T. Miller and C. Downey, "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

# Listener Preferences for High-Frequency Response of Insert Headphones

**THOMAS MILLER,*AND CRISTINA DOWNEY,** *AES Member*
(Tom.Miller@Knowles.com) (Cristina.Downey@Knowles.com)

*Knowles Electronics, LLC, Itasca, IL, USA*

The frequency response of a headphone is very important for listener satisfaction. Listener preferences have been well studied for frequencies below 10 kHz, but preferences above that frequency are less well known. Recent improvements in the high-frequency performance of ear simulators makes it more practical to study this frequency region now. The goal of this study was to determine the preferred headphone response for insert headphones for the audible range above 10 kHz. A new target response is proposed, based on listener preference ratings in a blind listening test. The results show a clear preference for significantly more high-frequency energy than was proposed in a previous popular headphone target curve. The preferred response is also affected by the listener's hearing thresholds, with additional high-frequency boost being preferred for listeners with age-related hearing loss.

## 0 INTRODUCTION

Listener preferences for headphones are most strongly determined by their frequency response [1], but the response at very high frequencies has not been closely studied. This study aims to find the preferred response of headphones above 10 kHz. Researchers have proposed various target frequency response curves, with the work by Sean Olive and colleagues at Harman showing good correlation to listener ratings. However, Olive et al. noted that their work is mainly applicable up to 12 kHz due to technical limitations at the time of the study [2]. The preferred curve at higher frequencies has not been closely studied or validated with listening tests. This report focuses on listener preferences from 10 kHz to the upper limit of hearing, to be used in combination with the recommendations of Olive et al. at lower frequencies. It expands on results shared earlier by these authors [3].

Recent technical developments have increased interest in testing preferences above 10 kHz. Ear simulators are now available with tight tolerances up to 20 kHz. More insert headphones are including dedicated high-frequency drivers, and wireless headphones universally have electronic equalization available to enable tailoring of the high frequency response. The goal of this research is intended to help headphone designers better match their designs to listener preferences, providing a target curve that can be measured using an ear simulator that corresponds well with high subjective ratings from listeners. A further goal of this research is to learn how normal age-related hearing loss alters the preferred response of earphones.

SEC. 1 of this report covers previous work on target headphone responses. SEC. 2 explores factors important to the design of this preference test. SEC. 3 describes the test results. SEC. 4 examines the accuracy and validity of the results, followed by a discussion and conclusions in SEC. 5 and 6.

## 1 RELATED WORK

There is no direct way to predict the optimal headphone response curve, because a single curve cannot match the complex transmission path from sound in a room to the eardrum when one is not wearing earphones. Sounds in a room reach our ears via both direct and reflected paths, with each path being filtered differently by our hearing system depending on its direction of arrival. This, in turn, is continuously altered as our heads are always moving as we respond to sounds. Therefore, an empirical approach is needed to find a headphone response curve that forms a suitable approximation to open ear listening.

Several headphone response curves have been proposed for headphones. Many of these attempt to mimic how sound from a source reaches the eardrum using a single or a combination of several head-related transfer functions (HRTF) for a listener. For example, one can use the free-field response

*To whom correspondence should be addressed, e-mail: Tom.Miller@Knowles.com. Last updated: March 31st, 2023

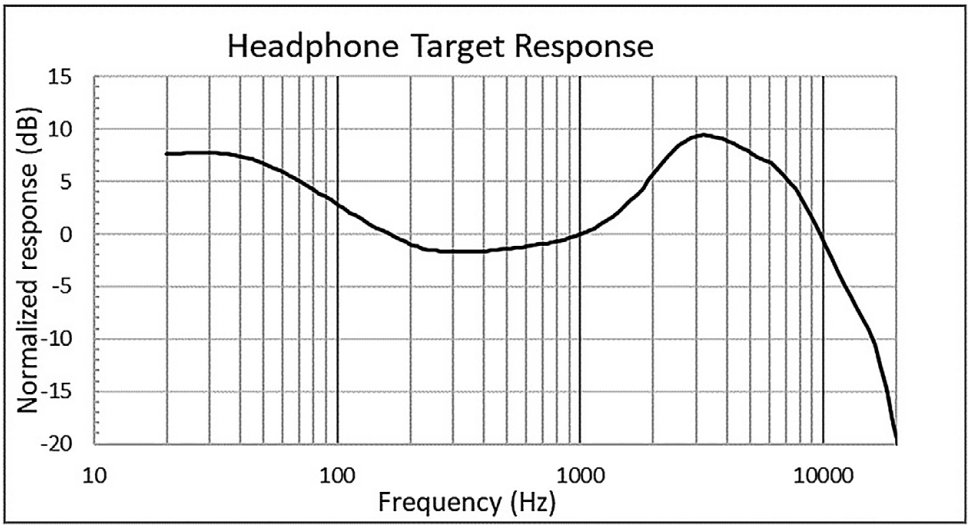


Fig. 1. Olive et al. recommended headphone frequency response.

for a sound source aligned directly in front of the listener, or at an azimuth of 22°, to mimic the typical direction of first arrival from a stereo loudspeaker [4].

Another proposed approach is to use the diffuse-field response, which is the average of HRTF curves for all possible arrival directions [5], similar to listening in a highly reverberant space. This method avoids assuming any specific direction for a sound source but fails to match the sound for any particular direction of arrival. An in-between approach is an average of the frontal hemisphere or some narrower range of angles. This might approximate the effect of including some early reflections along with the direct sound [6]. Another approach is to use a mixture of the diffuse-field response and a range of free-field response curves, adjusted to approximate the balance of diffuse and direct energy arriving at ear the in a given listening scenario [7,8]. The appropriateness of these proposed curves have not been verified with controlled listening tests.

Many recommendations for headphone responses use measurements made at the entrance to the ear canal. This is not practical for insert headphones, which block the entrance to the ear canal. These devices must be measured at the eardrum reference point (DRP) using an ear simulator. Converting pressure at the ear canal entrance to the DRP requires applying a correction specific to the ear simulator or an individual's ear. Recommendations made in this paper will be referenced to the DRP.

Olive et al. [1] recommended a curve based on the sound arriving at the DRP of a mannequin from speakers in a reference room, with further adjustments based on listener recommendations (Fig. 1). Olive et al. demonstrated that this correlated well to subjective ratings, with slightly different bass responses desired for over-the-ear and in-ear headphones (in ear shown).

The curve shown in Fig. 1 is known in the headphone industry as the "Harman curve." It has a boost below 200 Hz that mimics the average influence of room resonances in the Harman Reference Room to the sound of speakers [9]. Like the diffuse-field curve, it includes a boost from 3 to 8 kHz that mimics the effects of the open ear canal resonance and pinna. The curve above 10 kHz shows a steep attenuation with frequency. Olive et al. stated that their group chose not to equalize their headphone above 12 kHz because of the uncertainty in the headphone measurements and in the presentation to the listeners. Therefore, it seems likely the Harman curve above 12 kHz represents the response of their reference headphones used in testing rather than an idealized response based on listener preferences.

## 2 TEST METHOD OVERVIEW

A goal of this investigation was to build on the work of Olive et al., providing new information in the 10-kHz to 20-kHz region. The methods described in this article were matched as closely as possible to the methods used by Olive et al., including the use of Multiple Stimulus with Hidden Reference and Anchor (MUSHRA) double-blind testing with the virtual headphone method. Additional steps were required to overcome issues specific to high-frequency operation. These issues include the selection of appropriate music, the effect of the subject's hearing thresholds, headphone measurement accuracy, and the choice of appropriate test curves.

### 2.1 High-Frequency Controls

To reduce bias, the music chosen for testing was selected to have spectra similar to typical popular music. By definition, this genre represents the music with which the most listeners will have had some experience and are most likely to hear in future headphone usage. An average of the spectra of 200 popular songs from two most recent decades was used as a reference for selecting song excerpts and is shown in Fig. 2. This was inspired by the work of Elowsson and Friberg [10] and by Pestana et al. [11].

Ten music recordings were randomly selected from each of the years from 2001 to 2020, for a total of 200 recordings. Unique to this work is the use of gating, to exclude portions

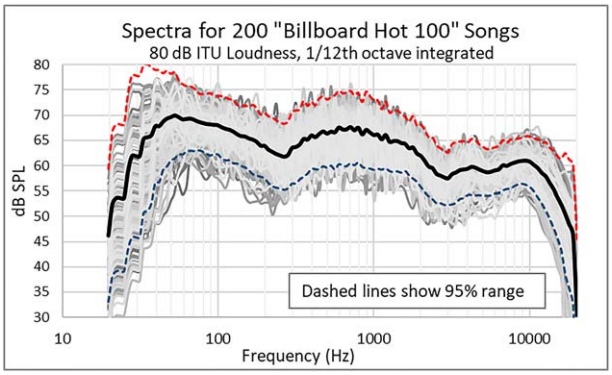


Fig. 2. Spectra of 200 Pop songs. Heavy black line indicates the group average.

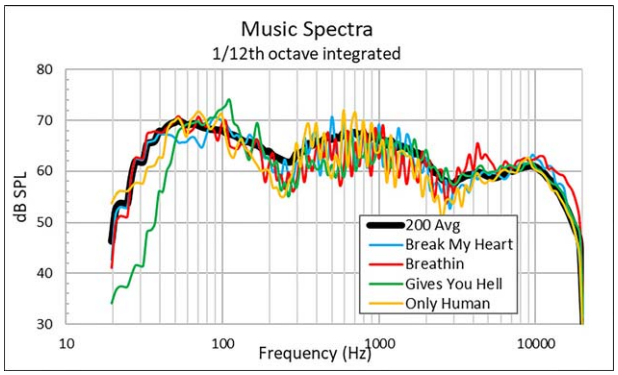


Fig. 3. Test excerpts. Spectra are similar to 200-song average.

of songs in which high-frequency instruments were not playing, such as during the introductions of songs. Data are presented using 1/12th octave integration. The upper and lower dashed lines are based on standard deviation, showing that 95% of the songs fit within ±6 dB of the average curve. The dashed lines are not a perfect fit to the data, because the distribution of spectra is not Gaussian, especially at the lowest frequencies.

Another criterion for song selection was to have the high-frequency energy supplied by familiar acoustic sources for which the listeners would have strong internal references, such as percussion instruments and voice. Short clips were used of "Breathin'" by Ariana Grande, "Break My Heart" by Dua Lipa, "Only Human" by the Jonas Brothers, and "Gives You Hell" by the All American Rejects. Clips were taken from the chorus sections of the songs, where the greatest number of instruments were likely to be playing and the spectra were most consistent. Fig. 3 shows the songs are representative of the 200 song average. One clip had significantly less energy below 50 Hz, but that was not expected to alter judgements of high-frequency energy because that still includes most bass guitar notes.

The presentation level is known to alter impressions of timbre, as is reflected in the equal loudness curves in ISO 226 [12]. Therefore, it is important to use a consistent presentation level. Music was presented at 80 dB SPL, to be consistent with the SPL used by Olive et al. in their

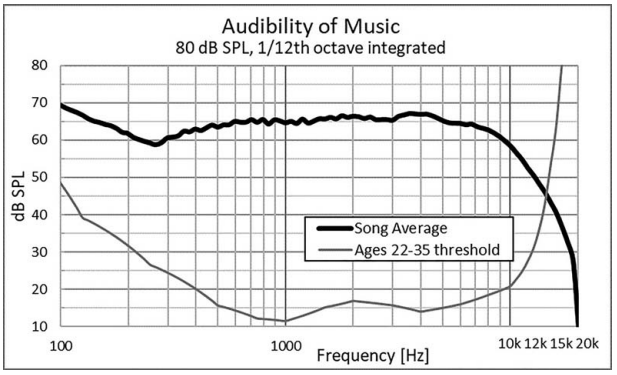


Fig. 4. Audibility of 80-dB SPL average pop music.

development of the Harman curve that is used as a basis for this experiment.

The music spectral analysis can also be used to predict the audible bandwidth of the musical energy. This allows the experiment designer to know the bandwidth that can impact user ratings of headphone quality. In Fig. 4, the upper curve shows the music spectrum, and the lower curve is the typical hearing threshold data at the DRP for people in the age range of 22 to 35 years old [13]. The music content and the audibility curves cross near 17 kHz, although the exact crossing point depends on the headphone response shape and the loudness of the music. Above this frequency, the music content is not loud enough to be heard.

In this graph, the music spectrum is adjusted to show the energy at the DRP. While the curve in Fig. 2 shows the spectrum of the music if replayed into free space through a system with perfectly flat frequency response, Fig. 4 shows the spectra of the same music at the DRP, for music played through a headphone having a response matching the Harman curve from Fig. 1. The headphone loudness has been adjusted to match 80-dB SPL presentation in a free space by using the inverse of the diffuse loudness curve.

The high-frequency response of headphones varies depending on the method used to measure them. A general background of headphone measurement is provided in [14]. Different designs of ear simulators provide different responses. None of them can match the precise acoustic properties of an individual's ears, due to their complexity and to the great variations between individual ears [15]. Standardized ear simulators provide a reasonable, if imperfect, method for comparing headphones. Headphones in this study were measured using the GRAS Sound and Vibration RA0401 ear simulator, which is based on IEC 60318-4 [16]. This device has a tolerance of ±2.2 dB from 10 to 20 kHz [17]. This tight tolerance is achieved by damping the 12-kHz resonance that occurred in previous IEC 60318-4–compliant ear simulators. All headphone target curves in this paper are based on measurements made using this ear simulator.

Another ear simulator that provides excellent high-frequency response is the 4620 made by Brüel and Kjær (and described in ITU P.57 4.3 [18]). It is beyond the scope of this paper to determine which of these devices better represents a typical ear. Both provide useful mimicry of human


Fig. 5. Headphone measurement using funnel-shaped adapter.

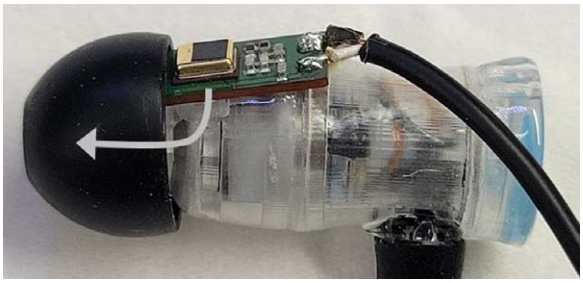

Fig. 6. Custom Knowles headphone used for listening tests. Arrow indicates path for MEMS microphone to sense ear canal sound.

anatomy. However, the differences in ear canal geometry and the acoustic impedance of this simulator are different enough that one cannot use a simple calibration to convert measurements made on this simulator to reliably replicate curves made on the GRAS ear simulator. Therefore, experimenters will need to use a device with an impedance and geometry similar to the GRAS ear simulator to replicate the results presented here.

Measurements for this study were made using a metal funnel-shaped adapter, as shown in Fig. 5. This avoided variability that can occur when using pinna with an ear simulator. The insertion depth and alignment of headphone opening axis to the ear canal axis are more variable when using a pinna, causing variation in the high-frequency response. Although the shape and material properties of this adapter are quite different that that of a rubber pinna, the effective canal length is similar, so that the acoustic resonances in the 10- to 20-kHz region are generally similar to that measured with a GRAS rubber pinna. Although not a perfect solution, the errors introduced by using this coupler were smaller than the potential errors that can occur with a poor fit to the rubber pinna.

A headphone with extended high-frequency response was selected for the listening tests. This was a custom headphone using a 7-mm-diameter moving coil driver made by UTM for the woofer and a Knowles WBFK balanced armature tweeter (Fig. 6). Each driver had a separate outlet tube to the rubber ear tip to ensure a smooth and extended high-frequency response. A Micro-Electro-Mechanical System (MEMS) microphone was added to enable checking if the headphone was properly sealed to the subject's ear. The change in bass response that occurs with a leak could alter the subjects' judgements of treble to bass balance.

Each headphone was equalized to have a flat response before applying the various test curves. Equalization gain

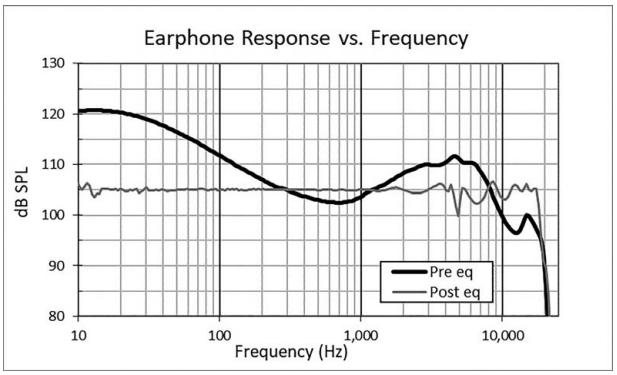


Fig. 7. Headphone response before (dark) and after calibration equalization (light).

was limited to 20 dB to avoid requiring excessive voltage, providing flat response up to 18 kHz. A typical post equalization response is shown in Fig. 7. Although this falls short of the 20-kHz goal, it exceeds the limit of music audibility determined earlier, so 18-kHz bandwidth was sufficient for this test.

A unique impulse response representing the corrective equalization was created for each earpiece of each headphone. The calibration also included the response of the listening test headphone amplifier interacting with the headphone.

Playback was through a StarTech ICUSBAUDIO2D external sound card and headphone amplifier connected via USB audio interface to a laptop computer. Playback was controlled using a Cycling 74 Max patch. All signal processing and playback were done at a sample rate of 48 kHz. Any signal processing effects within the Windows Sound system were disabled.

The music recordings were convolved with the headphone calibration impulse response and an impulse response representing each target filter to create the program excerpts used in the test

$$h_{\text{final}} = h_{\text{cal}} * h_{\text{target}} * h_{\text{music}}. \quad (1)$$

The calibration and target transfer functions were each 8,192-point impulse responses.

## 2.2 Curve Selection

An inherent limitation of preference testing is that subjects can only compare a limited number of alternatives with good repeatability. Larger numbers of choices cause both fatigue and confusion, leading to more random variation in the results. This test limited the choices to eight, including the anchor and the hidden reference. A series of less rigorous small-group listening tests were used to narrow down the range of possible curves to be used later with more rigorous testing of a larger group of subjects. As stated before, frequencies above 17 kHz were not expected to be audible. The region from 10 to 17 kHz has a ¾ of an octave-wide frequency span, so it cannot contain many peaks and valleys unless they have very high Q. Such narrow features were not expected to be needed, especially

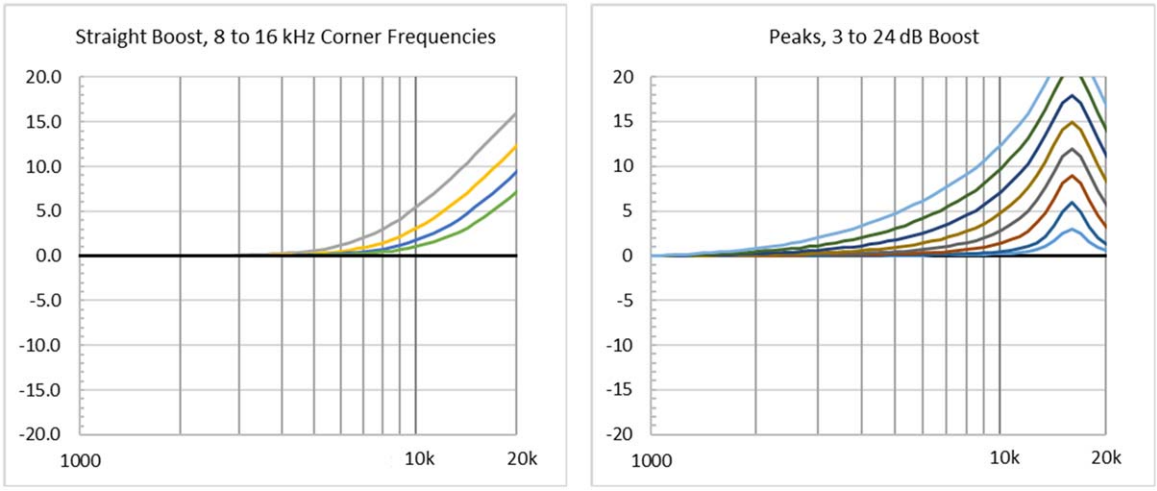


Fig. 8. Filter prototypes for testing. Shelving filters on left, peak filters on right.

after examining HRTF curves. Therefore, only simple filter shapes were explored.

The target curves were formed by superimposing filter shapes onto the Harman curve, rather than independently creating new responses. This ensured smooth transitions from the original low-frequency curve to the experimental high-frequency curves. Shelving and peaking filters were explored. The use of a MUSHRA test interface (described in SEC. 2.3) allowed trying a wide variety of curves during the initial exploration.

Shelving filters were found to apply gain inefficiently. High-order filters were needed to provide significant gain in the 12- to 14-kHz region while minimizing changes at frequencies below 10 kHz. However, such high-order filters produced very high gain near 20 kHz (Fig. 8), where signals are inaudible. This required very high dynamic range. Peak filters were more effective for creating useful change within the audible range while not requiring extreme gain at inaudibly high frequencies.

Digital peaking filters with a center frequency near the upper band limit show considerable frequency warping compared with the equivalent analog filter, reducing the frequency where a given amount of boost occurs. This creates additional gain in the 10-kHz region. To avoid frequency warping, prototype filters were created using a 96-kHz sample rate, then their impulse responses were down-sampled to 48 kHz for later use in convolution filtering of the music. This provided a filter shape more similar to that of an analog filter.

Initial trials of peak filters explored a range of gain, Q values, and center frequencies. The best-liked filters provided a strong boost at the upper end of the audible frequency range, with progressively less boost when approaching 10 kHz. A peak filter with a center frequency of 16 kHz with a Q of 4 was found to be an effective way to achieve this. The remaining variable to be tested was the appropriate filter gain.

The final large-group trials presented music with a range of different boosts and no change to the center frequency of 16 kHz or the Q of 4. The combined effect of the filter

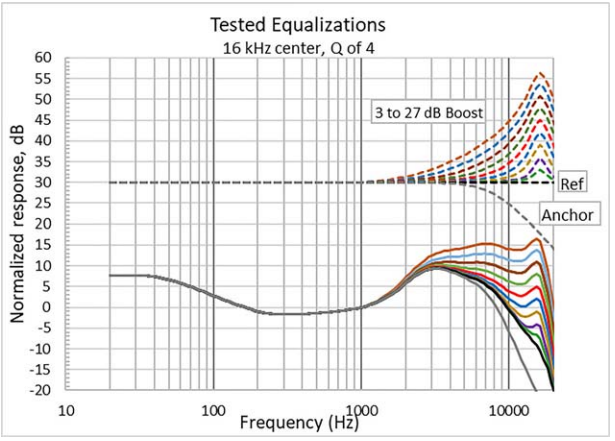


Fig. 9. Curves used in large group testing. Modification curves shown as dashed lines, offset by 30 dB for readability. Combined response of Harman curve and modification curves shown as solid lines.

variations is shown in Fig. 9. The upper dashed lines are the modification curves, and the lower solid lines are the combined effect of the Harman and the modification curves, which were presented to the subjects. The Harman curve is shown in bold.

The higher gain curves show a spillover effect below the 10–20-kHz region of interest, which is the result of keeping the Q constant. This was considered acceptable for two reasons. First, increasing the filter sharpness while increasing the gain risks confounding two perceptual effects with one adjustment—brightness and ringing. Pretrials had already shown the listeners did not like high Q filters. Second, although the primary goal was to retain the Harman curve for frequencies below 10 kHz for listeners with normal hearing, it was expected that listeners with reduced hearing ability may desire some compensation in the octave below 10 kHz, so some alteration of frequencies below 10 kHz was considered acceptable for high gain equalizations.

The lowest curve shown in Fig. 9 is the anchor curve, an 8-kHz, second-order, low-pass–filtered version of the

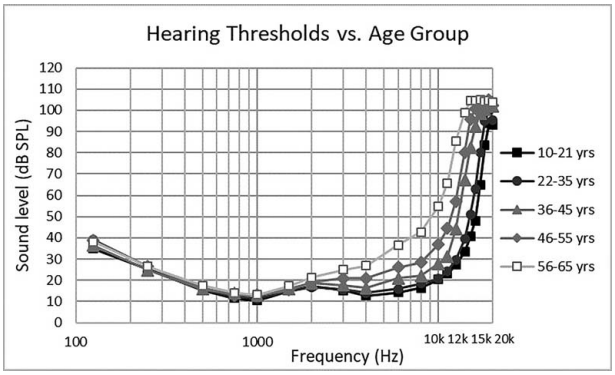


Fig. 10. Average hearing thresholds rise as people age.

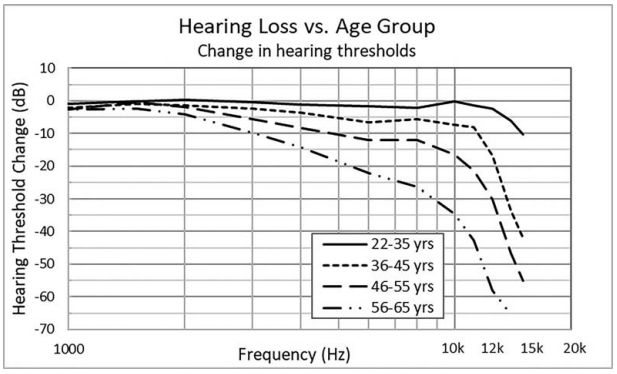


Fig. 11. Change in hearing thresholds with age, relative to 10–21-year-old group.

Harman curve. The anchor curve removes most of the energy in the high-frequency region being explored, serving as a worst case example. Subjects who prefer the anchor likely do not want to hear high-frequency sound, which makes them outliers from most music listeners. The filter design for the anchor curve was limited to second order with Butterworth damping to avoid the risk of creating ringing artifacts near the corner frequency.

The second lowest curve in Fig. 9 is the Harman curve, which was selected as the hidden reference to be included in all tests. It was selected because it has been studied in great detail already [19] and is known to receive good ratings from listeners. The ITU MUSHRA test recommendation was originally created for the measurement of audio impairment, not for preference testing. However, its ability to avoid many sources of bias in subjective testing is equally useful for preference testing. The role of the hidden reference, though, changes for preference testing. In impairment testing, the hidden reference is used to determine how accurately the subjects judge a sound with no impairment. For preference testing, there is no unimpaired choice possible. Instead, the reference reveals if the subjects prefer to depart from the best previously known practice.

The test was structured to control for hearing ability, as this was expected to influence preferences for high-frequency response. Fig. 10 shows average hearing threshold data vs. age for people who have clinically normal hearing [13]. These data were collected by Lee and associates using ear canal depth-compensation calibration of the signal source. This improves the accuracy of high-frequency measurements by accounting for how sound reflected from the ear drum affects the pressure at the calibrating microphone. The change in high-frequency thresholds between age groups is shown in Fig. 11, displayed in a format resembling audiograms. There is a gradual loss of hearing up to about 10 kHz, with a steeper change above that. The elevated thresholds not only make some sounds inaudible, but the perceived loudness of sounds near threshold levels is reduced, a phenomenon called recruitment [20]. Sounds near threshold are affected much more strongly than sounds well above threshold.

Subjects were assigned to groups based on similarity of their audiograms to the curves of Fig. 11, rather than using their actual age, yielding a fairly even distribution across

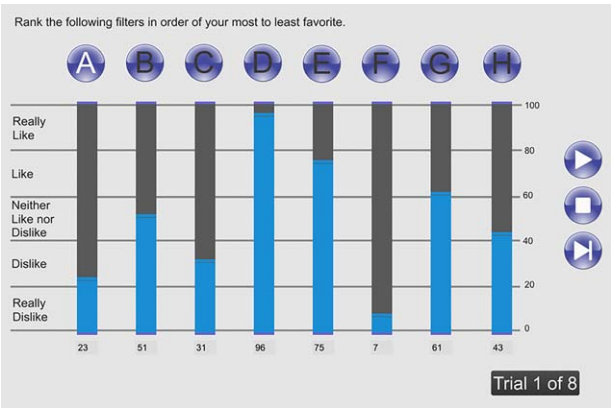


Fig. 12. Typical MUSHRA user interface (from Institute of Sound Recording).

the four age groups. They were tested using a wideband audiometer that provided measurements up to 20 kHz. Although some subjects had significant high-frequency loss, only two subjects were hearing aid users. A total of 133 subjects were tested. Subjects were primarily Knowles employees, and were all in the United States. With just a couple exceptions, none had training in listening testing.

### 2.3 Test Method

Testing was done using the virtual headphone [21] method with the MUSHRA method [22], similar to that used by Olive et al. The virtual headphone method uses equalization of a reference headphone to replicate the sound of alternative designs. The user interface used in testing is shown in Fig. 12 [23], with each slider representing a different equalization of the headphone. Subjects could rapidly switch between different equalizations with no interruption to the music and compare any particular equalization to another in rapid succession. The number of equalizations was limited to eight to avoid overwhelming the subjects. This included six curves with various amounts of treble boost, the Harman curve as the reference, and the low-pass–filtered anchor curve.

To reduce the required number of curves, a somewhat different selection of curves was provided for each age

Table 1. Filter boosts used in MUSHRA test.

| Age | Boost Selections (3-dB steps) |
|---|---|
| 22–35 years | 3–18 dB |
| 36–45 years | 6–21 dB |
| 46–55 years | 9–24 dB |
| 56–65 years | 12–27 dB |

group, based on preferences observed in preliminary trials (Table 1). A step size of 3 dB between equalizations was chosen as a compromise between having fine amplitude resolution and presenting a wide range of gain choices within the eight alternatives presented to each subject. The range of boosts presented to each group was progressively increased by 3 dB. Presenting tailored stimuli to each group enabled increasing the number of choices likely to receive high scores, increasing test resolution.

Subjects were presented with clips of four songs lasting up to 80 seconds in length, in looped form to provide continuous output. The subject would rate all equalizations of one song before proceeding to the next one. Each song was presented twice for a total of eight trials. The order of songs was randomized, and the order of equalizations was randomized each time a new song was presented. Subjects were asked to rate the curves based on their personal preference, giving their most preferred selection a very high rating, and their least preferred selection a very low rating. They were instructed that they should expect to hear differences in the uppermost octave of music, impacting the sound of cymbals and the upper edge of vocals. Subjects were not informed that the selections included the Harman curve or the hidden anchor.

The seal of the headphone to the ear was confirmed before and after each listening session by playing a sine sweep through the earphone and measuring the sound level in the ear using the microphone in the headset. A drop in level at low frequencies would indicate the presense of a leak. If the leak could not be fixed by reinsertion in the ear, then different size and/or different material ear tips (rubber vs. foam) were substituted to enable a good seal.

## 3 TEST RESULTS

Fig. 13 shows the subjects' preference ratings on the *y*-axis and the amount of boost relative to the Harman curve on the *x*-axis. Subject's scores were rescaled before being combined with others in their group, and group scores were rescaled before graphing. Subjects were instructed to use the full range available. However, most of the subjects were new to subjective testing and thus did not consistently use the full range. Rescaling assured all subjects had equal weight within a group average. Scaling of group scores assures that greater agreement between subjects within one group does not create the appearance of one group entering higher ratings than another. Different age groups were presented with different ranges of treble boost. The dots in each curve indicate the levels presented to each group.

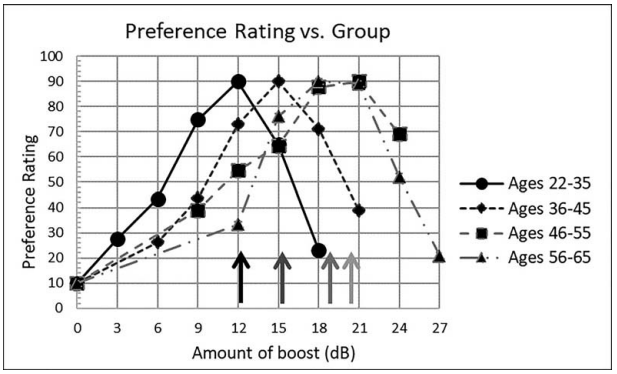


Fig. 13. Preferred boost level. Normalized average of normalized preferences.

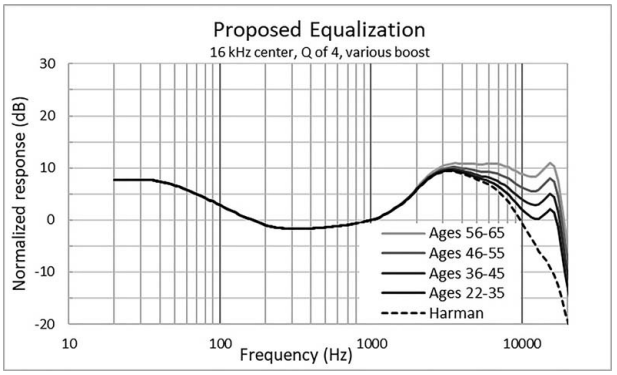


Fig. 14. Preferred response curves for the four age groups.

All groups preferred a significantly higher amount of treble than specified in the Harman curve. People with little hearing loss wanted about 12 dB more level at 16 kHz. Each group wanted progressively more treble boost to compensate for increasing hearing loss, approximating 3 dB for each 10 years of age over the age of 30. Within any group, the treble boost could be adjusted ±3 dB (one step size in the test material) with only a small drop in the group's average quality rating.

Fig. 14 shows the frequency response curves that each group preferred. Although the preferred curves are much higher above 10 kHz than the Harman curve, the levels are not so high as to be impractical to obtain with reasonable hardware. The preferred level at 16 kHz for the youngest group is similar to the level at 1 kHz. The preferred 16-kHz level for the oldest cohort is similar in level to that of the 3-kHz peak.

About 7% of subjects in each age group gave their highest ratings to the anchor curve. As it appeared that these subjects preferred not to hear any treble, it was not practical to include their input to group averages that showed the appropriate level of treble, so these responses were removed from results shown after Fig. 13. These outliers will be discussed further in SEC. 3.2.

Fig. 15 shows quartiles and the minimum and maximum preferences within each age group, after removing those that preferred the anchor. The mean climbs with hearing loss in the first three groups but not in the last. This is due to

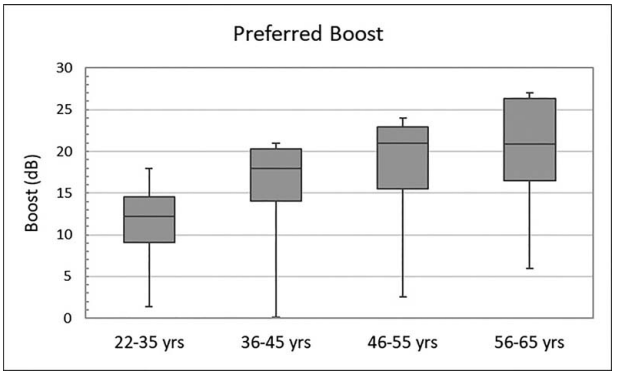


Fig. 15. Distribution of preferred boost level, showing minimum, maximum and quartiles. Anchor responses excluded.

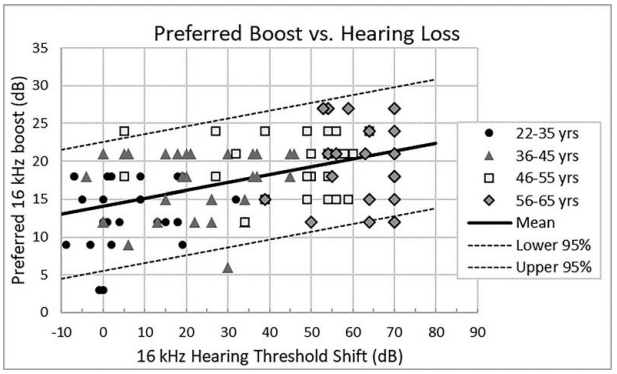


Fig. 16. Scatter plot of all subjects' preferred 16-kHz boost vs. 16-kHz hearing threshold shift.

a few subjects with high hearing loss reporting preferences that deviated far below the group average.

### 3.1 Effect of Grouping

The scatter plot in Fig. 16 shows the effect of hearing loss on preferences while avoiding the age-related grouping used in the previous plots. Age is a small but still important factor in preferences. Subjects who preferred no boost or the anchor have been removed for this graph. Shading and symbols indicate the age groups that subjects were assigned to. A line fit to the data using a regression analysis has a slope of 1-dB rise in the preferred boost for each 10 dB of increase in high-frequency hearing threshold. Dashed lines show upper and lower bounds adjusted to contain 95% of the population. Adjustment is based on fraction of responses, not on mean and standard deviation, because responses did not fit a Gaussian distribution. These lines are ±8.5 dB from the average. This is generally consistent with the rise in preferred boost for the different age groups, showing the use of age grouping was a valid test method.

Fig. 17 shows the variation in $R^2$ value of the linear regression when using hearing loss at different frequencies to predict preferences. The regression quality improves greatly after censoring the subjects who preferred no treble boost or the anchor, because their preferences were independent of hearing loss. The $R^2$ value indicates that

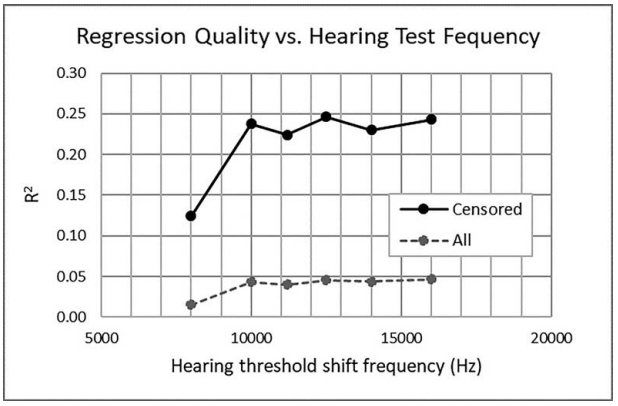


Fig. 17. Quality of regression fit for different hearing assessment frequencies.

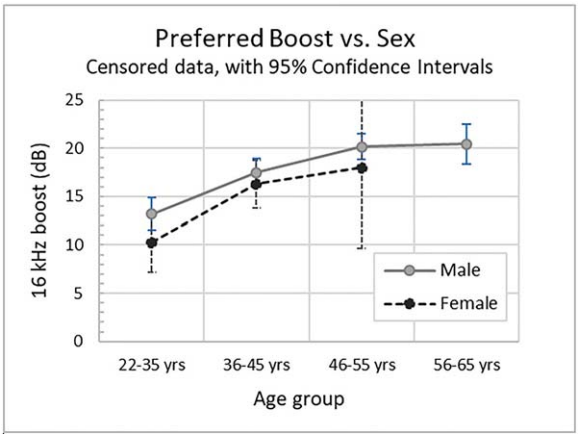


Fig. 18. Effect of sex on preferred response, with confidence intervals.

about 25% of the variation in preferences is linked to high-frequency hearing loss. The $R^2$ value is similar for all hearing threshold frequencies of 10 kHz or higher. This indicates a single frequency above 10 kHz from a hearing test can be effective to guide the selection of an appropriate headphone frequency response.

### 3.2 Secondary Contributing Factors

An Analysis of Variance (ANOVA) analysis was used to show correlation between self-reported background factors and the subjects' preferred boost and is summarized in Table 2. The $F$ value indicates the amount of variance, and the $p$ value indicates the probability that the variance was due to chance. The censored data exclude outliers. The only factors considered statistically significant ($p < 0.05$) were hearing threshold, sex at birth, and preferred musical style. The effect of hearing threshold is particularly strong for censored data. The effect of musical style is only marginally significant.

The difference in preference between men and women is not easily explained by physical or cultural differences. The female subjects generally preferred 3 dB less treble than the men. Fig. 18 shows that the impact of sex is nearly independent of hearing loss, so this effect is not related

Table 2. ANOVA analysis summary, statistically significant values in bold.

| | Uncensored | | Censored | |
|---|---|---|---|---|
| | *F* value | *p* value | *F* value | *p* value |
| 12.5-kHz threshold | **4.9** | **0.003** | **23.2** | **9.5e-12** |
| Sex | **9.9** | **0.002** | **9.1** | **0.003** |
| Preferred music style | 1.9 | 0.092 | **2.4** | **0.037** |
| Cultural heritage | 1.6 | 0.179 | 2.4 | 0.060 |
| Preferred headphone type | 0.3 | 0.790 | 2.5 | 0.068 |
| Listening test experience | 0.9 | 0.401 | 1.9 | 0.162 |
| Hours/day usage | 1.3 | 0.275 | 2.8 | 0.032 |
| Music vs. telephone usage | 0.7 | 0.532 | 0.1 | 0.970 |

Table 3. Number of subjects preferring boost vs. their sex.

| | No boost | Boost |
|---|---|---|
| Male | 7 | 94 |
| Female | 6 | 26 |

to age-related losses that tend to be greater in men. Cubic interpolation of user responses was used to obtain finer resolution in this graph. A curve was fit to each subjects' preference ratings vs. boost using a cubic fit, and the peak location of this curve was used to estimate boost that each subject would most prefer. The small number of female subjects in the high-hearing-loss groups weakens the statistical strength of the trends for the female group, causing the large confidence interval for the 46–55 age group and lack of data for the oldest age group.

Questionnaire data generally were not correlated to why some subjects preferred low amounts of treble. The exception to this is sex. Women were three times more likely than men to not want any treble boost relative to the Harman curve (Table 3). This may be linked to women generally preferring less treble overall, as was shown in Fig. 18.

## 4 ANALYSIS

The difference between the Harman curve and the curves preferred in this testing is substantial. The following section covers reasons why the differences may have occurred and what the sources of variability were.

### 4.1 Similarity to HRTF Curves

The high-frequency preferences shown in Fig. 13 for subjects with little hearing loss are likely to be attributable to differences in the sound pressure at the eardrum for an open ear and when using an insert headphone. Fig. 19 shows a comparison of the preferred response curve for the youngest group of subjects to an average of frontal incidence measurements of a KEMAR head and torso simulator. A close match would indicate that subjects preferred headphones that provided a high-frequency response similar to that experienced when listening to a stereo speaker system having a flat high-frequency response.

The KEMAR ear drum pressure was measured by the authors at an elevation of 0° for a range of azimuth angles from 0° to 40° for the ear nearest to the sound source, mimicking the range of angles one would experience with small head movements and assuming timbre assessment is dominated by the ear nearest the speaker. The KEMAR system was fitted with the same ear simulator as was used for the headphone measurements to ensure matching high-frequency sensitivity. The sound source was a 100-mm-diameter speaker, placed 1 m from the KEMAR, and time-windowing was used to approximate an anechoic measurement.

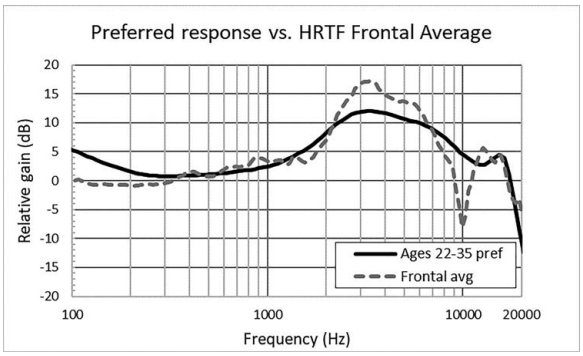


Fig. 19. Preferred curve (bold) is similar to KEMAR 0° to 40° incidence measurements (dashed).

With the exception of the bump at 3 kHz and the notch at 10 kHz, the preferred curve is remarkably similar to the KEMAR data. Listeners may prefer headphones that mimic what they hear at high frequencies from loudspeakers, though more research is needed to confirm this. The differences near 3 kHz and 10 kHz in the two curves warrant further study. Because it is possible to obtain a variety of curves by adjusting the range of angles included in the average and other measurement details, this similarity should be taken only as potential support for the measured data, not as proof of an underlying principal.

### 4.2 Bias

Although many sources of bias were removed in this testing, there were some that could not be avoided. Increasing the level of the high-frequency information raises several bias issues: context bias, loudness bias, and annoyance bias, with context bias likely being the greatest risk.

Context bias is related to the range of choices that were provided. Subjects were likely to form a mental average of the choices provided and make that their reference. Although the choices included the original Harman curve and a low-pass-filtered anchor curve, there were more curves with boost than attenuation. In addition, the curves presented to the groups having higher hearing loss had more boost than curves presented to the group without hearing loss. This could have provided an upward bias for the greater-hearing-loss groups. An analysis of the results showed that the average boost preferred by each age group exceeded the average of the choices presented to each group, even after excluding the reference and anchor choices. Therefore, context bias was considered unlikely to be a strong factor.

Loudness bias occurs because subjects often prefer louder signals in listening tests. The effect was small in this test, because only one of ten octaves of audio content was affected by the equalization changes. The perceived loudness of the most strongly equalized stimulus measured only 2.5 dB higher than the Harman reference stimulus, despite the 27 dB of high-frequency gain. Loudness was determined by calculating the loudness of an equivalent diffuse field signal in a room that would produce the same signal at the ear as an earphone matching the Harman curve.

Annoyance bias potentially has the opposite effect. Several subjects felt that 80 dB SPL was louder than what they preferred for listening. Those subjects may have chosen less boost simply to reduce the annoyance. The annoyance indicated by these listeners was not great; this seemed to also be a small risk.

Another potential source of error was the limited range of equalization choices that were presented. Some subjects may have preferred boost beyond what was available, bringing group averages downward. This was most likely to impact group of subjects with the greatest hearing loss.

The issues of context bias and limited range of choices could be avoided in future research by using the method of adjustment in place of the MUSHRA method [24]. In a method of adjustment study, the starting point of each trial is random, avoiding context bias. This method also avoids range limitations, gain quantization, and the need for grouping subjects.

### 4.3 Variability

The test results showed large variability. This can be attributed to several sources, including lack of training in subjective listening testing, differences in anatomy, and lack of familiarity with the music.

A known source of variability is the lack of training. Nearly all of the test subjects were new to listener preference testing, increasing their variability. Subjective tests that rely on an internal reference rather than on comparison with a known good reference are quite difficult and are more likely to have variation. Subjects self-reporting a low experience level showed a larger within-subject standard deviation between repeated trials (Fig. 20). The standard deviation for inexperienced subjects varied from 0 to 10 dB, with an average value of about 3 dB. This accounts for most of the variation seen in the testing.

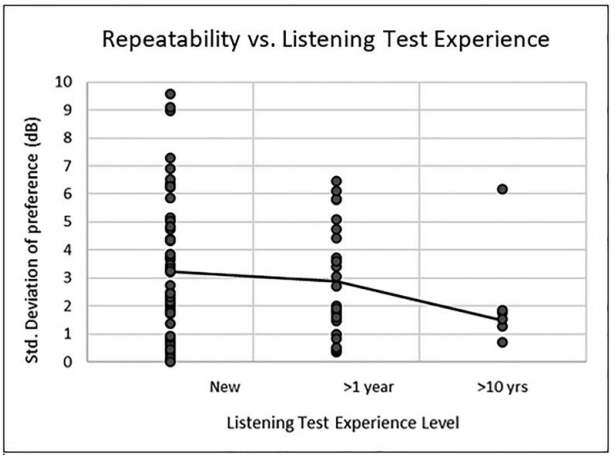


Fig. 20. Variation generally decreased with self-reported experience. Black line shows the mean of the variability for each group.

Human anatomy is another known source of variability. Variations in pinna and ear canal shapes alter the sound arriving at the eardrum both with and without headphones. Therefore, each subject may need a different headphone response to properly replicate their open ear listening experience.

Familiarity was probably the smallest source of variability. Some subjects may have been unfamiliar with the style of music provided, or not have had much experience with high-quality music playback systems. This would cause the subjects to have a more poorly defined internal reference for judging the headphones. The use of familiar instruments within the music was used to guard against this but may not have been sufficient.

## 5 DISCUSSION

Because of the limited nature of testing, the preferences shown in this test have only been confirmed for a narrow range of conditions: people living in the United States, listening to popular music with insert headphones, sealed tightly to the ear, at a listening level of 80 dB SPL, in quiet surroundings, for durations less than an hour. However, the results are expected to be applicable to a much broader range of headphone usage.

A similar response curve is also likely to be preferred for over-the-ear or on-the-ear headphones, provided the response is measured at the eardrum position. That assures that the same sound would be heard for all types of earphones. This appears to match the experience of Olive. In an article in which Olive summarized much of his headphone research [19], the same reference curve was used to predict quality ratings of around-the-ear, on-ear, and-in-ear headphones, implying that Olive found the same curve to be useful for all three types of earphones, aside from some differences in the level of bass.

The effect of music genre is expected to be small. The key features of the preferred response curve can be explained by room acoustics, human anatomy, and human perception, none of which are tied to the style of music. The effect of nationality is also expected to be small, based on research by Olive et al. [2,24].

On the other hand, presentation level is expected to alter listener preferences, especially for those having elevated (impaired) hearing thresholds. Changes in listening level affect the perceived loudness of low and high frequencies more strongly than middle frequencies due to hearing thresholds, as is described in the equal loudness curves in ISO 226 [12]. Users are expected to prefer a greater amount of high-frequency boost than chosen in this experiment when listening at low levels, especially if they have elevated hearing thresholds. Conversely, a smaller amount of high-frequency boost may be desired at elevated listening levels. A dynamically adjusted equalization could accommodate a wide range of music levels.

## 6 CONCLUSION

A set of target curves for insert headphones has been presented, based on subjects' preferences in a double-blind listening test. The target curves are an extension of previous work by Olive et al., providing additional boost with a 16-kHz peaking filter. The preferred curve for subjects with no hearing loss has a relatively flat response above 10 kHz, at a level similar to that in the mid band of the response. Subjects with typical age-related hearing loss preferred greater high-frequency energy, increasing roughly 3 dB at 16 kHz for each 10 years over the age of 30. Because hearing loss varies greatly within an age group, hearing thresholds measured above 10 kHz would be a more effective predictor of preferences than their age. Even testing a single frequency above 10 kHz provides useful information to select a response curve. The preferred amount of high-frequency energy varied over a range of $\pm 8$ dB among users with similar hearing loss, so it would be helpful to provide user control over the high frequency response of the headphone.

The large variation in user responses show that further research is needed. A larger and more diverse pool of subjects, perhaps with greater training, would help confirm the validity of the suggested curves. A method of adjustment test is suggested for further experiments, to avoid some limitations found using the MUSHRA test.

## 7 ACKNOWLEDGMENT

The authors thank the subjects who participated in the listening tests, Shehab Albahri and Knowles for supporting this research, Jordan Williams for test administration, Emily Dorne for editing, and Sean Olive and Todd Welti for their advice.

## 8 REFERENCES

[1] S. E. 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 Convention of the Audio Engineering Society* (2017 Oct.), paper 9878.

[2] S. E. Olive, T. Welti, and E. McMullin, "The Influence of Listeners' Experience, Age, and Culture on Headphone Sound Quality Preferences," presented at the *137th Convention of the Audio Engineering Society* (2014 Oct.), paper 9177.

[3] T. Miller and C. Downey, "A Wideband Target Response Curve for Insert Earphones," presented at the *153rd Convention of the Audio Engineering Society* (2022 Oct.), paper 10615.

[4] H. Møller, C. Jensen, D. Hammershøi, and M. F. Sørensen, "Design Criteria for Headphones," *J. Audio Eng. Soc.*, vol. 43, no. 4, pp. 218–232 (1995 Apr.).

[5] C. Struck and S. Temme, "Headphone Response Target Equalization Trade-offs and Limitations," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9374.

[6] J. R. Sank, "Improved Real-Ear Tests for Stereophones," *J. Audio Eng. Soc.*, vol. 28, no. 4, pp. 206–218 (1980 Apr.).

[7] C. Struck, "Free Plus Diffuse Sound Field Target Earphone Response Derived From Classical Room Acoustics Theory," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), paper 8993.

[8] F. Fleischmann, A. Silzle, and J. Plogsties, "Identification and Evaluation of Target Curves for Headphones," presented at the *133rd Convention of the Audio Engineering Society* (2012 Oct.), paper 8740.

[9] S. E. Olive, T. Welti, and E. McMullin, "Listener Preference For Different Headphone Target Response Curves," presented at the *134th Convention of the Audio Engineering Society* (2013 May), paper 8867.

[10] A. Elowsson and A. Friberg, "Long-Term Average Spectrum in Popular Music and its Relation to the Level of the Percussion," presented at the *142th Convention of the Audio Engineering Society* (2017 May), paper 9762.

[11] P. Pestana, Z. Ma, J. Reiss, A. Barbosa, and D. Black, "Spectral Characteristics of Popular Commercial Recordings 1950-2010," presented at the *135th Convention of the Audio Engineering Society* (2013 Oct.), paper 8960.

[12] ISO, "*Acoustics — Normal Equal-Loudness-Level Contours*," *Standard 226* (2003 Aug.).

[13] J. Lee, S. Dhar, R. Abel, et al., "Behavioral Hearing Thresholds Between 0.125 and 20 kHz Using Depth-Compensated Ear Simulator Calibration," *Ear Hear.*, vol. 33, no. 3, pp. 315–329 (2012 May/Jun.). http://dx.doi.org/10.1097/AUD.0b013e31823d7917.

[14] IEC, "*Sound System Equipment - Part 7: Headphones and Earphones*," *International Standard 60268-7* (2010 Jan.).

[15] A. Christensen, W. Hess, A. Silzle, and D. Hammershøi, "Magnitude and Phase Response Measurement of Headphones at the Eardrum," in *Proceedings of the AES 51st International Conference: Loudspeakers and Headphones* (2013 Aug.), paper 3-3.

[16] IEC, "*Electroacoustics - Simulators of Human Head and Ear - Part 4: Occluded-Ear Simulator for the Measurement of Earphones Coupled to the Ear by Means of Ear Inserts*," *International Standard 60318-4* (2010 Jan.).

[17] M. Wille, "High Resolution Ear Simulator," White Paper, GRAS Sound & Vibration (2017 Oct.). https://www.grasacoustics.com/files/783-High%20Resolution%20Ear%20Simulator.pdf.

[18] ITU-R, "*Artificial Ears*," *Recommendation ITU-R P.57* (2021 Jun.).

[19] S. E. Olive, "The Perception and Measurement of Headphone Sound Quality: What do Listeners Prefer?" *Acoust. Today*, vol. 18, no. 1, pp. 58–66 (2022 Spring). http://dx.doi.org/10.1121/AT.2022.18.1.58.

[20] L. Parrish, *The Effect of Age, Noise Level, and Frequency on Loudness Matching Functions of Normal Hearing Listeners with Noise Masking*, Master's thesis, Brigham Young University, Utah, USA (2016 Feb.).

[21] S. E. Olive, T. Welti, and E. McMullin, "A Virtual Headphone Listening Test Methodology," in *Proceedings of the AES 51st International Conference: Loudspeakers and Headphones* (2013 Aug.), paper 3-5.

[22] ITU-R, "*Method for the Subjective Assessment of Intermediate Quality of Coding Systems*," *Recommendation ITU-R BS.1534-1* (2003 Jan.).

[23] Institute of Sound Recording, "MUSHRA-MaxMSP," http://github.com/IoSR-Surrey/MUSHRA-MaxMSP (accessed Sep. 14, 2022).

[24] S. E. Olive and T. Welti, "Factors That Influence Listeners' Preferred Bass and Treble Balance in Headphones," presented at the *139th Convention of the Audio Engineering Society* (2015 Oct.), paper 9382.

## THE AUTHORS

Thomas Miller

Cristina Downey

Thomas Miller is an Engineering Fellow at Knowles Electronics, LLC, where he has been working since 1995. He received his BS in Electrical Engineering in 1979 from the University of Michigan. His research interest is in applications of transducers in headphones and hearing aids. He is a Fellow of the AES and a member of the ASA.

•

Cristina Downey currently works as an electroacoustic engineer at Knowles Electronics near Chicago, IL. She received her B.Sc. in Metallurgical and Materials Engineering from the Colorado School of Mines in 2015, M.F.A. in Recording Arts and Technologies from Middle Tennessee State University in 2018, and M.Sc. in Acoustics from the Pennsylvania State University in 2020. Her research interests include psychoacoustics, transducers, and digital signal processing for audio and music applications. In her free time, she also works with artists as a producer, recording engineer, and mix engineer.