HearAdvisor procedures for recording and evaluating hearing devices v 1.0

Andrew Sabin Ph.D., Steve Taddei Au.D., Abram Bailey Au.D.
Hear Advisor LLC

0.0 Introduction

It can be challenging for a hearing aid consumer to assess how well a device performs before purchasing. Similar marketing claims are used across manufacturers and it is rarely an option to audition hearing aids in advance of purchase. This is particularly problematic in the growing over-the-counter markets where an expert might not be involved in device selection.

Our goal is to help hearing aid customers by providing realistic recordings and understandable metrics that empower them to make informed hearing aid purchase decisions. In this report we describe our approach to addressing this problem via lab recordings and scientific metrics (presented on a simple 0-5 point scale).

Designing appropriate recording methods and metrics is a complex task. There is a nearly infinite set of combinations of hearing losses, environments, and device settings. To make this effort feasible, we had to make many choices (descriptions and rationales below). In all cases we attempted to leverage scientific research and to target the most likely scenarios. We tried to create realistic environments, repeatable procedures, and objective metrics. We acknowledge that this might not represent all hearing losses, environments, and settings and we welcome feedback.

1.0 Laboratory Setup

1.1 Room and Equipment

We designed and built an acoustic testing laboratory (Fig. 1) that was sufficiently quiet and non-reverberant for our tests. The walls and ceiling were filled with sound-absorbing material between studs (RockWool Safe ‘n Sound Insulation or heavy blankets) and the floor was carpeted. The resulting space had an ambient sound pressure level of 35.7 dB LAeq (A weighted) and a 4-frequency (0.5, 1, 2, and 4 kHz) average reverberation time (RT60) of 0.072 s. We used the reverberation time and room volume to estimate the critical distance (the distance at which the direct and reverberant signals have equal levels) to be 1.3 m (e.g., [1]). We installed a ring of 8 speakers (Yamaha HS5) with a radius of 1 m - thus ensuring that at the center of the ring the direct sound from the speakers dominates any room reverberation.

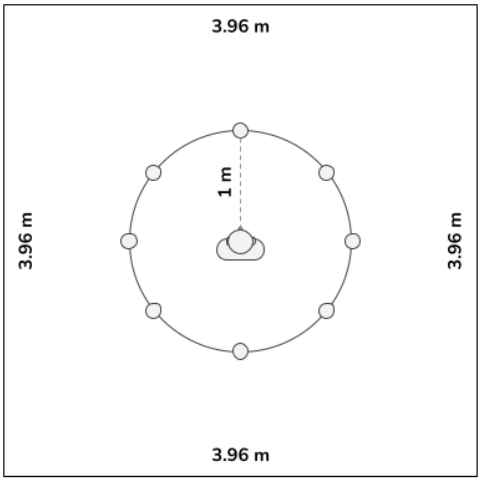

Figure 1. Test lab floorplan

Additionally, each speaker is equalized to be flat (+/- 2 dB) from 50 to 1500 Hz.

We placed an acoustic manikin (Kemar 45BA) in the center of the speaker ring. The height of its artificial pinnae was aligned with the high-frequency drivers (tweeter) of the speakers. The manikin has anthropometric pinnae, VA-style tapered ear canals, and wears a wig. Our coupler mic (Standard IEC 60318-4, 711 coupler) is accurate up to 10 kHz and accordingly we remove any energy above that frequency via downsampling to 20 kHz. The microphone outputs are digitized by a high quality audio interface (Antelope Orion Studio).

2.0 Recordings

2.1 Acoustic Scenes

We created a set of acoustic scenes that represent a wide range of environments and conversations. For backgrounds, the 12 recordings from the ARTE database [2], were decoded to the 8 channels of our 2D speaker ring (HOA Order = 3). Each background’s presentation level was adjusted to match the published values that were observed in the real environment. To ensure consistency with each background, any 5-second segment that was more than 5 dB from the average of all 5-second segments was removed.

For the speech, we recorded a custom set of conversations by hiring actors to read scripts in a recording studio. To elicit potential Lombard effects, the associated background sounds were played into the actor's headphones during recordings. For each background there were 3 scripts (1, 2, and 3 talkers). Each script was performed twice (rotating actors).

The speech recordings were combined with the backgrounds in realistic spatial locations. In one-talker scenes, the talker was placed in the 0-degree (on-axis) speaker. In two-talker scenes, the talkers were placed in the -45 and 45 degree speakers. In three-talker scenes, the talkers were placed in the -45, 0, and 45 degree speakers. In addition, to match the reverberation between the talker and background, the talkers were convolved with the multichannel impulse response from the ARTE database that was matched to the scene (without "direct sound enhancement"). When talkers were not at 0 deg azimuth, the multichannel impulse response was rotated to match the direct path to the location of the talker.

The presentation level of each talker was set to follow the relationship described in [3] between environment SPL and signal-to-noise ratio (their Fig. 3B) as observed in individuals with hearing loss. Each scene had at least 15 seconds of just background at the beginning to allow the hearing aid to adapt. Then each script was played back-to-back. In total there are 72 scenes (12 environments X 3 numbers of talkers X 2 actor variations). Average scene duration was 34.9 seconds (s.d. 4.5 sec).

2.2 Device Insertion

We attempted to insert devices in a way that creates a symmetrical fit and the appropriate acoustic seal. The tester monitored the real time insertion loss (see Section 3.4) while inserting the powered-off device. The devices were adjusted until they had the expected overall insertion loss shape for their coupling (e.g., low pass for semi-occluding) and the between ear difference was < 5 dB at 1 kHz. The resulting measurement (Real-Ear Occluded Gain) was used in our measure of occlusion (see Section 4.2). Water based lubricant was used for devices that had difficulty creating a seal on the manikin's ear.

2.3 Music Streaming

We assessed streaming music quality by playing five genres of recorded (royalty-free) music from a paired smartphone to the manikin wearing the device. On average, the segments were 33.7 seconds (s.d. 5.9 sec). The phone volume was adjusted to match a reference level using real-time spectral analysis of the eardrum mic of the manikin (see Section 3.4). That reference level was derived by presenting a custom steady noise whose spectrum was matched to that of the average across music signals via the speaker ring at 70 dB SPL (a common level [4]). The tester adjusted the streaming level from the phone until the hearing aid's level matched that of the reference curve at 1 kHz (⅓ octave filter) within 5 dB.

2.4 Post-Processing

Minimal post-processing was applied to make sure the recordings were suitable for presentation over headphones. First, we performed a diffuse field equalization to remove the acoustic effects of the manikin's anatomy from the recording. As is standard, we fit a filter to the spectral difference between Kemar's eardrum microphones and a flat reference mic. Each measurement was taken with the microphone in the center of the speaker ring, while the speakers were emitting uncorrelated white noise. The resulting filter shape largely matches the published values [5]. Some deviation is expected due to the 2D ring of speakers here vs 3D environment in their report. Finally, we needed to choose a convention of mapping dB SPL in the laboratory to dB FS in a sound file for online presentation. For this decision there is a tradeoff between clipping (when recordings are too loud) and noise or extreme volume settings needed from the playback system (when recordings are too quiet). Through trial-and-error we determined that a mapping of 0 dB FS = 100 dB SPL is a compromise that minimizes (but doesn't eliminate) clipping while allowing our scenes to be well above the noise of typical playback systems. In all online pages containing these recordings, aided recordings are placed alongside unaided ones and the users are instructed to

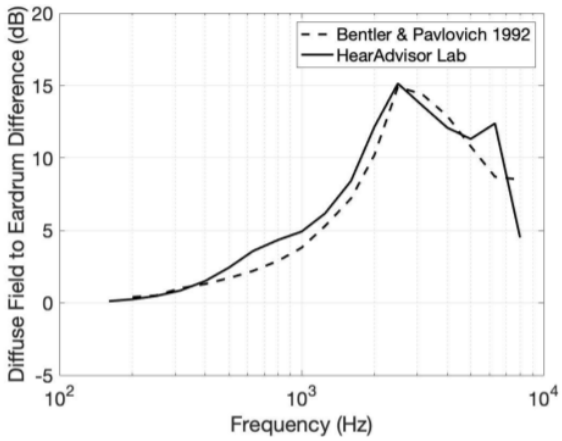

Figure 2. Diffuse-field equalization curves.

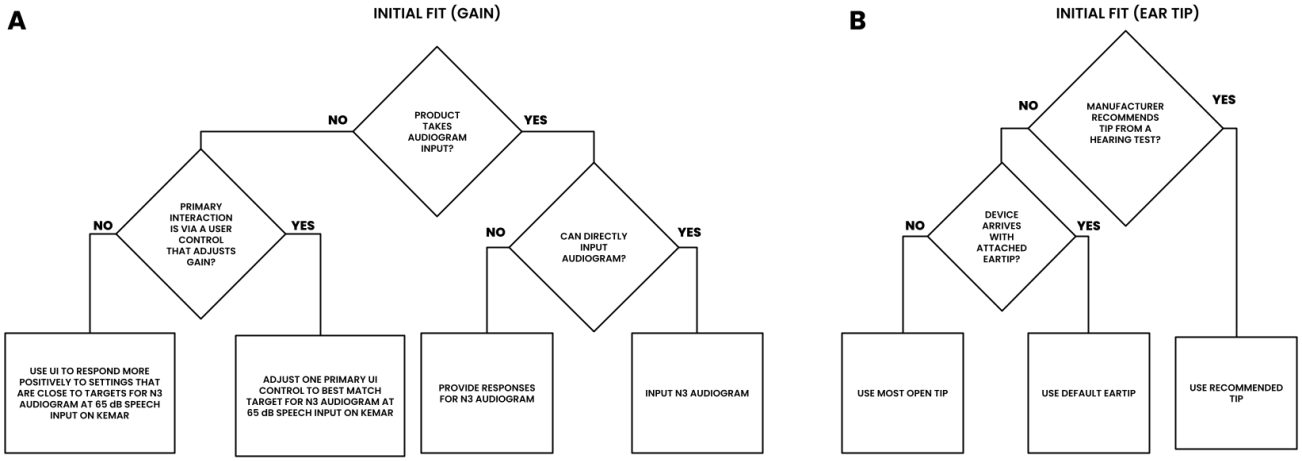

Figure 3. Fitting flow chart for initial fits: (a) Gain parameters and (b) ear tip selection.

(a) wear good headphones and (b) adjust their volume to make the unaided condition have a realistic unaided sound level.

3.0 Device Settings

3.1 Hearing Loss

We chose to target the standard sloping moderate hearing loss from [6] (N3 configuration; see Table 1). We chose this loss because (a) it is reasonably common (b) it is near the middle of the overall aided population, and (c) it is a loss that can be appropriate for either prescription or OTC devices. We approached fitting decisions by considering what would be appropriate for the average consumer with a binaural sensorineural hearing loss with the N3 configuration. Prescription targets are computed using NAL-NL2 [7] prescription with unspecified gender, non-tonal language, and prior hearing aid experience. Each hearing aid was recorded in two configurations "Initial" and "Tuned," that differ in their gain settings and ear couplings (Sections 3.2 and 3.3).

250	375	500	750	1k	1.5k	2k	3k	4k	6k
35	35	35	35	40	45	50	55	60	65

Table 1. N3 Audiogram. Values in Hz (Top) and dB HL (Bottom)

3.2 Initial Fit

Our goal for the initial fit was to approximate the settings that a user would experience if the person fitting the device just followed basic instructions. To standardize this process we came up with the flow chart shown in Fig. 3A.

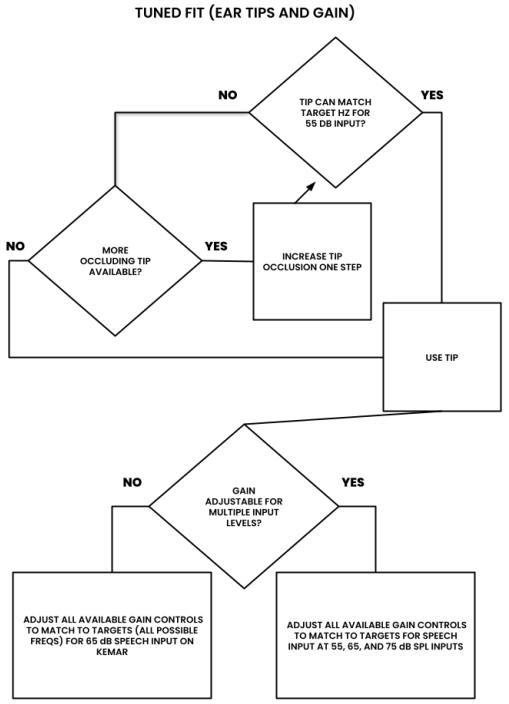

Figure 4. Fitting flow chart for tuned fits – ear tip selection and gain.

The initial node is whether the fitting software takes an audiogram input. If the audiogram can be entered digitally, a manufacturer's "first-fit" is performed - as is done in many clinics (e.g. [8]). Alternatively, if the device requires an in-situ audiogram, then the tester provides responses consistent with an N3 audiogram (see Section 3.5). If audiogram entry is not possible, then the next node is whether there is a primary control for gain (e.g., presets or volume). If so, that single control is adjusted to the closest match to the target insertion gain (see Section

3.4). If not, the tester provides software responses that favor insertion gains that are more similar to the prescription.

We also considered ear tip (acoustic coupling to ear canal) to be part of the fitting procedure since many manufacturers offer several choices. Our flow diagram for ear tip is in Fig. 3B. We reasoned that most individuals fitting the device would start with the most open ear tip to minimize occlusion. Therefore we used the most open ear tip for most initial fits. The only exception was when the fitting software prescribed an ear tip. In this case the prescribed ear tip was used.

3.3 Tuned Fit

Our goal for the Tuned fit was to approximate the settings that a user would experience if the person fitting the device performed a more thorough fitting. Specifically, the tester adjusted all available parameters and ear tips to match prescriptive targets for quiet and loud inputs. The flow chart for this fit is shown in Fig. 4. We first evaluated if the prescriptive targets could be matched with the ear tip from the initial fit. We defined a success as a four frequency (0.5, 1, 2, 4 kHz) SII-importance-weighted [9] average absolute deviation from prescription of less than 5 dB (with a 55 dB International Speech Test Signal – ISTS, [10]). For a fit to be successful it had to be stable - meaning no audible feedback. If a successful fit was not possible with the initial ear tip, it was swapped for a more occluding one. The tester iterated until either a successful fit was achieved or there were no more occluding ear tips available. Once the "Tuned" ear tip was selected, the tester adjusted all available parameters to match to targets for speech inputs at 55, 65, and 75 dB SPL.

3.4 "Real Ear" Measures

Several aspects of our initial and tuned fittings rely upon the equivalent of Real Ear Measures (REMs, e.g. [11]). We replicated this procedure on the manikin via real-time spectral analysis (⅓ octave filterbank) of the eardrum microphone. During fitting, we presented a calibrated speech signal [10] and computed the difference at the eardrum mic between the aided and unaided conditions. The resulting value (Insertion Gain) could be visually compared to the prescriptive targets via a custom interface. This interface was used (a) when performing some initial fits (without an audiogram) (b) all tuned fits, and (c) while measuring occluded response during device insertion.

3.5 In Situ-Audiograms

Several devices were fitted using an audiogram that was performed using the device itself (i.e., an in-situ audiogram). In this case, we provided responses that were consistent with the N3 hearing loss (Table 1). We performed real-time spectral analysis (⅓ octave filterbank) on the eardrum microphone during the test. The spectrum could be visually compared to the sound level equivalent to the auditory thresholds of the N3 hearing loss via a custom interface. The tester responded positively for signals above threshold and negatively for signals below threshold. We validated this approach using the Mimi app for iOS [12] using wired apple earpods (with their correction for these earbuds). The resulting audiogram was within < 5 dB of N3 at all frequencies.

3.6 Noise Settings

Finally, we also considered the settings that influence conversation in noise. The flow chart for this setting is in Fig. 5. The first node is whether the device offers automatic switching (e.g., via scene classification). If yes, then the automatic algorithm is used for all fits. If not, we chose to leave the speech-in-noise program on for all environments, given the high value hearing aid consumers place on performance in noise [e.g., 13].

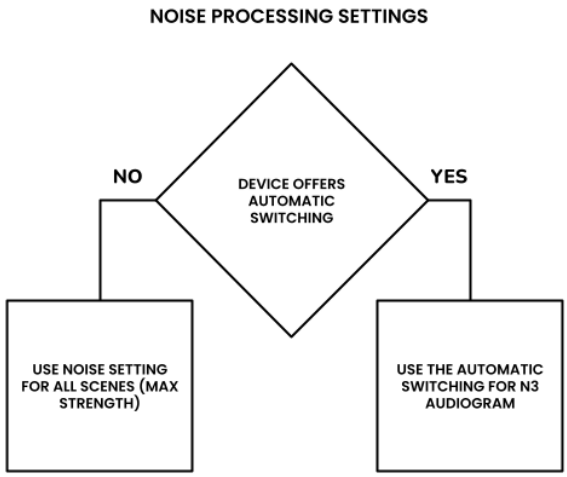

Figure 5. Flow chart for noise settings

4.0 Metrics

4.1 Speech Perception Benefit Metric

The first dimension we quantified was the speech perception benefit. We attempted to quantify the

expected improvement in speech intelligibility for each device/fit combination via acoustic measurement. For all recordings, we computed Hearing Aid Speech Perception Index v2 (HASPIv2 [14]). We chose this metric because it models the impaired auditory system and predicts intelligibility for a wide range of acoustic environments. We computed HASPIv2 using an N3 audiogram (Table 1) and RAU-transformed [15] the output. We averaged across both ears and computed the difference between unaided and aided recordings (ΔHASPIv2). We report this value separately for quiet/moderate (< 70 dB SPL) and loud (> 70 dB SPL) environments. To do so, we averaged the ΔHASPIv2 score separately for all quiet/moderate vs all loud acoustic scenes. The resulting values are mapped to our 5-point scale via linear scaling.

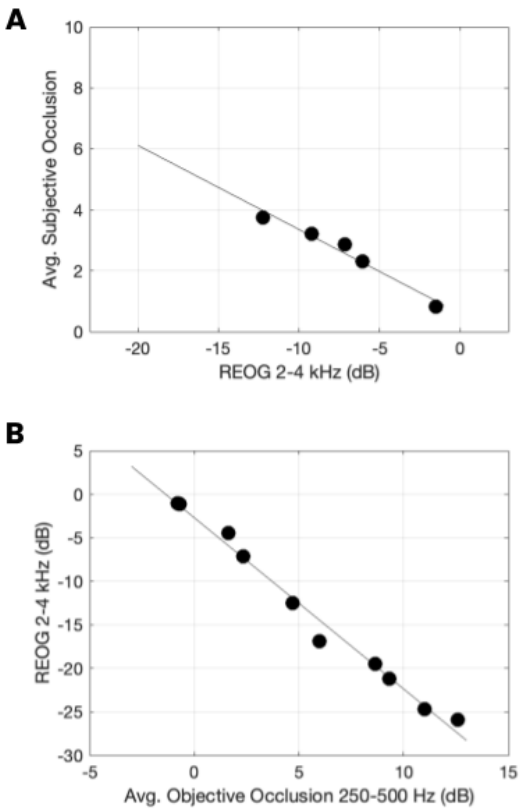

Figure 6. Occlusion Metrics. (a) Relationship between REOG and *subjective* occlusion, replotted data from [18] (b) Relationship between REOG and *objective* occlusion, replotted data from [20]

4.2 Occlusion Metric

Another dimension that we considered was own-voice quality. This is a common complaint of hearing aid users (e.g., [16]) and can be attributed to two factors: (1) occlusion which arises from bone-conducted sound becoming trapped in the ear canal and (2) amplification which arises from the receiver-transmitted sound. For these initial metrics, we chose to focus on occlusion which is the larger contributor to user complaints (e.g., [17]). Our goal was to estimate *subjective* occlusion (a user's rating of their perceived occlusion from 0-10 - as in [18]) via acoustic measurements. We chose to focus on measurement of *Real Ear Occluded Gain* (REOG – the difference in spectrum between the open ear and the aided ear with the device off). We reasoned that acoustic coupling that generates high subjective occlusion usually also generates more negative REOG (excluding deep-insertion devices). Using the data from [18], we observed that the group average REOG from 2-4 kHz for instant-fit tips had a strong linear correlation to the group average *subjective occlusion* (Fig. 6A). We used this relationship to map from REOG to estimated subjective occlusion. This procedure was used for all devices without active occlusion cancellation (AOC, e.g. [19]),

For devices with AOC, the passive REOG measurement does not capture the influence of cancellation on subjective occlusion. With this in mind, we instead chose to measure *objective occlusion* on vocalizing humans. Objective occlusion here is the difference in probe mic spectra between the open ear and the aided ear while the participant was vocalizing a sustained /i/ as in [20] ($REAR_{VOC}$ - $REUR_{VOC}$). We then wanted to use this *objective* measurement to estimate *subjective occlusion* on the same scale as the non-AOC devices. Our approach to this estimation had two steps: (1) map from *objective* occlusion to a matched REOG value, and then (2) map from that matched REOG value to *subjective* occlusion. For part (1) we replotted data from [19] (their Figs 4 and 7) and observed a strong linear correlation between group average objective occlusion in the 250-500 Hz band and the group average REOG from 2-4 kHz (Fig. 6B). For part (2) we used the matched REOG value to estimate subjective occlusion via the same relationship as in the non-AOC devices (Fig. 6A). Finally, we then converted the estimated subjective occlusion value to our own-voice metric by inverting and scaling the estimated subjective occlusion value.

4.3 Music Streaming Metric

The next dimension that we considered was the quality of streamed music. Our goal was to estimate subjective sound quality via an acoustic measurement. Accordingly we leveraged the Hearing Aid Audio Quality Index (HAAQI, [20]). This metric uses the same model of the impaired auditory system as our speech metrics [14] and was

designed to match subjective music sound quality ratings from individuals with hearing loss (from [21]). HAAQI attempts to capture the influence of both linear and nonlinear distortions. However, we observed that some primary nonlinear distortions (e.g., from streaming audio compression codecs) were not well captured by HAAQI. With this in mind, we reported only the portion of the metric that quantifies linear distortions ($HAAQI_{LIN}$). We then computed a device's average $HAAQI_{LIN}$ across both ears and all recordings of streamed music (see Section 2.3). Lastly, the resulting values were scaled to be on a 5 point scale.

4.4 Feedback Metric

The final dimension that we considered was the degree to which feedback ("squealing") was a problem - another common complaint from hearing aid users [15]. There is already some influence of the device's feedback properties in the Speech Perception Benefit score (Section 4.1), because we only considered stable settings (without audible squeals) to be appropriate. However, we additionally tested the quality of the feedback canceller in two challenging cases: (a) repeatedly moving hands near the device for 10 seconds - simulating a hairbrush motion, and (b) using a hand to cup the ear 10 times in a row. The library scenes were played in the background to ensure that the devices were amplifying the signal. We recorded from the manikin during these cases and then performed blind listening tests to subjectively rate each recording into a 0-3 scale where: 3 pts = no feedback, 2 pts = mild feedback, 1 pt = moderate feedback, and 0 pts = strong feedback. Agreement was very high across 3 expert raters who each did two repetitions. There was perfect agreement (same score) on 77% of trials and strong agreement (all scores within 1 point) on 100% of trials. With this in mind, we simply averaged the points across raters for each challenge case. The final metric was computed by summing the number of points across a device's 2 challenge cases and then scaling to a 5 point scale.

5.0 Results

5.1 Group Average Metrics

We ran eighteen popular hearing devices (12 BTE RICs, 3 earbuds, 3 in-canal devices) through this procedure and computed metrics on each of our 5 point scales. In Table 2 we report the group average and standard deviations across all devices for each metric. Values are separated for initial and tuned fits. Scores for individual devices are on hearingracker.com. The average speech benefit

	Speech Benefit in Quiet/Mod	Speech Benefit in Noisy	Own Voice Not Boomy	Does not squeal	Streaming Music Quality
Initial Fits: **Avg** (s.d)	**3.1** (1.2)	**1.6** (0.8)	**2.8** (1.2)	**4.4** (0.7)	**3.8** (0.8)
Tuned Fits: **Avg** (s.d)	**3.9** (0.7)	**2.2** (0.8)	**2.5** (1.2)	**3.9** (0.8)	**3.9** (0.7)

Table 2. Summary Statistics Across 18 Devices. All metrics are on a scale from 0-5 points where higher is better.

scores in quiet/moderate environments were reasonably high (above 3). This was also the metric that had the largest absolute improvement from "initial" to "tuned" fits (0.8 points), likely due to improvements in audibility resulting from better fitting to prescriptive targets. Additionally, this was the only metric where there was a substantial reduction in standard deviation between initial and tuned fits (0.5 points), likely reflecting the differences in philosophies behind first fits that are reduced when devices are "tuned" to the prescription. Speech perception benefits in loud environments were roughly half the size of those in quiet environments. This decrease is likely due to a smaller influence of increasing audibility and a greater influence of increasing signal-to-noise ratio (where hearing aids often underperform) on our metric. The occlusion ("Own Voice Not Boomy") metric had the highest group standard deviation (1.2 points), primarily reflecting the wide spread in ear tip coupling across devices. The average decrease from initial to tuned fits (0.3 points) reflects the use of more occluding ear tips that were needed to match to prescriptive gains. The feedback metric ("Does Not Squeal") indicated good performance across most devices. There was a modest decrease in the metric from initial to tuned fit (0.4 points), due to more gain being selected for the later fit. Finally the streaming music quality metric, performance is very similar between initial and tuned fits (0.1 point improvement) potentially due to counteracting effects of increased occlusion (improving the metric) and increased spectral tilt (decreasing the metric).

6.0 Conclusion

In this whitepaper we describe our method for (a) recording hearing aids in realistic and repeatable environments (b) deriving hearing aid settings using audiological practices (c) processing those recordings for presentation over the internet, and (d) computing perceptually-relevant metrics on those recordings. At present, we have performed these procedures on 18 popular hearing devices and the resulting content is

published on Hearing Tracker. Our hope is that these recordings and metrics can help provide clarity to hearing aid consumers who are attempting to make purchase decisions based on a device's audio performance. Our intent is to update our database as new devices are released and to update our methods/metrics as better ones emerge.

7.0 Acknowledgements

We'd like to thank Jim Kates Ph.D. for use of the HASPIv2 and HAAQI metrics. Bill Rabinowitz Ph.D and Dianne Van Tasell Ph.D. provided helpful conversations. We'd also like to thank the creators of the ARTE database for publishing their data. Finally, we note that Andrew Sabin Ph.D. is a scientific advisor to HearAdvisor LLC and is also employed by a company that participates in the hearing device market. As a result, he does not directly participate in evaluation of any products that are associated with Bose Corporation.

8.0 References

[1] Marshall Chasin, A., & Gastmeier, B. (2021).Critical Distance: How Far Can Musicians and Choir Members Be Spaced from Each Other?. *Hear. Rev.* 28(3):27-29.

[2] Weisser, A., Buchholz, J. M., Oreinos, C., Badajoz-Davila, J., Galloway, J., Beechey, T., & Keidser, G. (2019). The ambisonic recordings of typical environments (ARTE) database. Acta Acustica United With Acustica, 105(4), 695-713.

[3] Wu, Y. H., Stangl, E., Chipara, O., Hasan, S. S., Welhaven, A., & Oleson, J. (2018). Characteristics of real-world signal-to-noise ratios and speech listening situations of older adults with mild-to-moderate hearing loss. Ear and Hearing, 39(2), 293.

[4] Worthington, D. A., Siegel, J. H., Wilber, L. A., Faber, B. M., Dunckley, K. T., Garstecki, D. C., & Dhar, S. (2009). Comparing two methods to measure preferred listening levels of personal listening devices. The Journal of the Acoustical Society of America, 125(6), 3733-3741.

[5] Bentler, R. A., & Pavlovic, C. V. (1992). Addendum to "transfer functions and correction factors used in hearing aid evaluation and research". Ear and Hearing, 13(4), 284-286.

[6] Bisgaard, N., Vlaming, M. S., & Dahlquist, M. (2010). Standard audiograms for the IEC 60118-15 measurement procedure. Trends in amplification, 14(2), 113-120.

[7] Keidser, G., Dillon, H., Flax, M., Ching, T., & Brewer, S. (2011). The NAL-NL2 prescription procedure. Audiology research, 1(1), 88-90.

[8] Anderson, M. C., Arehart, K. H., & Souza, P. E. (2018). Survey of current practice in the fitting and fine-tuning of common signal-processing features in hearing aids for adults. Journal of the American Academy of Audiology, 29(02), 118-124.

[9] ANSI S3.5-1997 ~1997!. American National Standard: Methods for the Calculation of the Speech Intelligibility Index ~American National Standards Institute, New York.

[10] Holube, I., Fredelake, S., Vlaming, M., & Kollmeier, B. (2010). Development and analysis of an international speech test signal (ISTS). International journal of audiology, 49(12), 891-903.

[11] Almufarrij, I., Dillon, H., & Munro, K. J. (2021). Does probe-tube verification of real-Ear hearing Aid amplification characteristics improve outcomes in adults? A systematic review and meta-analysis. Trends in hearing, 25, 2331216521999563.

[12] http://apps.apple.com/us/app/mimi-hearing-test/id932496645

[13] Manchaiah, V., Picou, E. M., Bailey, A., & Rodrigo, H. (2021). Consumer Ratings of the Most Desirable Hearing Aid Attributes. Journal of the American Academy of Audiology, 32(08), 537-546.

[14] Kates, J. M., & Arehart, K. H. (2021). The hearing-aid speech perception index (haspi) version 2. Speech Communication, 131, 35-46.

[15] Studebaker, G. A. (1985). A" rationalized" arcsine transform. Journal of Speech, Language, and Hearing Research, 28(3), 455-462.

[16] Jenstad, L. M., Van Tasell, D. J., & Ewert, C. (2003). Hearing aid troubleshooting based on patients' descriptions. Journal of the American Academy of Audiology, 14(07), 347-360.

[17] Kuk, F., Keenan, D., & Ludvigsen, C. (2005). Efficacy of an open-fitting hearing aid. Hearing Review, 12(2), 26-3

[18] Cubick, J., Caporali, S., Lelic, D., Catic, J., Damsgaard, A. V., Rose, S., ... & Schmidt, E. (2022). The Acoustics of

Instant Ear Tips and Their Implications for Hearing-Aid Fitting. Ear and Hearing, 43(6), 1771-1782.

[19] Sabin, A. (2020). Tech Trends in OTC Hearing Aids. *Hear Rev*, *27*(6), 18-19.

[20] Kuk, F., Keenan, D., & Lau, C. C. (2009). Comparison of vent effects between a solid earmold and a hollow earmold. *Journal of the American Academy of Audiology*, *20*(08), 480-491.

[21] Kates, J. M., & Arehart, K. H. (2015). The hearing-aid audio quality index (HAAQI). IEEE/ACM transactions on audio, speech, and language processing, 24(2), 354-365.

[22] Arehart, K. H., Kates, J. M., & Anderson, M. C. (2011). Effects of noise, nonlinear processing, and linear filtering on perceived music quality. International Journal of Audiology, 50(3), 177-190.