Audio Engineering Society

Convention Paper 9658

Presented at the 141st Convention
2016 September 29–October 2 Los Angeles, USA

Validation of a Virtual In-ear Headphone Listening Test Method

Todd Welti[1], Sean E. Olive[2], and Omid Khonsaripour[3]

Harman International, Northridge, California, USA

Email: [1] todd.welti@harman.com, [2] sean.olive@harman.com, [3] omid.khonsaripour@harman.com

Correspondence should be addressed to Todd Welti (todd.welti@harman.com)

ABSTRACT

Controlled, comparative double blind listening tests on different in-ear (IE) headphones are logistically challenging to conduct. One solution is to present listeners with virtualized versions of the headphones through a high quality IE replicator headphone equalized to match their measured frequency responses. To test the accuracy of the virtual headphone method, ten trained listeners evaluated the overall sound quality of both the actual and virtualized versions of twelve different IE headphones that were binaurally recorded on a standard coupler and reproduced through a calibrated replicator headphone. The results show the different models of headphones produced the main effect on perceived sound quality. The virtualized headphones were essentially rated the same as the actual headphones: the agreement in terms of Pearson correlation was r = 0.98.

1 Introduction

Scientific listening tests on headphones are difficult to conduct owing to the challenges in controlling listening test nuisance variables and their inherent biases. They include sighted and tactile biases, leakage effects, consistent fit across listeners, and auditory memory loss accrued from large time gaps between comparisons between different headphones. This is particularly challenging for in-ear (IE) headphones where a good seal is needed for good bass performance below 500 Hz [1].

Several authors have proposed different solutions for this problem including the auralization of the different headphones via binaural recording and reproduction of the different headphones. A compensated neutral replicator headphone with a consistent fit and seal is required so that it doesn't add its own distortions to the reproductions. In this way, listeners can make instantaneous comparisons among the different headphones in a controlled and double blind fashion. Hiroven at al. [2] employed this auralization method to evaluate six different headphones (four circumaural and two intra-concha models) using narrow and wide-band speech. The results were compared to sighted evaluations of the actual headphones to determine how accurate the auralizations were. While there was generally good agreement between the different test methods, there were notable differences, which they attributed to sighted biases, and errors related to headphone leakage and variations in fit during the recording and reproduction stages. Since the evaluations of the actual headphones did not control sighted-tactile biases and leakage effects, it was not possible to precisely determine how accurate the auralizations were.

Brielle and Voinier [3] chose a different evaluation approach where different circumaural headphones

were evaluated through a high quality replicator headphone equalized to match the measured frequency response of the headphones. Rämö and Välimäki [4] described a digital signal processing framework for implementing the virtual headphone including simulating the headphones isolation capabilities in noisy environments. This virtualization approach eliminates the need to make and edit recordings of every headphone and music program, an advantage that saves significant time and effort. In contrast to the auralization approach, the virtual headphone method requires each headphone to be measured only once to quantify its magnitude and phase response, and then real-time digital filters are applied to the replicator headphone to simulate their responses. The benefit is that any number of virtualized headphones can be compared immediately with any arbitrary test signal without the need to make additional recordings for each headphone.

Olive et al. [5] used a similar virtual headphone method to evaluate six different circumaural headphones that were evaluated using both actual and virtual headphones. The correlation between the listeners' preference ratings of the actual versus virtualized headphones was overall good ($r = 0.85$). However, there were some noted discrepancies between results for certain listeners and for certain headphone models. They attributed these errors to variance in headphone fit and seal (leakage) across subjects. Some of the discrepancies they suspected were related to tactile biases in the actual headphone tests where subjects could recognize the headphones by their weight and pressure on their ears.

Together these studies highlight the importance of removing sighted and tactile biases in order to achieve reliable and meaningful subjective evaluations of headphones. This can only be done by either auralizing or virtualizing the headphone as described above. While the impulse response based virtual headphone method is preferable for its superior speed and flexibility, its disadvantage is that it doesn't capture any nonlinear distortion in the headphone like recordings do. Whether these distortions are audible at typical playback levels and significantly contribute to the overall sound quality is a matter of debate. A recent investigation into circumaural headphones found that unless the distortion is quite high, it has little influence on the overall headphone preference rating [6]. However, to date there have not been any investigations into the influence of nonlinear and phase distortions on perceived sound quality of IE headphone and whether they limit the validity of their virtualization. That research question is the motivation behind this study.

In the current study, we report the results of well-controlled double blind listening tests where ten trained listeners evaluated the sound quality of twelve models of IE headphones. To validate the accuracy of the virtual headphone method, listeners rated recordings of both the actual headphones and the virtualized versions reproduced through a high quality, low distortion replicator headphone that was compensated to have a flat magnitude response as measured on a standard acoustical coupler. Leakage effects in both the recording and playback were eliminated by monitoring the signal in the coupler and inside the ear canal via a Microelectro-mechanical microphone (MEM) attached to the playback headphone as described in [7].

The recordings of the actual headphones include the magnitude, phase and nonlinear distortions while the virtualized versions only capture the magnitude and minimum phase response. Therefore, by comparing the sound quality ratings of the actual versus virtualized headphone, one can assess the audibility and influence of nonlinear and phase distortions on perceived sound quality.

This paper is organized as follows: section two describes virtual headphone method including the selection, recording, playback and equalization of the headphones. Section three describes the listening test method with the results discussed in section four. Section five summarizes and discusses the results with conclusions presented in section six.

2 Virtual Headphone Method

2.1 Selection of Headphones

A total of twelve different models of IE headphones were selected for the purpose of this study. Table 1 summarizes the brand, model and approximate street price of each model tested.

Table 1: Details on the Headphone Sample

Model	Brand	Approximate Street Price (USD)
AKG	K3003i	$1000.
AKG	Y23	$40.
AKG	K323 XS	$59
AKG	Y20	$24
Klipsch	S4	$60.
Panasonic	RP-TCM125	$ 11.
Panasonic	RP-HJE120	$ 6.
Polk Audio	AM5110A	$ 90.
RHA	MA 750i	$129.95
Sennheiser	Momentum	$ 100.
Shure	SE215	$99.
Sony	MDR 7550	$ 230.

The test sample included IE headphones from eight different manufacturers covering a price range from $6 to $1000, with the average price around $160.

2.2 Selection of Replicator Headphone

The Sennheiser Momentum was selected as the replicator headphone based on its measured smooth and extended frequency response and low distortion, and reliable seal. A large unobstructed sound port allowed us to mount a miniature MEM microphone (Knowles SPV1840 LR5HB) inside the left and right sound port so that the signal could be measured inside the listeners' ear to check for leakage (see Fig. 1). The presence of the microphone had an insignificant effect on its response. Together these features made the Sennheiser Momentum an excellent candidate for replicating the measured responses of the other headphones.

Figure 1 The modified replicator headphone shown modified with a small MEMS microphone attached to monitor and control leakage effects in the listeners' ear.

2.3 Headphone Virtualization Process

For each IE headphone the magnitude and phase response was measured for both left and right channels with a G.R.A.S. RA0045 externally polarized ear simulator according to IEC 60318-4 [8]. The Harman Audio Test System was used to generate a log sweep and measure the magnitude and phase response with 1/48-octave resolution. Each headphone was inserted into the GR0408 ear canal extender coupler until a good seal was attained. Generally, this put the tip of the headphone port approximately 2-4 mm from the coupler reference plane.

Virtualization was accomplished using the following general steps (refer to Fig. 2):

1. Multiple measurements were made of the target and replicator headphones to check for repeatability. These coupler measurements were normally quite repeatable, so averaging was not needed.
2. Measured responses of the replicator headphone were regularized using a non linear equation that smoothed high Q dips. For IE headphones the effect is minimal.
3. The target headphone response magnitude was divided by replicator headphone response magnitude.

4. Equalization was only attempted within the range of 20 Hz to 15 kHz. Beyond those limits, the virtualization filter was normalized such that it tended towards 0 dB gain.
5. A complex (minimum phase) response was generated using cepstral processing (Matlab rceps())
6. Resulting impulse response was exported as a WAV file FIR filter for use in Max/MSP virtualizer. The virtualization filter is normalized such that it tends to 0 dB gain.

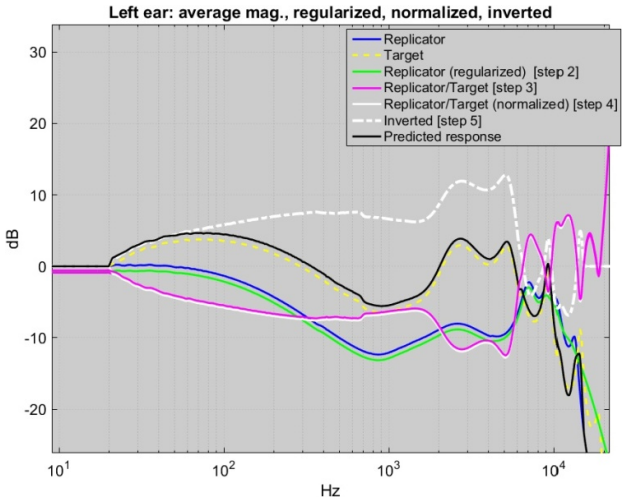


Figure 2 Example showing steps to generate virtualization filter for IE headphone.

Appendix 1 shows the measured magnitude response of the actual and virtual headphones based on the average of the left and right channels. There is a very good agreement between the two up to 10 kHz above which we did not attempt to aggressively equalize the headphone due to uncertainty in the measurements, headphone positioning in the ears, and the limited output of the replicator headphone above 12 kHz. Appendix 2 shows a typical example of the phase response of a headphone where there is good agreement between the actual and virtual headphone, and a worst-case example where the agreement is less good above 4 kHz.

2.4 Recordings of Real and Virtual Headphones

Binaural recordings were made of both the real and virtualized headphones by inserting them into the same G.R.A.S coupler and recording the two music loops used in the listening experiment. The headphone outputs were adjusted to produce the same average SPL (82 dB, C-weighted), which was maintained during playback. Care was taken to ensure that no external noise was recorded and that the recording chain was not overloaded or distorted. The recordings were edited to exactly the same length for each headphone.

2.5 Reproduction of Headphone Recordings

Since the recordings contained the transfer function of the actual and virtualized headphones including the coupler transfer function, it was necessary to compensate or equalize the replicator headphone so its transfer function was a flat response at the eardrum reference point (DRP). This was accomplished using a similar process as outlined above in steps 1-6, except that the "target" in this case was a flat magnitude response. For each subject, the seal of the IE replicator headphones was checked at the beginning and end of the listening test to ensure there was no leakage. This was done by measuring the in-ear frequency response of the headphone via the MEM microphone, and adjusting the fit until its response below 500 Hz matched the response measured in the coupler.

2.6 Harman Reference IE Target Response

An IE headphone target response based on some preliminary research described in [7] was included as a hidden reference in the listening test. The target response was based on the preferred target response for a circumaural headphones described in a previous paper [9]. Starting with this preferred target response, ten trained listeners adjusted the gain and frequency parameters of the low frequency shelving filter based on preferred sound. The preferred high frequency shelving filter parameters were defined in follow up (unpublished) study.

3 Listening Experiment

3.1 Listening Test Design

A listening test was designed where the twelve headphones listed in Table 1 were evaluated in two separate tests. In each test, five different headphones were evaluated with the same hidden anchor and reference headphone included in both tests. The headphones tested in each test are identified in Table 2 with a letter A through J in order to remove the identities of the products from the test results. The low hidden anchor was deemed to be poorest product among the headphone selected based on objective measurements and an informal listening test. The hidden reference was the Sennheiser Momentum equalized to a target response determined through experimentation as described in section 2.6.

Table 2: Headphones Evaluated in Each Test

Test One: Headphone	Test Two: Headphone
A	F
B	G
C	H
D	I
E	J
Hidden Low Anchor	Hidden Low Anchor
Hidden Reference	Hidden Reference

A multiple comparison test protocol similar but not identical to ITU-R BS. 1534-3 [10] (a.k.a. MUSHRA) was used that hallowed listeners to randomly access any of the seven headphones in each test and rate each one using a 100-point sound quality scale. The most notable departure from the MUSHRA standard was no explicit unhidden reference was employed. In codec testing the choice of a reference is obvious (the original unimpaired signal). However, for loudspeaker or headphone testing, there is no obvious reference, and including one would be controversial and likely introduce bias. Both Tests One and Two consisted two sessions that were repeated several days apart to measure the reliability of the responses. In each session, listeners completed four trials using two music programs discussed in section 3.3. Half of trials in each session included the recorded virtual headphones, while the other half were the recorded actual headphones.

3.2 Listening Test Software

A custom listening test software application was written in MAX/MSP [11] to administer the tests. The graphical user interface is shown in Fig. 3. The listener could switch among the stimuli in random order using the buttons A through G and adjust their ratings using the sliders. Semantic definitions were given for the odd-intervals of the scale to provide meaning to the ratings. The presentation order of the stimuli was randomized in each trial. Listeners were instructed to rate the headphones based on the overall perceived sound quality considering attributes related to timbre, spatial and distortion attributes.

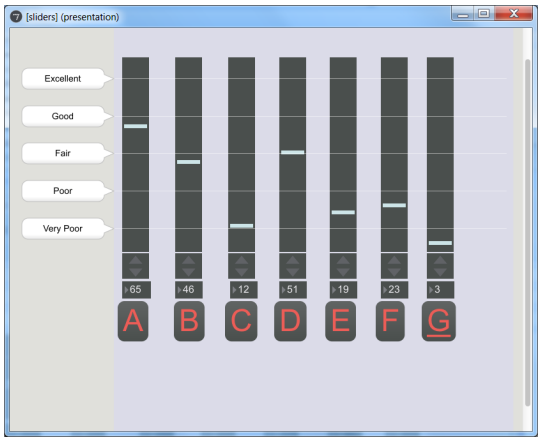


Figure 3 The graphical user interface of the listening test software.

3.3 Selection of Program

Listeners evaluated the headphones using two tracks of popular music that included female and male vocals. Details of the tracks are summarized in Table 3. Both tracks were sourced from compact disc and digitally copied and edited into 20-25 s loops.

The tracks were well recorded with spectral content that is dense and extended across the audio bandwidth as shown in Fig. 4. This facilitates listeners in identifying resonances and spectral imbalances across a wide range of the headphones.

Table 3: Details on the listening test programs.

Program/Artist/Track/ Album	Description
JW – Jennifer Warnes / Bird on a Wire / BMG Records, 1989, B00000DN6J	Female Pop Vocal
SD - Steely Dan / Cousin Dupree/ Two Against Nature / Giant Records/WEA, 2000, B00004GOXS	Male Pop Vocal

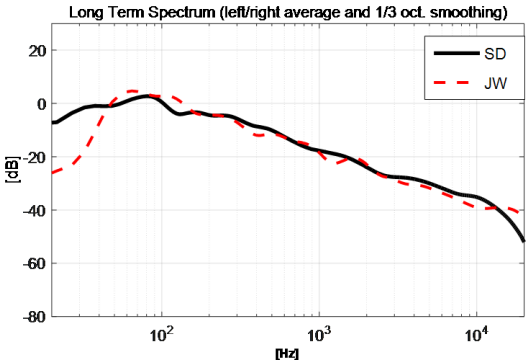


Figure 4 The long-term spectrum of the programs SD and JW averaged across left and right channels and 1/3-octave smoothed.

3.4 Selection of Listeners

A total of ten trained listeners participated in the tests, all employees of Harman International. The panel included listeners ranging in age from 22 to 56 years old (median: 33 years, SD = 3 years). All listeners had normal audiometric hearing and were considered trained based on their participation in formal listening tests and completing level eight or higher in all r training tasks defined in the listener training software, "Harman How to Listen" [12].

3.5 Absolute and Relative Playback Levels

The headphones were adjusted for equal loudness according to ITU-R 1770.3 [13]. The average playback level was 82 dB, C-weighted.

3.6 Listening Test Equipment

The test software ran on an Apple Mac Mini (2.3 GHz Intel Core i7 running OSX 10.8.5). The sound files (16-bit/ 44.1 kHz WAV files) were digitally sent to a Benchmark DAC2 USB headphone amplifier. The output of the headphone amplifier was sent to a passive switch box to allow the signal to be sent to the replicator headphone or the Harman Audio Test System used to measure the headphone inserted into the listeners' ears.

4 Results

4.1 Statistical Analysis

The listening test results were analysed using a 7 x 2 x 2 x 2 repeated-measures analysis of variance (ANOVA) model where the independent fixed factors were Headphone (7 levels), Program (2 levels), Session (2 levels), and Test Method (2 levels: Actual and Virtualized Headphone). The dependent variable was sound quality rating. Separate analyses were done for Headphone Tests One and Two. All statistical tests were performed at a significance level of 0.05.

For Headphone Test One, the main effect was Headphone; $F(6,9) = 162.77$, $p < 0.0001$. None of the other factors including Test Method were statistically significant. There was significant interaction between the factors Program and Headphone.

For Headphone Test Two, the main effect was also Headphone; $F(6,9) = 59.0$, $p < 0.0001$. There were no other significant effects or interactions between the other factors.

To summarize, the headphones were the only significant factor on listeners' sound quality ratings. The test method (virtual versus real headphone

presentations) had no significant effect on how listeners rated the sound quality of the headphones.

4.2 Headphone Effect

The mean sound quality ratings and 95% confidence intervals are shown for Headphone in Figs. 5 and 6 for the Tests One and Two, respectively. The ratings are averaged across all subjects, programs and sessions. The range of mean ratings across both tests ranged from a highest grade of 59 (Fair-to-Good) for Headphone A to the lowest grade of 7 (Very Poor) given to the low anchor headphone. The ratings for the hidden reference headphone ranged from 55 to 42. On average, listeners were not overly generous in giving any of the headphones high ratings. Also, as to their favourite headphone, the listening panel was about evenly split between Headphone A and the hidden Reference based on their sound quality, and this split explains the larger confidence intervals. In contrast, as to their least favourite headphone, there was unanimous consensus among the listening panel: listeners rated the Anchor headphone their least favourite in both tests.

The tighter spread of headphone ratings in Test Two also suggests that headphones may have been more similar in their sound quality compared to the headphones in Test One.

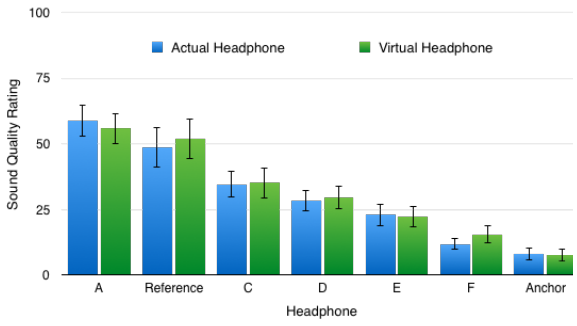


Figure 5 Mean sound quality ratings and 95% confidence intervals for Headphone Test One.

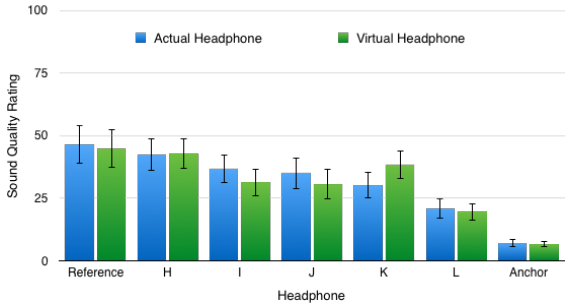


Figure 6 Mean sound quality ratings and 95% confidence intervals for IE headphones in Test Two.

4.3 Lack of Effect From Headphone Method

The main purpose of this study was to validate the accuracy and use of virtual headphone listening tests by comparing sound quality ratings of virtualized headphones versus the actual headphones. It was necessary to test using auralizations to control nuisance variables (e.g. sighted and tactile biases, leakage, etc.). From a statistical standpoint, there was no evidence found that the virtual headphones produced significant differences in ratings compared to the actual headphones in Test One; Headphone Method; $F(1,9) = 0.082$, $p = 0.77$), or Test Two; Headphone Method $F(1,9) = 0.09$, $p = 0.78$.

Figs. 5 and 6 graphically illustrate the good agreement in sound quality ratings between the virtual and actual headphone. In cases where there are small differences in mean ratings, they are within the margin of the 95% confidence interval. Based on the Pearson product-moment correlation coefficient the agreement between actual versus virtualized headphone ratings for Tests One and Two were: $r = 0.99$ and 0.95, respectively. When calculated across both tests, the correlation between actual and virtualized headphone ratings was $r = 0.98$.

4.4 Effect of Test Method For Individual Listeners

The lack of effect due to Test Method does not preclude the possibility that for some listeners there may have been significant differences in sound

quality between the virtual and actual headphones. To explore this question, the Pearson correlation coefficient was calculated for each individual indicating the agreement between Test Methods for each test (see Fig. 7). For Test One, the correlation was very high ($r > 0.88$) across all listeners. In Test Two, the correlation was slightly lower but still above 0.8 for all listeners, although there was more inter-listener variance ranging from 0.88 (Listener 1) to 0.8 (Listener 363). The lower correlation values observed in Test Two were likely related to the different headphone sample tested, which were apparently more similar in sound quality than the test sample used in Test One. This is indicated by the tighter range of headphone ratings in Test Two, which produced a much lower Headphone F-statistic that observed in Test One.

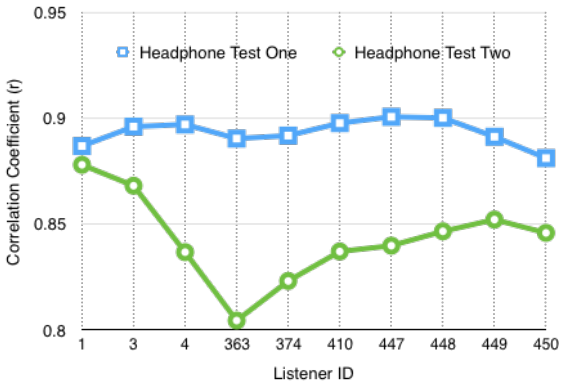


Figure 7 The Pearson correlation coefficient is shown for individual listeners based on the agreement between ratings given to actual versus virtualized headphones.

5 Summary and Discussion

In the previous section, we provided experimental evidence that can be summarized as follows:

1. The different models of headphones accounted for the main effect on listeners' sound quality ratings.
2. There were no significant differences between the sound quality ratings of the actual and virtual headphones. The agreement between the actual and virtual headphone ratings in Test One was $r = 0.99$ and $r = 0.95$ in Test Two.
3. The agreement between the actual and virtual headphone ratings was generally consistent across all listeners.

Together, these findings provide scientific evidence to support the use of virtual headphones for research and testing on the basis that they produce similar sound quality ratings as the actual headphones. Our ultimate goal is to conduct research and listening tests using real-time virtualized headphones (instead of recordings of auralizations of them) since doing so increases the efficiency, control of nuisance variables, and flexibility in IE headphone evaluations.

5.1 Subjective Versus Objective Measurements

The good agreement between the sound quality ratings of the actual and virtual headphones was not surprising given the good agreement in objective measurements (see appendices 1 and 2). The measured magnitude and phase responses of the actual and virtual headphones were very similar up to 10 kHz. The main point is that these errors had little or no impact on the perceived sound quality and ratings of the headphones.

Several authors [14,15,16] have reported that the measured frequency response of a loudspeaker is generally the best predictor of its perceived sound quality compared to its measured nonlinear distortion or phase response. While fewer similar studies exist for headphones, this one supports this idea. At the time of publication, we had not completed the nonlinear distortion measurements on the headphones, but they will be presented at a later date. Nevertheless, this study agrees with a previous study on circumaural headphones [6] where we found that nonlinear distortion had little influence on their perceived sound quality, except when the playback levels were very high.

While this study provides strong evidence to support the use of virtualized headphone listening tests there are several limitations in the study that make it

problematic to generalize the conclusions beyond the conditions we tested.

5.2 Limitations in Headphone Sample Size

This study only tested a small sample of twelve models of IE headphones, a while they covered a broad price range from $6 to almost $1000, our sample may have overlooked models that contain significantly higher levels of phase distortion that may have produced different results. We are currently in the process of testing another larger sample of IE headphones that will address this factor.

5.3 Limitations in the Bandwidth of the Coupler and the Replicator Headphone

As stated previously, the accuracy of the virtualized and actual headphones was limited up to 12-15 kHz. There remains a question whether the listening test results would be different if the auralizations were accurate up to 20 kHz or higher. We suspect that for most listeners and music signals, the accuracy of virtualized headphone reproduction beyond 12 kHz is less important due to the decreased sensitivity in hearing.

5.4 Limitations in Playback Levels and Program Selection

This study used comfortable and safe playback levels, and only two music programs that may have limited the audibility of nonlinear and phase distortion in the headphones. Previous loudspeaker studies have shown that the audibility of nonlinear and phase distortion depends on many factors including the test signal, the frequency range where the distortion occurs, and input signal level to the device [17,18,19]. We are in the process of measuring distortion in the headphones to possibly explain why distortion didn't appear to be a factor in how listeners rated them. In terms of group delay, the threshold of audibility of phase distortion in headphones while listening to clicks has been found to be about 1.6 msec independent of the center frequency of the delayed (1, 2, or 4 kHz) [18]. At higher frequencies, and when listening to music with reverberation added, listeners were less sensitive to phase distortions. Given that most IE headphones we tested exhibited well-behaved minimum phase behavior up to 6 kHz and beyond (see appendix 2) we believe that phase distortion was not a factor.

5.5 The Control of Leakage Doesn't Reflect Real-World Use of Headphones

Leakage was monitored and carefully controlled in this study during the recording and playback of the headphone auralizations. Leakage was treated as a nuisance variable and eliminated in order to allow valid and reliable assessments of the effect of headphone method. To our knowledge, this is the first published study where listening tests on IE headphones were reported and leakage was not a factor that corrupted the results. For that reason, we believe these results are more reliable and valid than those from previous studies. It also explains why the correlations between the actual and virtualized headphones were so high and consistent between listeners.

These results highlight how important headphone fitment and leakage are when designing a headphone for optimal performance and perceived sound quality. Good performance cannot be consistently achieved without paying close attention to minimizing the effects of leakage. This also highlights a limitation of the current virtual headphone method: to be a more useful tool for headphone validation and benchmarking, it must incorporate the leakage effects measured on actual headphones and real listeners.

6 Conclusions

A controlled double blind listening experiment was conducted in which ten trained listeners evaluated the overall sound quality of twelve IE headphones binaurally recorded and reproduced through a high quality replicator IE headphone. Listeners evaluated both the actual and virtualized versions of the headphones using two wideband music programs. Based on the statistical evidence the conclusions are:

1. The different models of headphones produced the main effect on sound quality ratings.

2. The Headphone Method had no effect: there was no significant difference between the sound quality ratings of the actual versus virtualized headphones. The Pearson correlation coefficient indicating agreement between the actual and virtualized headphone ratings was r = 0.98.
3. The good agreement between the virtual and actual headphone sound quality ratings was relatively consistent across all listeners.
4. Objective measurements (magnitude and phase response) of the actual and virtualized headphones showed good agreement up 10-12 kHz. The errors were generally small and based on the listening test results did not have a significant influence on perceived sound quality. The absence or presence of nonlinear distortions in the headphones (still to be confirmed with measurements) had no significant effect given that the distortions would have been present in the actual headphones but not the virtualized versions. That fact that both actual and virtual headphones were rated similarly suggests that nonlinear distortion was not a factor in listeners' sound quality ratings.

This observation is consistent with previous loudspeaker and headphone studies where the magnitude response was found to be the most reliable predictor of perceived sound quality compared to measurements of distortion and phase [6,14,15,16]. Future research should focus on developing perceptually meaningful distortion measurements that would identify whether the headphone has audible distortions and exclude them from virtual headphone listening tests. With that caveat and the limitations outlined in section 5, this study provides evidence that the virtual headphone listening test method can produce accurate and reliable sound quality ratings similar to those produced with the actual headphones.

7 Acknowledgements

The authors would like to thank Harman International for their support of this research. We also thank all of the ten listeners who participated in the tests.

8 References

[1] Todd Welti, "Improved Measurement of Leakage Effects for Circumaural and Supra-aural Headphones", presented at the 138th Audio Eng. Convention, Warsaw, Poland (May 7-10, 2015).

[2] Toni Hirvonen, Markus Vallgamaa, Juha Backman, Matti Karjalainen, "Listening Test Methodology For Headphone Evaluation," presented at the 114th Audio Eng. Soc., Convention, preprint 5736, (March 2003).

[3] Francoise Briolle and Thierry Voinier, "Transfer Function and Subjective Quality of Headphones: Part 2, Subjective Quality." 11th International Audio Eng. Soc. Conference, (May 1992).

[4] Jussi Rämö, Vesa Välimäki, "Signal Processing Framework for Virtual Headphone Listening Tests in a Noisy Environment " presented at the132nd Audio Eng. Soc. Convention, preprint 8640, (April 2012).

[5] Sean E. Olive, Todd Welti, and Elisabeth McMullin," A Virtual Headphone Listening Test Methodology," presented at the 51st International Audio Eng. Soc. Conference, Helsinki, Finland, (August 22-24, 2013).

[6] Steve Temme, Sean Olive, Steve Tatarunis, Todd Welti, and Elisabeth McMullin," The Correlation between Distortion Audibility and Listener Preference in Headphones," presented at the 137th Convention of the Audio Eng. Soc. (October 2014).

[7] Sean E. Olive, Todd Welti, and Omid Khonsaripour, "The Preferred Low Frequency Response of In-Ear Headphones," presented at the 2016 International Audio Eng. Society Conference on Headphone Technology, (August 2016).

[8] G.R.A.S. RA0045 Externally Polarized Ear Simulator According to IEC 60318-4 (60711), http://www.gras.dk/ra0045.html (June 2016).

[9] Sean E. Olive, Todd Welti, and Elisabeth McMullin, "Listener Preferences for In-Room Loudspeaker and Headphone Target Responses," presented at the 135th Convention, Audio Eng., Soc., preprint 8994, (2013 October).

[10] International Telecommunication Union, " Recommendation ITU-R BS, 1534-3: Method for the subjective assessment of intermediate quality level of audio systems," https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1534-3-201510-I!!PDF-E.pdf. (October 2015).

[11] Cycling '74 Software, Max/MSP, https://cycling74.com, (June 2016).

[12] Harman: How to Listen, www.harmanhowtolisten.blogspot.com (June 21, 2016).

[13] Recommendation ITU-R BS, 1770-3, "Algorithms to measure audio programme loudness and true-peak audio level," https://www.itu.int/dms_pubrec/itu-r/rec/bs/R-REC-BS.1770-3-201208-S!!PDF-E.pdf, (August 2012).

[14] Wolfgang Klippel, "Assessing the Subjectively Perceived Loudspeaker Quality on the Basis of Objective Parameters," presented at the 88th Audio Eng. Soc. Convention, preprint 2929 (March 1990).

[15] Floyd E. Toole, *Sound Reproduction: The Acoustics and Psychoacoustics of Loudspeakers and Rooms*, Focal Press, (2008).

[16] Sean E. Olive, "A Multiple Regression Model for Predicting Loudspeaker Preference Using Objective Measurements: Part II - Development of the Model," presented at the 117th Audio Eng. Soc. Convention, preprint 6190 (October 2004) Paper Number: 6190

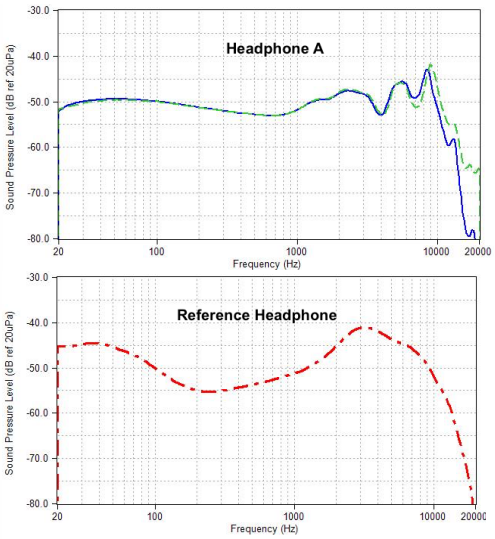
[17] Alexander Voishvillo, Alexander; Terekhov, Eugene Czerwinski, and Sergei Alexandrov, Sergei Graphing, " Interpretation, and Comparison of Results of Loudspeaker Nonlinear Distortion Measurements", J. Audio Eng. Soc., vol. 52 Issue 4 pp. 332-357; (April 2004).

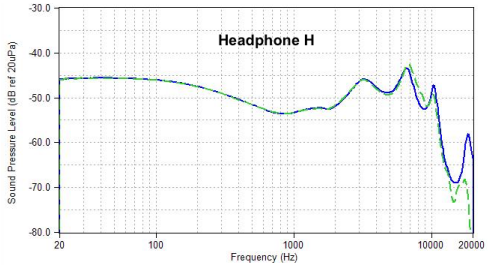
[18] Flanagan, Sheila; Moore, Brian C. J.; Stone, Michael A. Discrimination of Group Delay in Click-like Signals Presented via Headphones and Loudspeakers, JAES Volume 53 Issue 7/8 pp. 593-611; July 2005

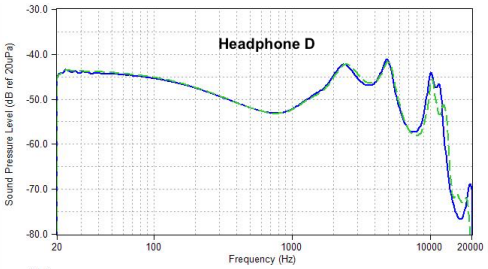
[19] Hideo Suzuki; Shigeru Morita, Takeo Shindo, "On the Perception of Phase Distortion" JAES Volume 28 Issue 9 pp. 570-574; September 1980.

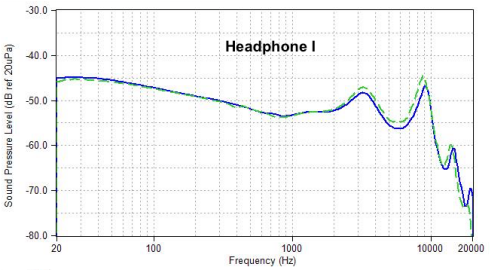
Appendix 1: Magnitude Responses of Actual and Virtual Headphones

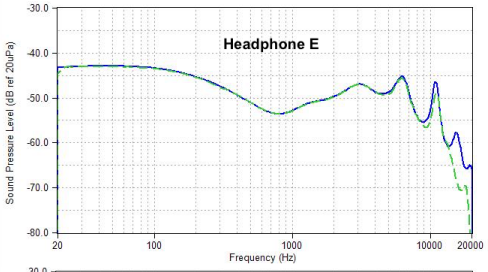
Below are the measured magnitude responses of the actual (solid curves) and virtualized (dotted curves) measured in the G.R.A.S. RA0045 externally polarized ear simulator according to IEC 60318-4. The headphones are plotted in descending order from the highest to lowest mean sound quality rating. Each curve is an average of the left and right channels of the headphone cha (highest to lowest rated sound quality). Note that only one curve is shown for the Reference Headphone (the Sennheiser equalized to a preferred target response) since it was used in both actual and virtual headphone listening trials.

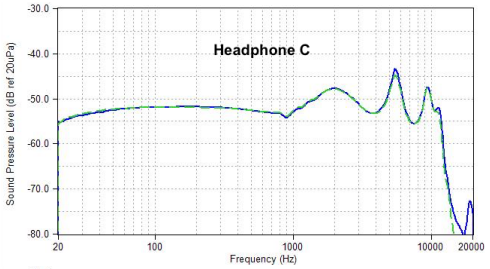


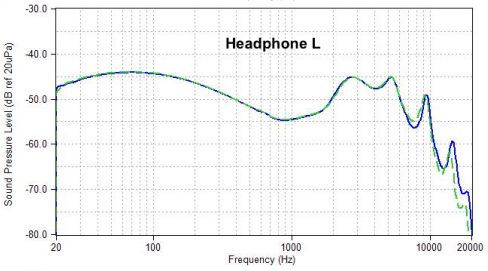

Headphone H
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)

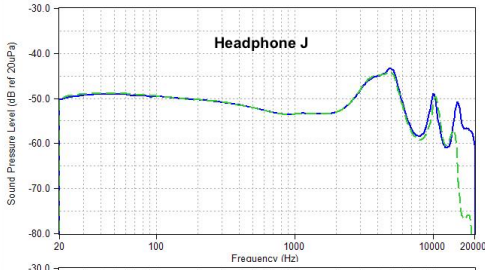

Headphone D
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)

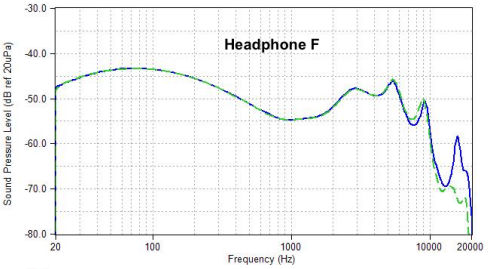

Headphone I
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)

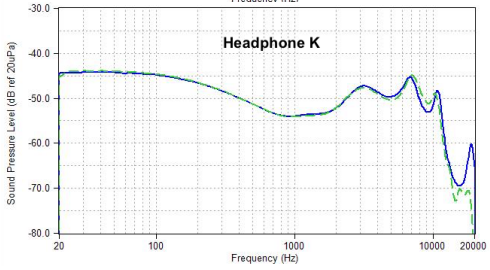

Headphone E
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)

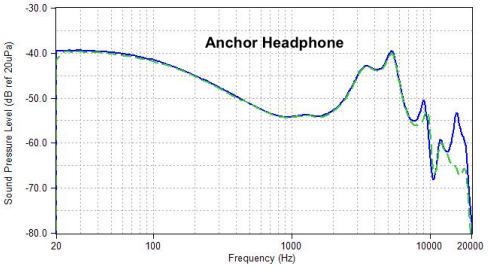

Headphone C
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)


Headphone L
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)


Headphone J
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)


Headphone F
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)


Headphone K
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)


Anchor Headphone
Sound Pressure Level (dB ref 20uPa)
-30.0
-40.0
-50.0
-60.0
-70.0
-80.0
20
100
1000
10000
20000
Frequency (Hz)

Appendix 2: Examples Of Measured Phase Response of Headphones

This section shows two examples of the measured phase response of the actual and virtualized headphones. The first example (Headphone K) shows good agreement between the phase response of the actual and virtualized headphone. This was typically the case for most headphones. The second example (Headphone I) is a worst-case example where the agreement is good up to 4 kHz at which point we see some small phase shift in the virtual headphone relative to the actual headphone. The cause of this phase error is a combination of small errors in the measurement and the filter regularization. If the magnitude responses are not quite matched, there will be a small phase error associated with it, assuming the response is minimum phase. The listening test results provide evidence these phase errors were not sufficiently large to cause audible effects on the sound quality.