Audio Engineering Society

Convention Paper 9919

Presented at the 144th Convention
2018 May 23–26, Milan, Italy

A Statistical Model that Predicts Listeners' Preference Ratings of Around-Ear and On-Ear Headphones

Sean E. Olive, Todd Welti, and Omid Khonsaripour

Harman International, 8500 Balboa Blvd., Northridge, CA, 91329

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

ABSTRACT

A controlled listening test was conducted on 31 different models of around-ear (AE) and on-ear (OE) headphones to determine listeners' sound quality preferences. 130 listeners both trained and untrained rated the headphones based on preference using a virtual headphone method wherein a single replicator headphone was equalized to match the magnitude and minimum phase responses of the different headphones. Listeners rated 8 headphones in each trial that included high (the new Harman AE/OE target curve) and low anchors. On average, both trained and untrained listeners preferred the high anchor to 31 other choices. Using machine learning a model was developed that predicts the listeners' headphone preference ratings using deviations in magnitude response from the Harman target curve.

1 Introduction

There is little consensus amongst headphone manufacturers on what the ideal headphone target frequency response curve should be for optimal sound quality. This was the conclusion of Breebaart [1] who measured the frequency responses of 283 headphones ranging in price from $4 to over $5000, including around-ear (AE), on-ear (OE), and in-ear (IE) types. He found little or no correlation between their price and their measured response when referenced to target curve shown to produce good sound [2]. We reached a similar conclusion in a recent study [3,4] where the correlation between listeners' headphone sound quality preferences and price was very low (r = 0.14).

Several researchers [5-7] have investigated the recommended diffuse-field headphone target curves in the current IEC, ITU, and EBU standards [8-10], and found that listeners prefer the sound quality of alternative target curves. The headphone industry and their standards organizations could benefit from more scientific research into what makes a headphone sound good and how to objectively define and measure it. That is the main focus of our research, and this current paper.

In a previous paper [4] we reported the results from controlled listening tests where listeners rated the sound quality of 30 different models of IE headphones according to preference. Included in each trial was a hidden high anchor (the Harman IE headphone target curve) that both trained and untrained listeners preferred overall. A statistical model based on how far the headphone deviated from the target response curve accurately predicted how listeners rated its overall sound quality [1].

[1] The Pearson correlation coefficient for the predicted versus observed headphone sound quality preference ratings was r = 0.91 with a mean square error of 5.5% ratings.

In this paper, the authors use a similar approach to test and validate new preferred target response curve optimized for AE-OE headphones. Controlled double blind listening tests were conducted on 31 different headphones from 18 different manufacturers using both trained and untrained listeners. Machine learning was then applied to the subjective and objective headphone measurements to model and predict how listeners would rate the sound quality of a headphone based on its measured magnitude response. In this way, the overall sound quality of the headphone can be determined from a simple acoustic measurement without the need to conduct expensive and time-consuming listening tests.

This paper is organized into six sections. Section 2 describes details on how the listening tests were conducted, with the results reported in section 3. The statistical model is developed in section 4, followed by a discussion and conclusions presented in sections 5 and 6.

2 Listening Tests

2.1 Headphone Selection

A total of 31 models of AE/OE headphones from 18 different manufacturers were selected for these tests (see appendix 1). The chosen headphones covered a broad price range from $60 to $4000 USD, and included both open and closed-back designs with either dynamic (n = 26) or magnetic planar (n= 5) type transducers. Ten of the 31 headphones were wireless (Bluetooth) and 5 models had Active Noise Cancellation (ANC).

2.2 Virtual Headphone Method

A virtual headphone listening test method was used to provide rapid multi-way comparisons among the different headphones in a controlled, repeatable and double blind manner. The method has been successfully used in previous papers for virtualizing both AE/OE headphones [7, 14] and IE headphones [3,4,11].

The accuracy of the virtualization method was validated by comparing listeners' sound quality ratings of actual AE/OE headphones to virtualized ones [14] and by comparing recordings of IE headphones to recordings of virtualized ones [12]. The agreement between the ratings of actual versus virtualized ranged from almost perfect (r = 0.98) for IE headphones [12], to very good (r = 0.85) for AE/OE headphones [14] where leakage effects and visual/tactile biases in actual headphone tests likely played a role. Together these validation studies provide evidence that the virtual headphone method produce valid and meaningful results.

The virtual headphones only simulated the magnitude and minimum phase part of the headphones and excluded any nonlinear or excess phase distortions that were present in the actual headphones. However, the validation studies together with other headphone investigations suggest that the magnitude response is the dominant factor in how good or bad a headphone sounds.

The magnitude and minimum phase response of each headphone was measured using a G.R.A.S. 45 CA coupler equipped with our custom pinnae optimized to better simulate leakage on human ears [13]. The final measurements were based on average of 3 re-seats of the headphone. The measured magnitude response of each headphone was then simulated over a replicator headphone (AKG K712) chosen for its low distortion, relatively smooth and extended frequency response. The open-back design of the replicator headphone provided a natural leak thus ensuring a more consistent response at low frequencies across listeners. How the headphone fits and mechanically couples to the listeners' head can influence its response below 200 Hz. The replicator headphone was modified using a stiff, curved piece of wire to increase clamping force, which preliminary testing showed would decrease variability of leakage.

An FIR filter was then designed and applied to the replicator headphone to simulate the measured response (see section 2.3 in [12]). The match in measured magnitude response of the actual versus virtualized headphone was good (± 1 dB) up to 12 kHz, above which we did not attempt aggressive equalization due to errors related to the sensitivity of the position of the headphone on the coupler and its accuracy beyond that range.

2.3 Listener Selection

The listening panel consisted of 130 Harman employees located in Novi, Michigan and Northridge, California. The panel included both trained (n = 28) and untrained (n =102) listeners with 78% of the sample male and 22% female. The trained listeners were tested for normal audiometric hearing and had successfully completed level eight for all tasks in the training software "Harman How to Listen" [15]. The listeners ranged in ages from 13 to 65 years old with approximately 37% being under the age of 30, 33% were 30-45 years, and 15% between 45-65 years old. 15% of the listeners failed to report their age. All listeners were paid for their participation.

2.4 Listening Test Procedure

Five listening tests were conducted in two blocked sessions each lasting about 30 m on average. In the first session, listeners completed two listening tests and three tests in the second session. The sessions were generally conducted on separate days. Each test was comprised of 6 trials (3 programs x 2 observations) wherein a total of 8 headphones were evaluated including the hidden high anchor (i.e. the new Harman AE/OE target) and a low anchor. High and low anchors were common to all tests.

Listeners were given written and verbal instructions prior to the test and were encouraged to use the full range of the scale. The test administrator ensured that the headphones were properly seated on the listeners' head to minimize leakage and produce consistent responses.

Altogether each listener provided a total of 240 preference ratings (5 tests x 8 headphones x 3 programs x 2 observations = 240 ratings), generating 31,200 ratings (130 listeners x 240 = 31,200 ratings) from the entire subject pool. The presentation order of the tests, trials, programs and headphones was randomized to circumvent order and learning biases.

2.5 Program Material

Table 1 summarizes the three music programs used in these tests. The stereo tracks were digitally copied from compact disc and edited into brief 15-25 s loops to facilitate listeners' judgement of the headphones according to ITU recommendations [16]. In a pilot test [11], these three programs produced the most discriminating listener preference ratings among ten different programs used to evaluate a subset of the headphones tested in this paper. They also produced no significant program effects or interactions with the headphone preference ratings.

Program	Artist/Track/ Album	Description
SD	Steely Dan/ *Cousin Dupree* / Two Against Nature	Pop/Jazz with male vocal
JW	Jennifer Warnes/ *Bird on a Wire* / Blue Raincoat	Pop with female vocal
BSG	Stu Philips / *Theme from Battle Star Galactica*	Classical Orchestra

Table 1. Description of programs used in these tests.

2.6 Listening Test Software

A custom listening test application was written in MAX/MSP [17], to administer the test, implement the FIR filters for the virtual headphones, and collect the listeners' ratings.

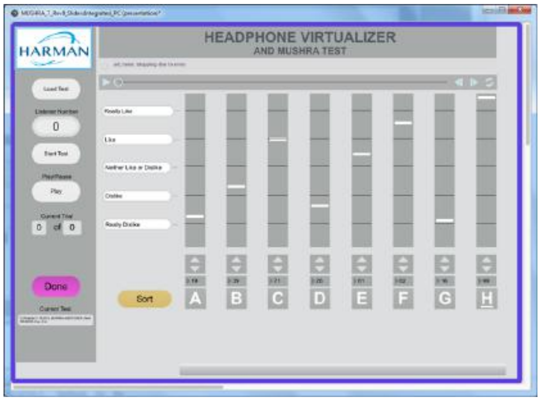


Figure 1 The graphical user interface for the headphone virtualizer and MUSHRA test.

Fig. 1 shows the graphical user interface used by the listener to randomly switch among the different virtualized headphones and rate them. Each headphone was rated on a 100-point preference scale. The software randomized the order in which the headphones and programs were presented in each trial. A "sort" button was included to allow the subjects to at any time sort the columns from low to high scores to facilitate the grading process.

2.7 Relative and Absolute Playback Levels

The relative levels of the virtual headphones were matched according to ITU-R BS 1770-4 loudness model [18]. The authors then performed fine-tuning of the level adjustments through informal listening. The absolute playback level was then set to produce an average level of 85 dB (slow, C-weighted), equivalent diffuse field level.

3 Listening Test Results

3.1 Statistical Analysis

Each of the five listening tests was analysed separately using a repeated measures analysis of variance (ANOVA) where the dependent variable was preference ratings and the independent variables were Headphone (8 levels), Program (3 levels) Training (2 Levels) and Gender (2 levels). All statistical tests were performed at a significance level of 5%.

Table 2 is an abbreviated ANOVA table summarizing the F-statistics for the main effects and interactions. A bold F-value value indicates the effect was highly significant (i.e. p-value is <0.05). In cases where the F-value was not significant the p-value is shown below it.

As expected, Headphone was the dominant effect on listeners' preference ratings followed by Training. Gender also had small effect in all five tests. Program had a small effect that was statistically significant in three of the five tests. There were also some small interaction effects between Headphone and the factors Training, Program and Gender. In the following sections we limit our discussion to Headphone and Training effects and interactions that is the focus of this paper.

Factor	Test 1	Test 2	Test 3	Test 4	Test 5
Headphone	**166.97**	**230.20**	**369**	**164.08**	**228.3**
Training	**122.42**	**136.27**	**101.59**	**136.99**	**86.76**
Program	2.93, p =.54	0.614, p =.54	**3.15**	**4.76**	**6.27**
Gender	**19.21**	**33.91**	**15.31**	**42.53**	**20.51**
Training * Headphone	**9.11**	**16.09**	**5.93**	**11.33**	**11.2**
Program * Headphone	**4.88**	**6.41**	**3.69**	**4.10**	**4.82**
Gender * Headphone	**16.88**	**4.04**	**10.95**	**12.28**	**7.88**

Table 2. Abbreviated ANOVA table showing the F-values for the significant effects (shown in bold) and the p-value where effects were not significant.

3.2 Headphone Effect on Preference

Fig. 2 plots the mean preference ratings and 95% confidence intervals for Headphone for each test. One observation is that Harman target curve received the highest preference rating in every test, except in Test Four where it was statistically tied with HP19. In Test Five, it was preferred overall but statistically tied with HP 25 and HP26. The low anchor was the lowest rated headphone in all tests except in Test Three, in which HP18 received slightly lower ratings.

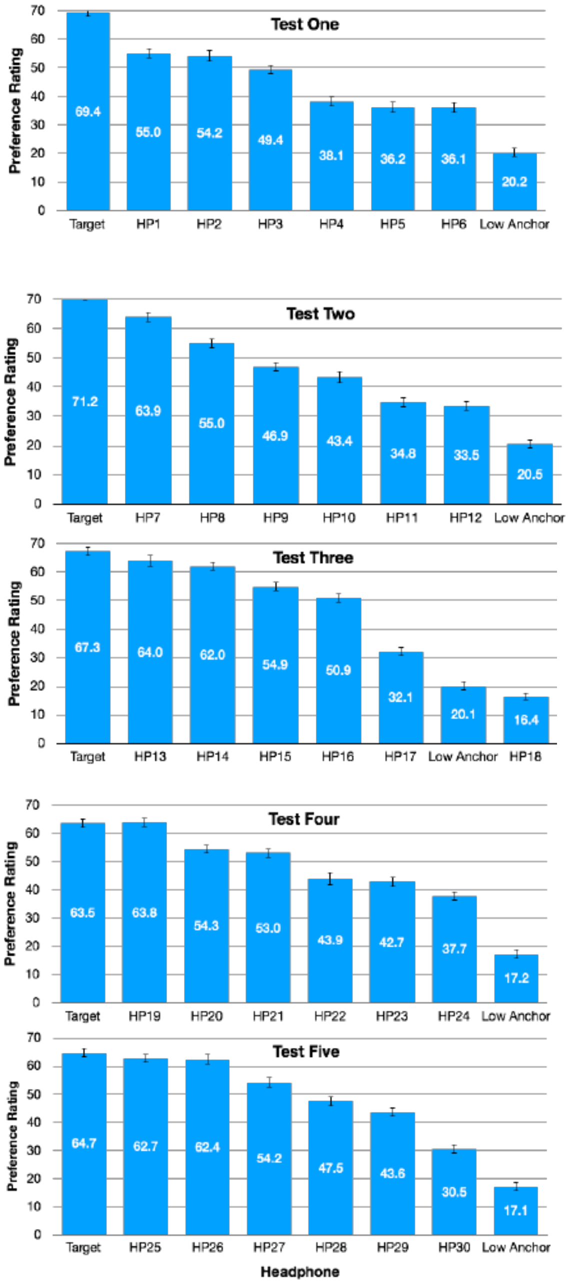


Figure 2. The mean preference ratings and 95% confidence intervals for Headphone shown for each of the five tests.

Looking at Fig. 2 we see that most listeners generally avoided using the upper range of the preference scale beyond 70. This compression effect was likely due to the well-known contraction bias (see section 4.3 of [19] wherein listeners are hesitant to give very high or low ratings near the top and bottom of the scale.

Depending on the test, listeners generally rated the eight headphones in each test into 4 to 6 different groupings based on preference. Beyond that the headphones were either too similar in sound quality or other factors like Program may have produced interactions that affected their ratings.

3.3 Effect of Listener Training on Headphone Preference

In section 3.1, Training was found to be the most dominant effect after Headphone. The effect can be summarized as follows: untrained listeners tended to scale their preferences 10 points higher than trained listeners. The average effect varied between 8.6 points higher in Test Five to 10.7 points in Test One. This behaviour is consistent with prior studies including the IE headphone study [3].

Of more interest was the significant interaction between Training and Headphones that suggests untrained listeners might have different sound quality tastes than those of trained listeners. The interaction effect is plotted in Fig. 3 for each test. It is evident than untrained listeners tend to use higher ratings, as noted above, and a slightly smaller scaling range. While the overall trend in headphone preference was similar between the two listening groups, there were some notable differences for specific headphones: HP13 in Test Three, HP22 in Test Four, and HP26 in Test Five, which for untrained listeners was preferred equally to the Target. It is clear that untrained listeners generally had more difficulty choosing a favourite among the top two or three headphones, and in some cases their preferences between two headphones were different than those of the trained listeners. The possible reasons for this are discussed in section 5.

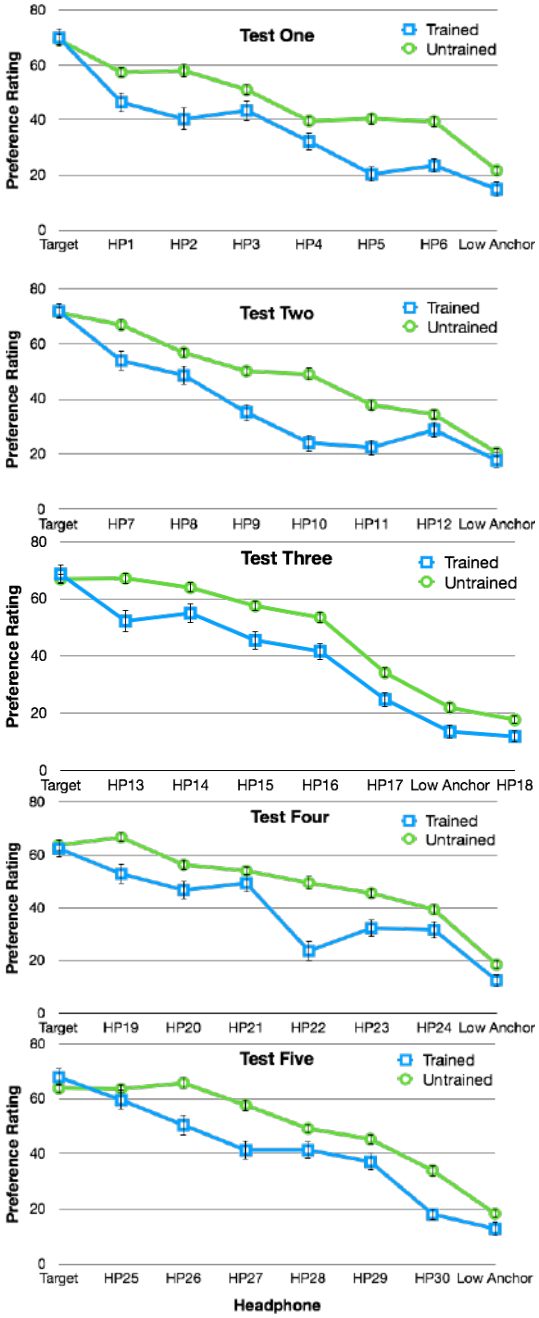


Figure 3 Mean headphone preference ratings and 95% confidence intervals for both trained and untrained listeners in each test.

4 A Statistical Model to Predict Headphone Preferences

In this section we present a statistical model that predicts listeners’ headphone preference ratings based on how much they deviate from the magnitude response of the Harman AE-OE target response. First, we linearly transformed the listener headphone ratings to occupy the entire 100-point preference scale since about 55-60% of the scale was used due to a contraction scaling bias (see section 3.3).

4.1 Relationship between Headphone Frequency Response and Listener Preference Ratings

To graphically explore the broad relationship between the frequency response of the headphones (left/right channels are averaged) and their subjective preference rating we plotted the average frequency response of headphones that fell into four distinctive categories of sound quality based on their preference rating: Excellent (90-100% preference rating), Good (65-76%), Fair (42-54%), and Poor (0-39%). Fig. 4 shows the average magnitude response (blue curve) of each category along with the target curve (green curve). The error response curve (red curve) and a linear regression line (black dotted line) showing the best fit to the error response curve are also plotted.

Looking at Fig. 4 it is clear that headphones generally received lower preference ratings as their frequency response deviated further from the Harman target curve. The headphones in the “excellent” category came closest to the target curve only deviating below 50 Hz where, on average, they fell off. The error response curve has small deviations and is generally flat with a slope approaching zero. The headphones in the “good” category came second closet to the target curve but were deficient in treble above 1kHz and deficient in bass below 100 Hz where they were flat instead of having a gentle rise. The error response curve has slightly larger deviations in it with a larger downward slope due to the deficit in treble.

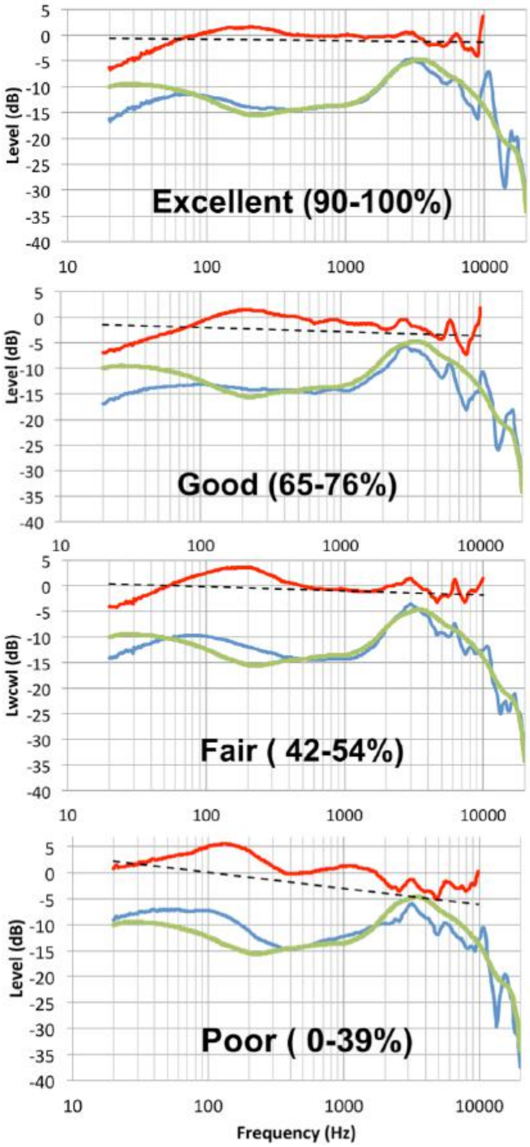


Figure 4 The average frequency response and error curves for headphones in four different categories based on their preference rating. The green curve is the target curve. The dotted curve is the regression line that best fits the error curve.

The headphones in the "fair category" have too much energy between 100 Hz and 500 Hz, which made vocals sound too muddy and colored. The error response has a reasonably flat slope, which could be misleading. However, if data below 50 Hz were ignored the slope would have a greater downward tilt.

Finally, the headphones in the "poor" category also have too much energy between 100 to 500 Hz but also too much below 100 Hz. This characteristic in combination with a deficit in treble above 1 kHz produced a sound profile described by listeners as exaggerated bass, boomy and dull. Our listening panel did not like this sound profile and gave the headphones very low ratings.

4.2 Selection of Independent Variables

Three independent variables were initially selected that provide different statistical measures of deviations in the error response curve discussed in the previous section. All three variables were used in a previous model to predict listener preference ratings of IE headphones [4].

The three variables are defined as follows:
ME – The mean error is based on the sum of the absolute values for each y-value in the headphone error curve from 50 Hz to 10 kHz divided by the total number of n values as defined in equation 1:

$$ME = \frac{\sum_{i=1}^{n} abs(y_i)}{n} \qquad (1)$$

SD - The standard deviation of error defined by the headphone error curve calculated from the y-values from 50 Hz to 10 kHz as defined in equation 2:

$$\mathrm{SD} = \sqrt{\frac{\sum\left(y-\bar{y}\right)^2}{\left(n-1\right)}} \qquad (2)$$

AS - The absolute value of the slope of a logarithmic regression line that best fits the y and x values defined in the headphone error response curve from 50 Hz to 10 kHz according to equation 3:

$$\text{AS} = \sqrt{\frac{\sum\left(\ln(x)-\ln(\bar{x})\right)\left(y-\bar{y}\right)}{\sum\left(\ln(x)-\ln(\bar{x})^2\right)}} \qquad (3)$$

The decision to exclude errors below 50 Hz in the model was based on the finding that these errors contributed little to the underlying variance in headphone preferences based on regression analysis. One possible reason for this is that the average response in all sound quality categories in Fig. 4 – except the "poor" category – drops off significantly below 50 Hz. Within the "poor" category of headphones there is excessive energy between 50 Hz and 500 Hz that contributes to their perceived poor sound quality.

4.3 Predictive Model

A linear model was developed initially using the three independent variables identified in the previous section. The regression analysis was performed using Partial Least Squares (PLS) due to the collinear nature of the explanatory variables. PLS reduces the independent variables to a set of uncorrelated principal components and then performs least square regression.

After an iterative process a linear model was found to produce the best goodness of fit based on the Pearson correlation coefficient of r = 0.86. The statistics for goodness of fit are summarized in table 3 and the equation for the mode is defined in equation 4:

$$\text{Predicted Preference Rating } = 114.49 - (12.62 * SD) - (15.52 * AS) \qquad (4)$$

The standardized coefficients for the variables in the model are weighted approximately equal: SD = -0.47, and AS = -0.434. Note that the model only has two independent variables (i.e. SD and AS) since including the third variable ME added little information to explaining the variance in preference ratings, and reduced the quality of the model.

Observations	32.000
Sum of weights	31.000
DF	29.000
R²	0.741
R	0.861
Std. deviation	6.933
MSE	44.962
RMSE	6.705

Table 3. Goodness of Fit Statistics.

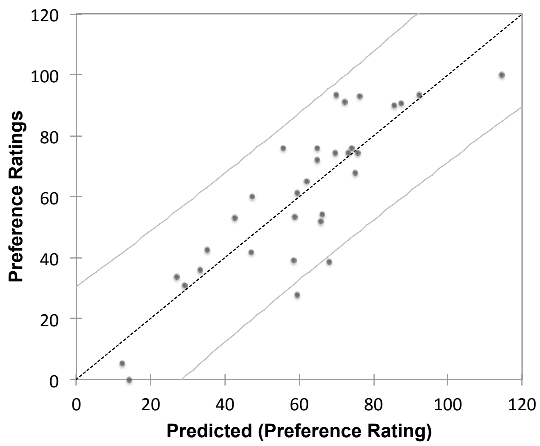


Figure 5 A scatterplot of predicted versus measured preference ratings for 32 headphones with the 95% confidence limits shown.

Fig. 5 shows a scatterplot of the predicted versus measured preference ratings of the headphones including the 95% confidence limits. Two of the headphones (HP6 and HP30) are outliers falling outside the 95% confidence limits. In this case, the model predicted higher ratings for HP6 and HP30 than what the headphones actually received in the listening tests. One possible reason for this is that both headphones have a frequency response that is not well represented by the explanatory variables SE and ME. Looking at their frequency response and error curves (see appendix 2) both HP6 and HP30 have 1-2 large medium Q resonances above 1kHz that may be perceptually underestimated by the statistical measures SD and AS in the model. Both metrics calculate average errors over a wide band and while they capture macro errors they don't capture micro errors contained within narrow bandwidth. The problem and a solution might be explored in the future.

If HP6 and HP30 are excluded from the model the goodness of fit improves to r = 0.90 and the root mean square error (RMSE) is reduced to 5.8 preference-rating points.

5 Discussion

5.1 The Case For Headphone Sound Personalization

A key objective in this study was to validate the sound quality of the new Harman AE/OE target curve using a relatively large listening panel and large sample of competitors' headphones. The listening test results presented in section 3 offer evidence the target curve sounds acceptable to most listeners. For trained listeners it was preferred to the other 31 models. For untrained listeners, the target curve was preferred to 27 of the models, with the other 4 models either equally preferred or slightly preferred to the target.

In a previous study [2] we provided evidence that the preferred level of the bass and treble in a headphone depends on the age, listening experience and possibly gender of the listener. Younger, less experienced listeners tended to prefer more bass and treble than older more experienced listeners. Older listeners (55+ years) on the other hand preferred much less bass and even more treble than the younger listeners. Headphone preference could also be explained by the listeners' degree of hearing loss. We hope to address these questions in a future paper.

5.2 Correlation between Headphone Price and Sound Quality

In the introduction of this paper we summarized two headphone studies that found weak correlation between headphone performance and its retail price. One study by Breebaart [1] was based on frequency response measurements of 283 headphones of all types, and the other one [3] was based on listening tests of in-ear headphones.

This current study also found little correlation between headphone price and sound quality based on listener preferences. Fig. 6 plots the headphone preference rating versus its retail price. The price in USD is plotted on a logarithmic y-scale for better clarity since most of the headphones fall in the $100 to $500 category. A regression line shows a poor fit between headphone price and preference rating (r = 0.17). This poor correlation is similar to what Breebaart reported in his study of 283 headphones, and similar to what we reported in our in-ear headphone study [2]. Together these three studies provide further confirmation that the headphone industry is not following best engineering practices when designing headphones for best sound, even in cases where there are fewer cost constraints. Based on the headphone samples tested in this study, the sound quality of headphones doesn't improve much beyond a $300 price point.

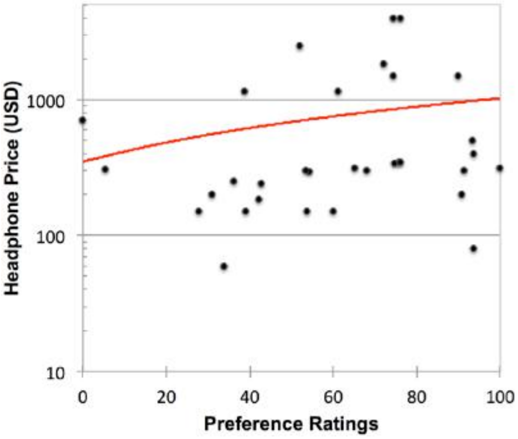


Figure 6 A plot of the 31 headphone preference ratings versus their retail price. The correlation between the two factors is r = 0.17.

6 Conclusions

In the previous sections we presented the results of a large controlled listening experiment where 130 listeners both trained and untrained evaluated 31 models of AE/OE headphones from 19 different manufacturers. The purpose of these tests was to validate the sound quality of a new AE/OE target and use the data to develop a statistical model to predict listeners' headphone preferences based on objective measurements of the headphones.

The following conclusions can be drawn from the results:

1. Headphone was the dominant effect on listener preference. The headphone equalized to the new Harman OE/AE target curve was preferred to 28 of the 31 models tested when combining results of both trained and untrained listeners. Four other models were equally preferred to the target headphone.
2. Untrained listeners tended to use higher preference ratings - about 10 points higher on average than trained listeners. This is consistent with previous studies. Trained listeners preferred the Harman target in all tests. Untrained listeners preferred the Harman target in two of the tests, and rated it about equal to four other headphones in three of the tests.
3. Program and Gender produced smaller effects and interactions with Headphones. The effects were small and not the focus of this paper.
4. Headphones received lower preference ratings as their frequency response deviated further from the response of the Harman target.
5. A statistical model based on these deviations can predict listeners' preference ratings with about 86% accuracy with 6.7% error (see table 3) using two variables: the standard deviation (SD) and the absolute slope (AS) of deviations described by the headphone error response curve.
6. Two outliers were found in the model (HP6 and HP30) that produced higher predicted preference ratings than observed. Both headphones have audible medium Q resonances that we believe are underestimated in the model. An updated version of the model will address this issue in the future.
7. There is poor correlation between the retail price of the headphone and its perceived sound quality based on preference. This confirms previous reports based on measurements [1], and listening test results on in-ear headphones [3].

The last point is symptomatic of a headphone industry in need of scientific guidance in how to optimize the design of headphones for best sound quality. Hopefully, this study will provide such guidance.

Finally, we wish to address the limitations of this study so that the results are not generalized to conditions outside those tested. This study did not address or simulate non-linear or excess phase distortions. Our experiences and others suggest these are not dominant factors in how a headphone sounds, but more research may change our views.

The study did not attempt to simulate masking effects of noisy listening environments, or headphone leakage effects, which can significantly affect the bass performance of the headphone. These limitations do not change or invalidate the results of this study. Good headphone design can mitigate leakage, and to some extent background noise (e.g. ANC and/or good noise isolation through proper seal), in which case the Harman target curve should produce good results.

7 Acknowledgement

Harman International supported this research. The authors thank Jonathan Pierce and Basheer Bazoun at Harman Lifestyle for their help in organizing the listening tests in Novi, Michigan. We are also deeply grateful to all of the Harman listeners in Northridge and Novi for their participation in this study.

References

[1] J. Breebaart, "No correlation between headphone frequency response and retail price," J. Acoustical Society of America, vol. 141, issue 6, (June 2017).

[2] S. Olive and T. Welti, "Factors that Influence Listeners' Preferred Bass and Treble Levels in Headphones," presented at the 139th Audio Eng. Soc., Convention, New York, USA, (2015).

[3] S. Olive, T. Welti, and O. Khonsaripour, "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 1 – Listening Test Results and Acoustic Measurements," presented at the 143rd Audio Eng. Convention, New York, USA (October 2017).

[4] S. Olive, T. Welti, and O. Khonsaripour, "A Statistical Model That Predicts Listeners' Preference Ratings of In-Ear Headphones: Part 2 – Development and Validation of the Model," presented at the 143rd Audio Eng. Convention, New York, USA, (October 2017).

[5] G. Lorho, 2009. "Subjective Evaluation of Headphone Target Frequency Responses," 126th Audio Eng. Soc. Convention, 126, Paper Number 7770. (2009).

[6] F. Fleischmann, F. Silzle, and A. Plogsties, " Identification and Evaluation of Target Curves for Headphones," 133rd Audio Eng. Soc. Convention, Paper Number: 8740 (2012).

[7] S. Olive. T. Welti, and E. McMullin, "Listener Preference For Different Headphone Target Response Curves" 134th Convention, Audio Eng. Soc., preprint no. 8867, (2013 May).

[8] ITU–R Recommendation BS.708: Determination of the electro–acoustical properties of studio monitor headphones, ITU-R. (1990).

[9] International Telecommunications Union, 2015. ITU-R BS 1116-3, "Methods for the subjective assessment of small impairments in audio systems," February 2015.

[10] European Broadcast Union Tech, 3276, 2nd edition, "Listening conditions for the assessment of sound programme material: monophonic and two–channel stereophonic," EBU (1998).

[11] S. Olive, T. Welti, and O. Khonsaripour, "The Influence of Program Material on Sound Quality Ratings of In-ear headphones," presented at the 142nd Audio Eng. Convention, Berlin, Germany, (May 2017).

[12] T. Welti, S.E. Olive, and O. Khonsaripour, "Validation of a Virtual In-Ear Headphone Listening Test Method," presented at the 141st Audio Eng. Soc. Convention, Los Angeles, USA, (September 2016).

[13] T. Welti, "Improved Measurement of Leakage effects for Circum-aural and Supra-aural Headphones," presented at the 138th Audio Eng. Soc., Convention, Warsaw, Poland, (May 2015).

[14] S. Olive. T. Welti, and E. McMullin, "A Virtual Headphone Listening Test Methodology,"51st Audio Eng. Soc. Int. Conference: Loudspeakers and Headphones, Helsinki, Finland (August 2013).

[15] S. Olive, "Harman How to Listen," www.harmanhowtolisten.blogspot.com," (April 2011).

[16] International Telecommunications Union, ITU-R 1534-3," Methods for the subjective assessment of intermediate impairments in audio systems," (October 2015).

[17] MAX/MSP software, Cycling 64, (March 2017).

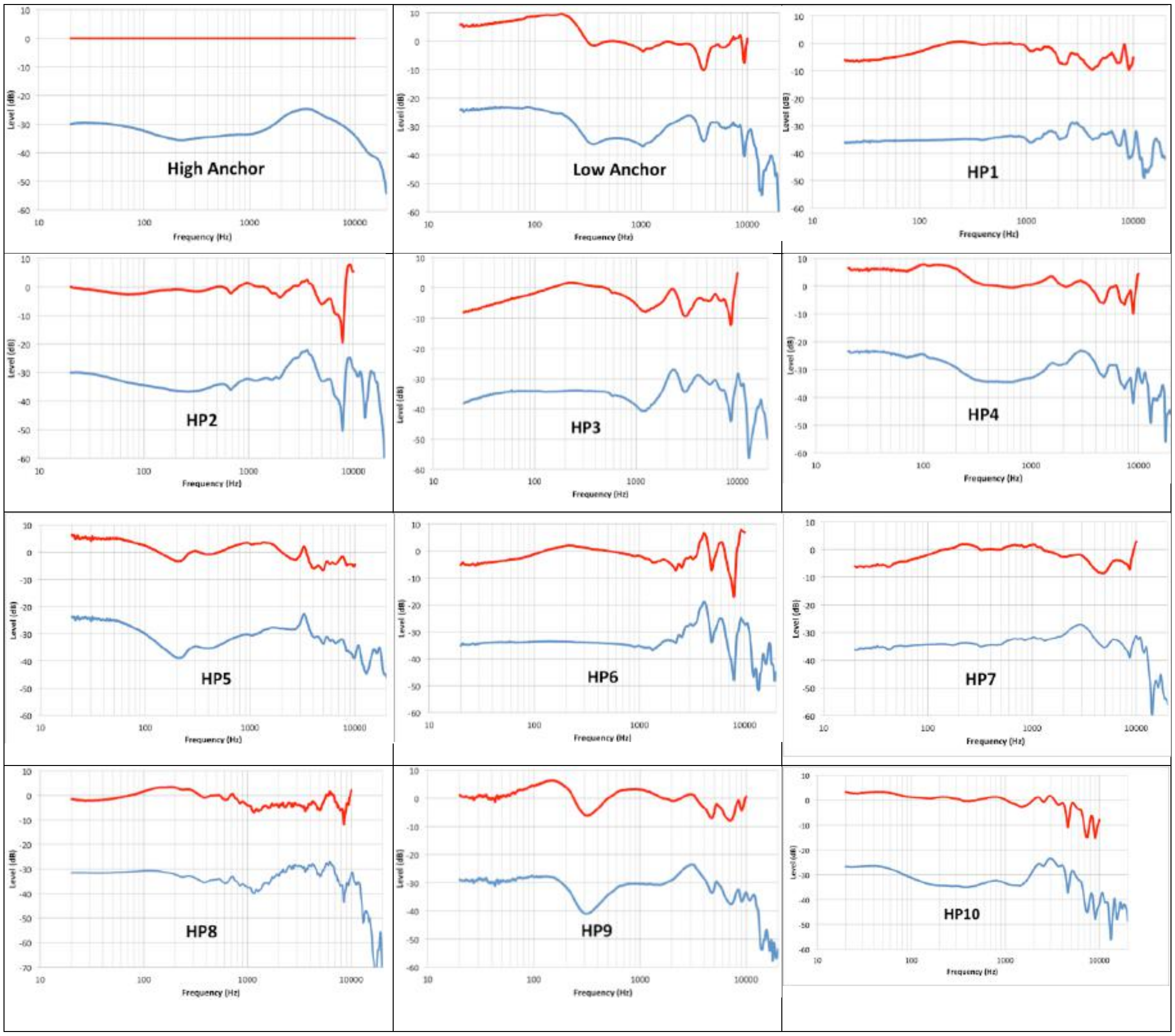
[18] International Telecommunications Union, ITU-R BS 1770-4, "Algorithms to measure audio programme loudness and true-peak audio level," (October 2015).

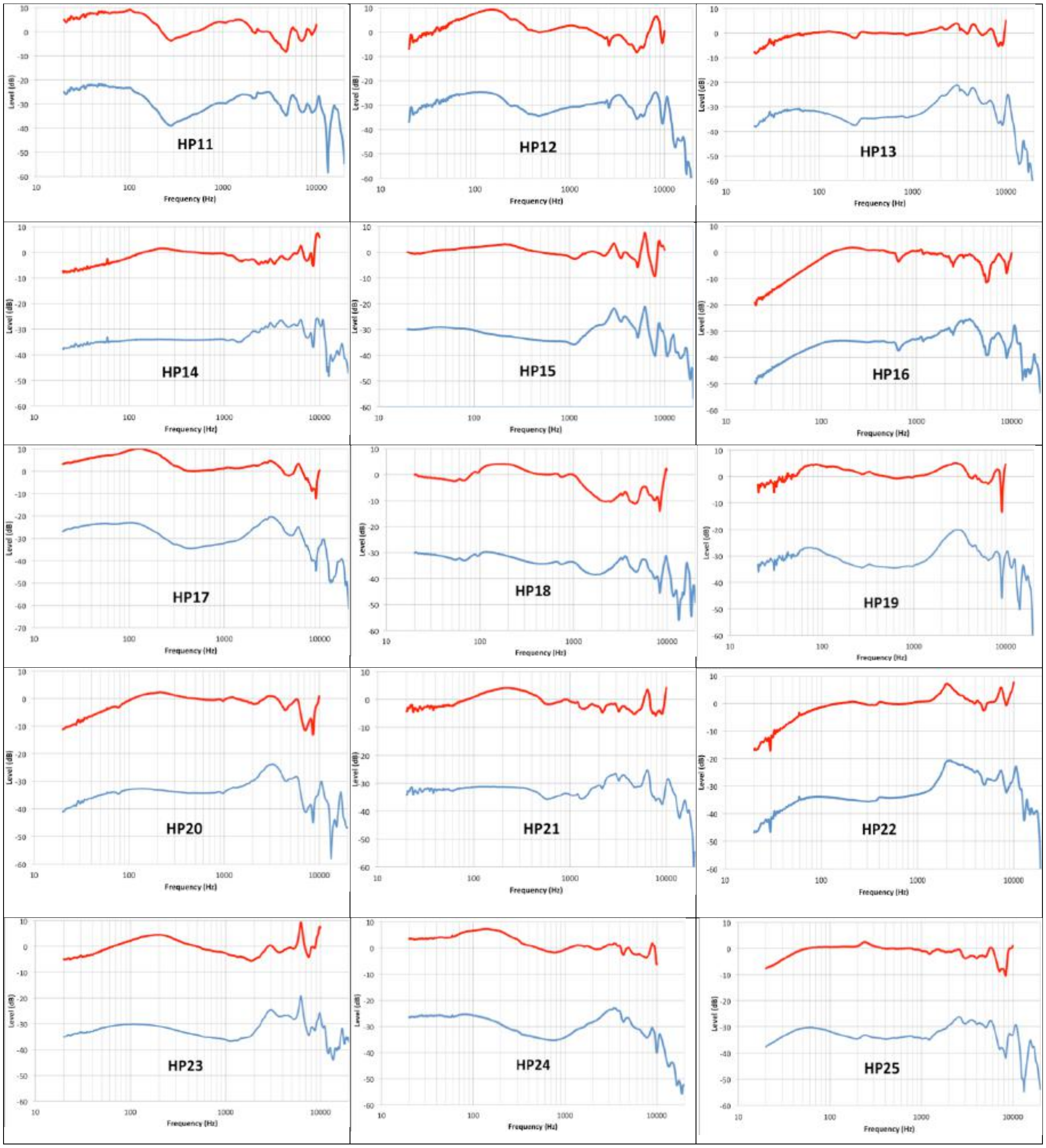
[19] S. Zielinski, F. Rumsey, and S. Bech, On Some Biases Encountered in Modern Audio Quality Listening Tests-A Review," J. Audio Eng. Soc., vol. 56 Issue 6, pp. 427-451, (June 2008).

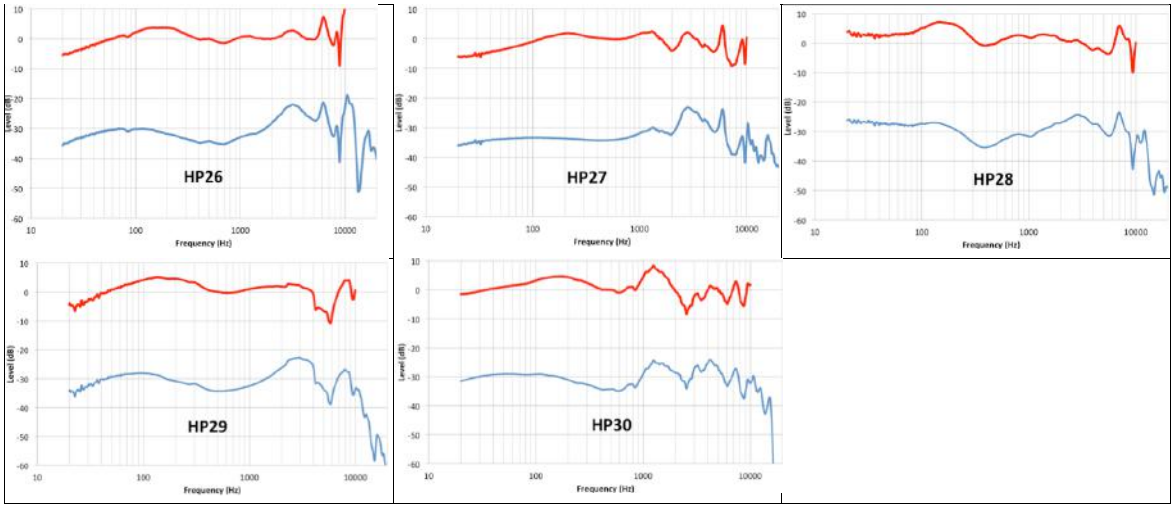
8 Appendix 1. Details on the Headphones in This Study

Brand	Model	Design	Type	Wireless (BT)	ANC	Driver	Price
AKG	K712	AE	Open	N	N	Dynamic	$315
AKG	K812	AE	Open	N	N	Dynamic	$1,155
AKG	K845BT	AE	Closed	Y	N	Dynamic	$240
AKG	N90Q	AE	Closed	N	Y	Dynamic	$1,500
AKG	Y50BT	OE	Closed	Y	N	Dynamic	$150
Audeze	LCD-4	AE	Open	N	N	Planar Magnetic	$4,000
AudioQuest	Night Owl	AE	Closed	N	N	Dynamic	$700
Beats	Solo2 Wireless	OE	Closed	Y	N	Dynamic	$184
Beyerdynamic	DT 990 Pro	AE	Open	N	N	Dynamic	$200
Bose	QC-35	AE	Closed	Y	Y	Dynamic	$350
Focal	Utopia	AE	Open	N	N	Dynamic	$4,000
Grado	Prestige Series SR325e	AE	Open	N	N	Dynamic	$295
HIFIMAN	HE400S	AE	Open	N	N	Planar Magnetic	$300
JBL	E55BT	AE	Closed	Y	N	Dynamic	$150
JBL	Everest 710BT	AE	Closed	Y	N	Dynamic	$250
JBL	Everest Elite 750NC	AE	Closed	Y	Y	Dynamic	$300
JBL	T450BT	OE	Closed	Y	N	Dynamic	$60
Meze	99 Classics	AE	Closed	N	N	Dynamic	$309
Mr. Speaker	Ether Flow	AE	Open	N	N	Planar Magnetic	$1,830
Oppo	PM-3	AE	Closed	N	N	Planar Magnetic	$400
Oppo	PM-1	AE	Open	N	N	Planar Magnetic	$1,160
Philips	Fidelio X1	AE	Open	N	N	Dynamic	$300
Pioneer	SE-Master 1	AE	Open	N	N	Dynamic	$2,500
Sennheiser	HD-25	OE	Closed	N	N	Dynamic	$150
Sennheiser	HD-650	AE	Open	N	N	Dynamic	$340
Sennheiser	HD-800S	AE	Open	N	N	Dynamic	$1,495
Shure	SRH840	AE	Closed	N	N	Dynamic	$200
Shure	SRH1540	AE	Closed	N	N	Dynamic	$500
Sony	MDR-1000X	AE	Closed	Y	Y	Dynamic	$350
Sony	MDR-100ABN	AE	Closed	Y	Y	Dynamic	$150
Sony	MDR-7506	AE	Closed	N	N	Dynamic	$80

Appendix 2. Headphone Measurements and Error Response Curve




HP11
HP12
HP13
HP14
HP15
HP16
HP17
HP18
HP19
HP20
HP21
HP22
HP23
HP24
HP25
Level (dB)
Frequency (Hz)
10
0
-10
-20
-30
-40
-50
-60
-70
10
100
1000
10000


HP26
HP27
HP28
HP29
HP30
Level (dB)
Frequency (Hz)
10
0
-10
-20
-30
-40
-50
-60
10
100
1000
10000