivers and two hybrid models with dynamic drivers and balanced armatures. Each measurement was repeated five times to account for small variations in insertion depth of the headphone into the ear simulator.

Fig. 1 shows the difference in magnitude between Type 3.3 and 4.3 for each model based on the average of five measurements. There are significant differences between the two ear simulators that vary up to 10 dB below 100 Hz and 30 dB above 6 kHz depending on the model of headphone. The magnitude and frequency dependency of these curves

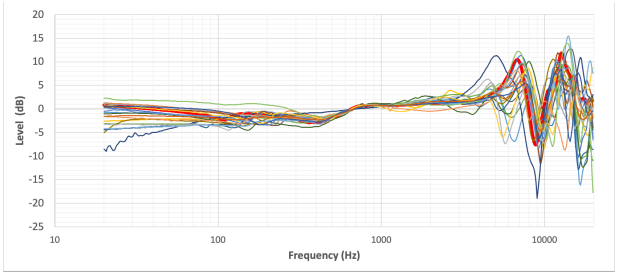


Figure 1 The difference in frequency response shown for 20 models of IE headphones measured on Types 3.3 versus Type 4.3 ear simulators

can be traced to differences between the ear simulators in their volume and transfer impedance which interact with the impedance of the headphone and its insertion depth. A single correction curve cannot account for the measurement differences between Type 3.3 and 4.3 test fixtures because the error changes with headphone model.

Fig. 2 shows the average difference between all 20 headphones. On average, these headphones have 2 dB less bass below 200 Hz when measured on the newer Type 4.3 and 2 dB more treble above 2 kHz with some large peaks at 7kHz and 14kHz and a dip at 9.5 kHz.

However, it is possible to derive a valid correction for the measurement differences between test fixtures for a single headphone. The same headphone (Sennheiser Momentum) used in the previous IE target curve study [6,7] was measured by two different people for the sake of reproducibility (see Fig. 3).

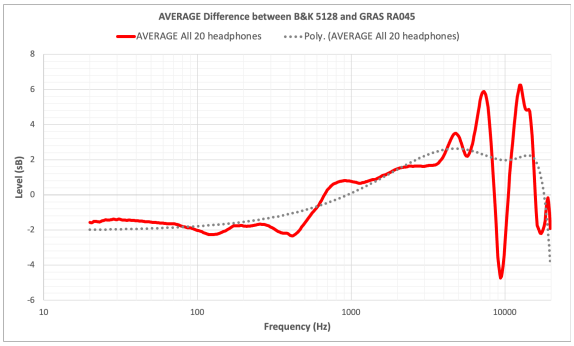


Figure 2 The average difference between 20 models of IE headphones measured on a Type 3.3 versus 4.3 ear simulator (solid curve) with a 2nd order polynomial curve (dotted) that best fits the average.

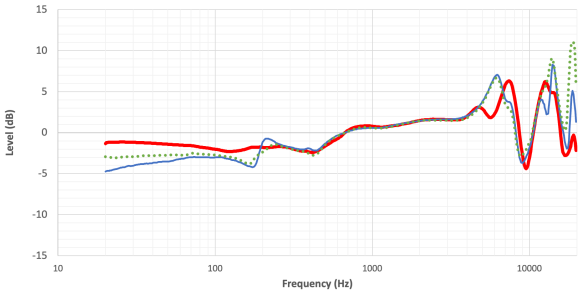


Figure 3 The average difference in magnitude response between Type 3.3 versus Type 4.3 fixtures for 20 models of IE headphones (red curve) and Sennheiser Momentum measured by two people (dotted curves).

The reproducibility is very good with only 1-2 dB differences below 100 Hz likely related to the seal. The measurement of the Sennheiser model on the Type 4.3 fixture has 3-4 dB less bass below 100 Hz, 1-5 dB more energy above 1 kHz, and is not so different from the average of the 20 headphones. This difference curve was the correction applied to the current Harman IE target curve to translate it from G.R.A.S RA045 (Type 3.3) to the B&K 5128 (Type 4.3). The sound quality validation of this target curve and four other targets is presented in the next section.

3 Listening Experiment

In this section, the setup and methodology of a listening experiment for five different IE target curves are described.

3.1 Selection of IE Target Curves

Five different IE target curves plotted in Fig. 4 were compared in controlled listening tests measured on a Type 4.3 ear simulator. Each target is discussed below.

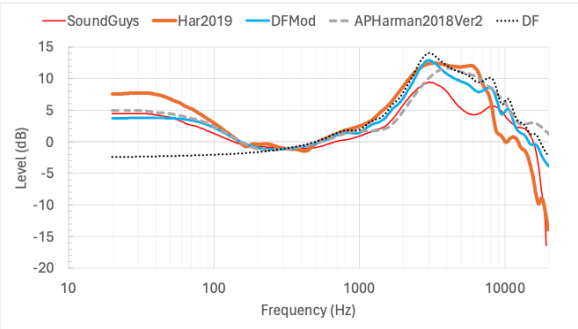


Figure 4 The five IE headphone target curves evaluated in this study.

APHarman2018Ver2 is an AE target curve that received the highest overall ratings among 21 different AE headphones evaluated by 30 Danish and 26 Japanese consumers [5]. It represents an approximation of the 2018 AE Harman Target curve in [4] translated from a GRAS45CA test fixture to the B&K 5128. While this target was specified for AE type headphones, it is included here because it represents a benchmark from a well-controlled study.

DF is the diffuse-field response of the B&K 5128 HATS provided by B&K. It approximates the current DF headphone calibration recommended by international standards for best sound quality.

DFMod is a modified version of DF with an added bass shelf (6.4 dB at 125 Hz) and high shelf cut (-1.4 dB at 2.5 kHz). It is a close approximation to the so-called Harman AE Target [4].

SoundGuys is the current target curve or both IE and AE/OE type headphones used by the headphone test organization SoundGuys [17]. The target is based on an average response of a compendium of popular headphones from well-known brands that include IE and AE/OE types both closed and open back designs. Soundguys argues that the target curve should be identical for all headphone types. In [5] it was among the top 10 highest rated targe curves.

Finally, *Har2019* is the current IE target curve translated from the Type 3.3 to Type 4.3 test fixtures as described in sec. 2. It is currently widely used within Harman to design IE headphones. For models aimed at younger demographics there is sometimes a slight modification made in the bass to satisfy their tastes based on research in [14].

3.1.1 Implementation of IE Target Curves

The five IE target curves were evaluated on the same Sennheiser Momentum IE headphone used in previous related studies [6,7] to allow more valid comparisons of the results between the two studies. Comparing different target curves through the same headphone facilitates immediate A/B comparisons between the targets addressing limitations in auditory memory, and controls nuisance variables like leakage and insertion depth. A small micro-electromechanical systems (MEMS) microphone was installed in the sound barrel of the headphone so that leakage could be detected and corrected before each listening test began.

The magnitude and phase responses of the headphone were measured on the B&K 5128 using the measurement software Room EQ Wizard (REW), and a minimum phase impulse response filter was designed to equalize the headphone to each target curve. The impulse response was then used to convolve the music programs and then adjusted for equal loudness according to ITU-R BS.1770-5 [18].

3.1.2 Program Selection

Three well-recorded music programs edited into short < 20 s loops were selected for these tests:

1. Jennifer Warnes, Bird on a Wire/ Blue Raincoat/ BMG Records, 1989, B00000DN6J.
2. Steely Dan, Cousin Dupree/ Two Against Nature/ Giant Records/WEA, 2000, B00004GOXS.
3. Steely Dan, Gas Lighting Abbey/Two Against Nature/ Giant Records/WEA, 2000, B00004GOXS.

The first two tracks have been identified in previous tests for producing the most discriminating and reliable programs for rating IE headphone target curves due to their dense and wide-band spectra [19].

3.1.3 Selection of Listeners

A total of 32 listeners participated, all Harman employees located in Novi, Michigan. Table 1 summarizes the age, gender, self-assessed hearing loss, and listening experience. Experienced listeners were defined as completing a formal training course and participating in formal listening tests. The age of listeners ranged from 21 to 50 years (M= 31.8, SD= 8.2).

Factor	Level	Count
Gender	Male	25
	Female	7
Listening Experience	Trained	27
	Moderate	4
	Some	1
Hearing Loss	None	28
	Mild	4

Table 1. The composition of the listening panel.

3.2 Test Procedure

A custom software program was written for Mac OS to administer the listening test and store the listener responses. The hardware included a MacBook Air M2 with a high audio-quality sound card (RME ADI-2/4 Pro SE). The program randomized the order of six trials and the target curves in each trial.

Each test was comprised of six trials (i.e. 3 programs x 2 Observations). In each trial the listener compared the five target curves and rated each on a 100-point preference scale. The order of the programs and target curves were randomized. Tied ratings were not allowed, and the listener was encouraged to use the entire range of the scale. On average, the test took 9.5 minutes to complete.

4 Results

4.1 Nonparametric Test

A Shapiro-Wilk normality test revealed the target curve preference ratings were not normally distributed ($p < 0.0001$). This was not unexpected given that listeners were encouraged to use the entire range of the scale, which can lead to asymmetrical or skewed distributions. The Q-Q normality plots indicated the DF target curve deviated furthest from normality, $W = 0.736$, $p < 0.000$, with the other target curves having W values between 0.942 to 0.965, where a value of 1 indicates a normal distribution.

Given the violation of normality, a non-parametric Freidman test based on ranks was performed at a significance level of 5%. The difference in ranks were significant, $Q(4) = 269.35$, $p < 0.0001$. A multiple pairwise comparison using Nemenyi's procedure found significant differences between four groups of target curves as summarized in Table 2. The rankings from most preferred to least preferred were SoundGuys, Har2019, DFMod, APHarman2018Ver2 and DF. Seven of the possible ten paired comparisons between target curves were statistically significant.

Target Curve	Frequency	Sum of Ranks	Mean of Ranks	Groups			
DF	192	275.0	1.432	A			
APHar2018Ver2	192	563.0	2.932		B		
DFMod	192	621.0	3.234		B	C	
Har2019	192	698.0	3.635			C	D
SoundGuys	192	734.0	3.766				D

Table 2. The statistics are shown for a non-parametric Friedman test for each target based on preference ranking. Groups indicate which pairs of targets are statistically significant from each other for a significance level of $p < 0.05$.

4.2 Repeated Measures ANOVA

Repeated measures ANOVA are generally known to be robust from violations of normality with minimal influence from Type I errors and the power of the F-statistic [20]. On this basis, a repeated measures ANOVA method explored potential effects and interactions between the preference ratings and the independent factors: Target Curve (5 levels), Program (3 levels), Gender (2 levels), Hearing Loss (2 levels) and Training (3 levels).

The main effect on preference was Target Curves, $F(4,188) = 26$, $p<0.0001$. There were also smaller but significant within subject effects for Gender, $F(1,185) = 12.43$, $p = 0.001$ and Hearing Loss, $F(1,185) = 7.08$, $p = 0.008$. Neither factor produced significant interactions between the Target Curves. No effects or interactions were produced by the remaining factors. The significant effects are summarized below.

4.3 Effect of Target Curve

A box plot in Fig. 5 displays for each target curve the median, mean (+), first and third quartiles, and the minimum and maximum values. The notch indicates the 95% confidence interval around the median.

The mean preference ratings in descending order were: SoundGuys (M = 65.97, SD = 29.6), Har2019 (M = 65.69, SD = 25.6), DFMod (M = 56.75, SD = 27.5), APHar2019Ver2 (M = 48.10, SD=29.6) and DF (M = 17.96, SD =2 5.1).

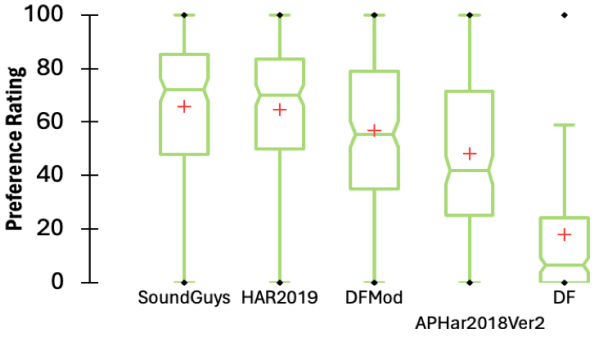


Figure 5 A box plot showing the median, mean, first and third quartiles and minimum and maximum values for the Target Curves.

It is noteworthy that rank order of the Target Curves is identical for both parametric and nonparametric tests, and similar conclusions can be drawn from both tests.

A Tukey post-hoc test indicates the differences in Target Curve preference rating means for all 10 paired comparisons are statistically significant with a confidence interval of 95%, except for the pair SoundGuys versus Har2019, $p = 0.991$.

4.4 The Effect of Hearing Loss

The mean target curve preference ratings and 95% confidence intervals are plotted in Fig. 6 separately for two groups of listeners who reported having no hearing loss (n =28) and mild hearing loss (n= 4).

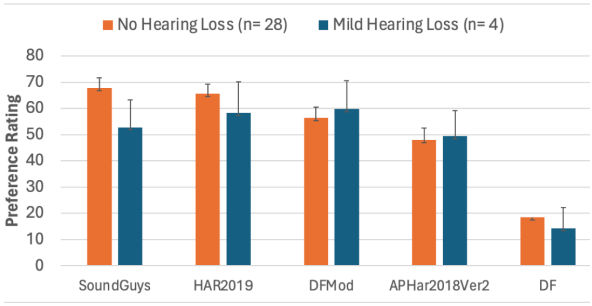


Figure 6 Mean preference ratings and 95% confidence intervals for target curves for listeners with reported no hearing loss and mild hearing loss.

Listeners with mild hearing loss tended to rate everything lower on the scale; on average 4.3 preference ratings lower than listeners with no reported hearing loss. Their larger confidence intervals can be explained by the smaller sample size (4 versus 28 listeners). The interaction between Target Curve and self-reported Hearing Loss was not statistically significant, F (4,740) =1.25, p=0.28.

4.5 Effect of Gender

The mean preference ratings and 95% confidence intervals are plotted for the interaction between Gender and Target Curve in Fig. 7. While the interaction between Gender and Target Curve was not significant, F (4,740) = 1.36, p = 0.245, the effect of Gender was, F (1,185) = 12.48, p = 0.001. While the rank order of the Target Curve was the same for males versus females, the latter gave relatively lower ratings than for the three lowest rated Target Curves. The confidence intervals are also larger for females which can be explained in part by the lower smaller sample size (i.e. 6 females versus 26 males).

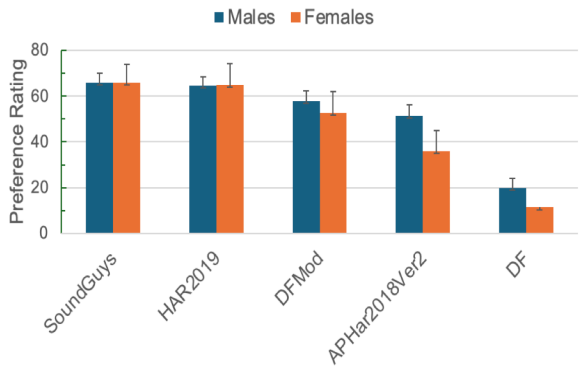


Figure 7 Mean preference rating and 95% confidence intervals for target curves shown for male and female listeners.

4.6 Segmentation of Listeners based on Target Curve Preferences

In a previous study on AE headphone preferences, Olive et al. reported there exists three different segments or classes of listeners based on dissimilarities in how they rated 31 different models of headphones [14]. Demographic factors such as age, listening experience, and gender explained membership in the classes and was related to differences in the preferred levels of bass and treble of the headphones. This section explores whether such relationships exist in this dataset.

An agglomerative hierarchical clustering (AHC) method using Ward's method generated the dendrogram plot in Fig. 8. The ordinate represents the magnitude of dissimilarity in target curve ratings and the abscissa represents each of the 32 listeners. There are two distinct classes of listeners, C1 and C2.

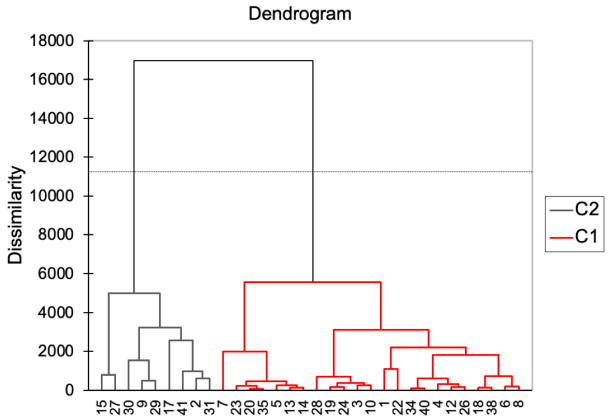


Figure 8 A dendrogram showing two main clusters C1 (right cluster) and C2 (left cluster) based on similarities in target curve preference ratings. The abscissa represents the different listeners.

C1 includes 72% (n= 23) of the listeners while C2 includes the remaining 28% (n = 9).

Fig. 9 show the mean preference ratings and 95% confidence intervals for each cluster. Listeners in C1 equally preferred SoundGuys and Har2019 overall whereas C2 listeners, gave the highest preference ratings to DFMod and preferred it slightly more than APHar2019Ver2. Both classes preferred DF the least although listeners in C1 rated

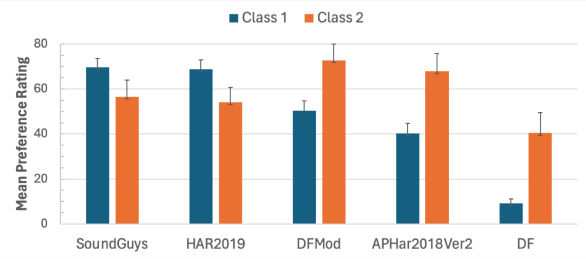


Figure 9 The mean preference ratings and 95% confidence intervals for the target curves in each class based on AHC cluster analysis.

it significantly lower than listeners in C2.

A discriminant analysis explored the relationship between class membership and the demographic factors: gender, listening experience, hearing loss and age. The analysis found none of the demographic factors were good predictors of class membership except age. Listeners in C1 were 10 years younger than C2 (M= 28.8 versus M= 39.33 years) and a unidimensional test of the means between classes were significant, $F(1,190) = 95.78$, $p < 0.0001$. Based on age, listeners could be predicted which class they belonged to with 94% accuracy.

5 Discussion

Sec. 2 provides confirmation that there are significant differences in the frequency response of headphones measured on Type 3.3 versus the newer Type 4.3 test fixtures, which more accurately reflect the acoustics of an average ear. Depending on the headphone model the magnitude of differences can range vary 10 dB at low frequencies to 30 dB at higher frequencies. Mansilla reported similar discrepancies using a balanced armature IE headphone (see Figs. 10 and 11 in [12]). Together these studies illustrate why target curves specified on Type 3.3 fixtures are not valid when applied to headphone measurements made on the newer Type 4.3 fixtures.

To the author's best knowledge, this is the first study to publish results from controlled listening tests comparing different IE target curves based on Type 4.3 measurements. It is reassuring that the Harman IE headphone target curve first reported in [6,7] still produces high preference ratings when translated from the Type 3.3 Type 4.3 test fixtures. The Har2019 target curve was equally preferred to the SoundGuys target curve, and both were preferred to the other three target curve tested.

The DF target curve was the least preferred target curve, which is consistent with previous studies [1,2,3,5]. It is baffling that the current recommended DF calibration of headphones specified in international standards [10,11] remains at odds with the scientific evidence that tells us there are more neutral and balanced options that listeners prefer.

5.1 Sound Profiles of Segmentation Classes

Segmentation analysis (see sec. 4.1) revealed that a single target curve cannot satisfy the tastes of all listeners. It identified two classes of listeners: Class 1 (72%) equally preferred the SoundGuys and Har2019 target curves and Class 2 (28%) preferred DFMod slightly to APHar2019Ver2.

Fig. 10 shows the spectral differences for the five targets relative to the average calculated from the five. The purpose here is to illustrate the spectral differences between the target curves and hopefully explain the different profiles or taste in sound for each class. The greatest spectral differences among the target curves occur below 200 Hz and above 1 kHz, which affects the perceived balance of bass and treble. Between 200 Hz and 1kHz the midrange is relative consistent among target curves.

Class 1 listeners preferred, Har2019 (red) and SoundGuys (black). Har2019 has 4 dB more bass than the average and about 3 dB more bass than SoundGuys. Above 1 kHz, there is a significant departure between the two targets; between 500 Hz and 5 kHz SoundGuys gradually drops 5 dB at 5 kHz before returning to the group's average level at 10kHz. In contrast, Har2019 rises 2 dB above the group average from 1kHz to 5 kHz before dropping off 4dB below at 10kHz.

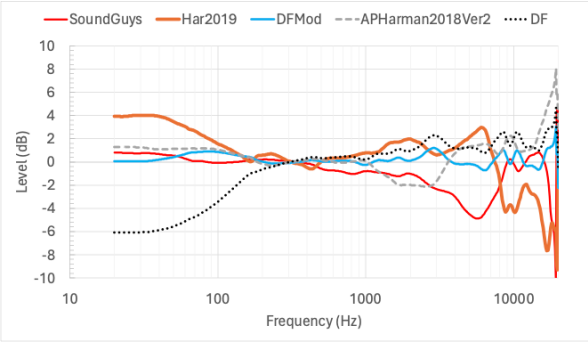


Figure 10 The spectral differences of each target curve relative to the average of all five.

In view of these large spectral differences, it is surprising that these two headphones received similarly high ratings from Class 1 listeners. One possible explanation there is a perceptual tradeoff between bass and treble levels whereby a reduction in bass may be compensated with an equivalent reduction in treble, and vice versa [21]. It seems that a constant delta of 8 dB between the bass and treble bands of the target curve can produce equivalent satisfaction and listener preferences.

Class 2 listeners preferred DFmod slightly more than APHar2018ver2. Looking at Fig. 10, DFmod comes closest to the average spectral balance of the five target curves deviating less than 1 dB up to 10kHz. APHar2019Ver2 has about 1 dB more bass than the average dropping 2 dB between 1-3kHz before rising 2 dB above the average.

Fig. 11 plots the average response of the top two rated target curves in Classes 1 and 2 with respect to the DF response of the B&K Type 5128 HATS, which was rated last in this study. The purpose is to show the contrast in target curve preferences between Class 1 and Class 2. Also, researchers can use these curves to equalize the DF response of their test fixtures for future target curve studies. The dotted red curve represents the difference in preferred spectral balance between Class 1 and 2, which can be summarized as follows: Class 1 listeners prefer about 2 dB more bass below 100 Hz, 1 dB less low treble between 3-6kHz, and about 2 dB less upper treble above 6kHz compared to Class 2 listeners.

The nature and magnitude of these differences are similar to the differences reported in [14] where three classes of listeners were found to have different preferences related to bass and treble levels. In that study, Class 1 (64% of listeners) preferred a target like Har2019, Class 2 (15%) preferred the same target with 4-6 dB more bass, and Class 3 (22%) preferred 2 dB less bass and 1 dB more treble. Class 3 included a disproportionate percentage of females and older listeners. In the current study, older listeners also tended to prefer target curves with 2 dB less bass and more treble than Class 1.

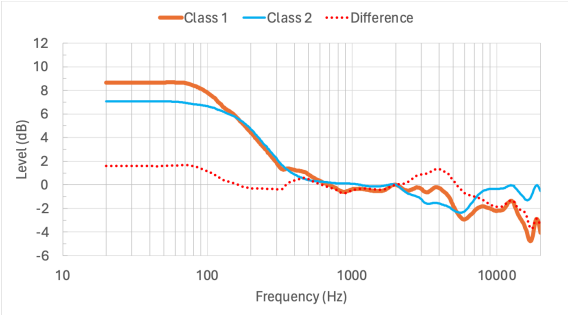


Figure 11 The relative differences of the average top two preferred target curves for Class 1 and 2 listeners relative to the DF of the B&K Type 5128 HATS.

Together these two studies suggest that at least two or three different segments of listeners exist that have different preferred target curves related to the relative balance of bass and treble. In the earlier study demographic factors such as age, listening experience and gender seemed to play a role whereas the current study, age was the main factor, likely because most listeners in this study were trained. For example, in a previous study (see Fig. 2 in [4]) trained listeners tended to prefer slightly less bass and treble in their headphones than untrained listeners. Given that this study included mostly trained listeners, the Har2019 target curve may be more preferred among untrained listeners than reported in this study.

Fortunately, the differences in listeners' target curve preferences here can be easily accommodated with a simple bass and treble filter if the headphone has a wide-band smooth response, free of resonances. For example, the DFMod target satisfies Class 2 tastes but Class 1's preferred sound profile met with 3 dB additional bass and 2 dB less treble. The Har2019 or the SoundGuys target curves are good default choices since they were preferred by 72% of the listeners in this study.

6 Conclusions

This study was motivated by the need to identify a preferred IE headphone target curve for the new Type 4.3 ear simulator (B&K Type 5128) based on controlled listening tests. To the author's best knowledge this is the first study to provide such data.

One conclusion from this study is that IE target curves defined on Type 3.3 test fixtures are not valid when applied to headphones measured on Type 4.3

ones. The differences in measurements between test fixtures are broadband and vary from 10 dB in the bass to more than 30 dB at higher frequencies depending on the headphone model (see sec. 2).

A second conclusion is that it is possible to translate a target curve based on a single headphone defined on Type 3.3 fixture to a Type 4.3, assuming the headphone is the same model used for the original target curve research. This is important since the target curve accounts for the acoustic interactions between the headphone impedance and the acoustics of individual ear canals during the listening tests. The assumption here is the more closely the ear simulator matches the average ear canal acoustics of the listeners, the better the correlation between headphone measurements and its perceived sound quality.

Using this method, a new target curve was defined for the Type 4.3 (B&K 5128) test fixture using an existing target curve defined on the older Type 3.3 (GRAS RA045) [6,7]. This new IE target curve (Har2019) was validated by evaluating it against four other target curves in controlled listening tests where 32 listeners gave comparative preference ratings on 100-point preference scale for three programs with two observations.

Nonparametric and parametric analysis of the results produced similar conclusions: overall, two target curves (Har2019 and SoundGuys) were preferred to the other three target curves. The least preferred target was the DF target curve.

Segmentation analysis revealed two classes of listeners based on different target curve preferences. Class 1 (72% of the listeners) equally preferred the top two rated targets. Class 2 (28% of listeners) preferred different target curves that had 2 dB less bass and 1-2 dB more treble than the targets preferred by Class 1. Discriminant analysis revealed Age to be the best predictor of class membership: Class 2 members on average (M= 39.33 years) were 10 years older than Class 1. Unfortunately, audiometry results were not available to confirm this, but age-related hearing loss may explain these different preferences among the two classes.

In summary, this paper provides more evidence that while a single target curve can satisfy most listeners, smaller segments of listeners exist whose tastes deviate from the average and may not be satisfied. This presents an opportunity for headphone manufactures to personalize the headphone to match the tastes and sound profile of the listener. The challenge is identifying those people and matching their headphone to their profile with a solution that is practical, effective and reliable.

7 Acknowledgement

The author thanks Harman International and his colleagues Jeff Baker, Etienne Rivet, Calum Smeaton, and Todd Welti for their support and feedback throughout this project. He also thanks the listeners in Novi, Michigan and Jonathan Pierce and Papa Adu-Boaheng for their assistance in recruiting listeners and administering the tests.

8 References

[1] Lorho, Gaëtan, “Subjective Evaluation of Headphone Target Frequency Responses’, 126th AES Convention, (May 2009).

[2] Fleischmann, F.; Silzle, A.; Plogsties, J., “Identification and Evaluation of Target Curves for Headphones”, presented at the 133rd AES Convention., (October 2012).

[3] Olive, S.; Welti, T.; McMullin, E., “Listener Preferences for Different Headphone Target Response Curves”, presented at the 134th AES Convention., (May 2013).

[4] Olive, S. E., “The Perception and Measurement of Headphone Sound Quality: What Do Listeners Prefer?” Acoustics Today, 18(1), p. 58, (2021), ISSN 15570215, doi:10.1121/at.2022.18.1.58.

[5] Ravizza, G., Villegas, J.; Volk, C.; Stegenborg-Andersen, T., “An Over-Ear Headphone Target Curve for Brüel & Kjær Head and Torso Simulator Type 5128 measurements “, presented at the 155th AES Convention, (October 2023).

[6] Olive, S.; Welti T.; Khonsaripour, O., “A Statistical Model that Predicts Listeners’ Preference Ratings of In-Ear Headphones: Part 1—Listening Test Results and Acoustic Measurements”, presented at the 143rd AES, (October 2017).

[7] Olive, S.; Welti T.; Khonsaripour, O., "A Statistical Model that Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 2—Development and Validation of the Model", presented at the 143rd AES Convention, (October 2017).

[8] Miller, T.; Downey, C., "Listener Preferences for High-Frequency Response of Insert Headphones," J. Audio Eng. Soc., vol. 71, no. 10, pp. 707–718, (2023 October). DOI: https://doi.org/10.17743/jaes.2022.0094

[9] International Electrotechnical Commission (IEC) (2009). Electroacoustics - Simulators of Human Head and Ear - Part 1: Ear Simulator for the Measurement of Supra-Aural and Circumaural Earphones. Report IEC 60318-1, IEC, Geneva, Switzerland.

[10] International Electrotechnical Commission (IEC) (2010). Sound System Equipment - Part 7: Headphones and Earphones. Report IEC 60268-7, IEC, Geneva, Switzerland.

[11] International Telecommunication Union Radiocommunication Assembly, (ITU-R) (1990). Determination of the Electroacoustical Properties of Studio Monitor Headphones. Report ITU BS.708, International Telecommunication Union, Geneva, Switzerland.

[12] Mansilla, J.," Headphone Measurements: The New Standard, Part 1," Head-Fi, https://www.head-fi.org/threads/headphone-measurements-the-new-standard-part-1.937301/ - post-15735920 (July 13,2020).

[13] Silver, G. "The shape of IEMS to Come", https://headphones.com/blogs/features/the-shape-of-iems-to-come, (October 4, 2023).

[14] Olive, S.; Welti T.; Khonsaripour, O., "Segmentation of Listeners Based on Their Preferred Sound Quality Profiles," presented at the 146th AES Convention, (March 2019).

[15] Bezzola, A., "In-Ear Headphones on Ear Canal Simulator vs Real Human Ear Geometries: Quantifying the Differences with Simulations," presented at the 155th AES Convention, New York City, USA, (October 2023).

[16] Bezzola, A., Celestinos, A., Blanes, E., "Personalized Equalization of Sound Pressure at Eardrum with Insert Buds," presented at the 156th AES Convention, Madrid, Spain (June 2024).

[17] Wykes, A. J. "The House Curve: How we make our target headphone responses," SoundGuys, https://www.soundguys.com/soundguys-house-curve-54290/ (August 21, 2021).

[18] Recommendation ITU-R BS.1770-5, *Algorithms to measure audio programme loudness and true-peak audio level*. (November 2023). https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-5-202311-I!!PDF-E.pdf

[19] Olive, S.; Welti T.; Khonsaripour, O., "The Influence of Program Material on Sound Quality Ratings of In-Ear Headphones," presented at the 142nd AES Convention, (May 2017).

[20] Blanca, M.; Arnau, J.; Garcia-Castro, F.J.; Bono R.," Alarcon R., "Non-normal data in Repeated Measures ANOVA: Impact on Type I Error and Power," Psicothema, 35(1), 21-29., (2023). https://doi.org/10.7334/psicothema2022.292

[21] Olive, S.; Welti T.; McMullin, E., "Listener Preferences for In-Room Loudspeaker and Headphone Target Responses, "presented at the 135th AES Convention, New York, USA, (October 2013).