Perceived Sound Quality Dimensions Influencing Frequency-Gain Shaping Preferences for Hearing Aid-Amplified Speech and Music

Jonathan M. Vaisberg[1,2,3] (iD), Steve Beaulac[1], Danielle Glista[1,4], Ewan A. Macpherson[1,4], and Susan D. Scollie[1,4]

Abstract

Hearing aids are typically fitted using speech-based prescriptive formulae to make speech more intelligible. Individual preferences may vary from these prescriptions and may also vary with signal type. It is important to consider what motivates listener preferences and how those preferences can inform hearing aid processing so that assistive listening devices can best be tailored for hearing aid users. Therefore, this study explored preferred frequency-gain shaping relative to prescribed gain for speech and music samples. Preferred gain was determined for 22 listeners with mild sloping to moderately severe hearing loss relative to individually prescribed amplification while listening to samples of male speech, female speech, pop music, and classical music across low-, mid-, and high-frequency bands. Samples were amplified using a fast-acting compression hearing aid simulator. Preferences were determined using an adaptive paired comparison procedure. Listeners then rated speech and music samples processed using prescribed and preferred shaping across different sound quality descriptors. On average, low-frequency gain was significantly increased relative to the prescription for all stimuli and most substantially for pop and classical music. High-frequency gain was decreased significantly for pop music and male speech. Gain adjustments, particularly in the mid- and high-frequency bands, varied considerably between listeners. Music preferences were driven by changes in perceived fullness and sharpness, whereas speech preferences were driven by changes in perceived intelligibility and loudness. The results generally support the use of prescribed amplification to optimize speech intelligibility and alternative amplification for music listening for most listeners.

Professionals often fit hearing aids using standardized, evidence-based prescriptive formulae, such as CAMEQ2-HF (Moore et al., 2010), DSL[i/o] (Cornelisse et al., 1995), DSL v5.0 (Scollie et al., 2005), and NAL-NL2 (Keidser et al., 2011) to provide individualized frequency-gain characteristics with the general goal of improving speech intelligibility. These formulae calculate prescribed gain as a function of frequency, level, and hearing loss for a signal with the long-term average speech spectrum (e.g., Cox et al., 1988; Holube et al., 2010) as the input. The use of prescriptive formulae is considered important in best-practice guidelines (American Speech-Language-Hearing Association Ad

[1]National Centre for Audiology, Western University, London, Ontario, Canada
[2]Graduate Program in Health & Rehabilitation Sciences (Hearing Science), Faculty of Health Sciences, Western University, London, Ontario, Canada
[3]Bose Corporation, Boston, Massachusetts, United States
[4]School of Communication Sciences & Disorders, Western University, London, Ontario, Canada

Corresponding Author:
Jonathan M. Vaisberg, National Centre for Audiology, Western University, 1201 Western Road, Rm 2262, London, ON N6G 1H1, Canada.
Email: jonathan.vaisberg@gmail.com

Hoc Committee on Hearing Aid Selection and Fitting, 1998; British Society of Hearing Aid Audiologists, 2012; Valente et al., 2006). Some of these guidelines also emphasize the need to provide tolerable and comfortable amplification (American Speech-Language-Hearing Association Ad Hoc Committee on Hearing Aid Selection and Fitting, 1998) which is consistent with studies of the trade-offs between intelligibility and sound quality in hearing aid fittings (Humes, 2003; Jenstad et al., 2007). Evaluating sound quality for hearing aids fit using an intelligibility-driven prescription is of particular importance because poor sound quality remains a significant barrier to device adoption (Abrams & Kihm, 2015).

Sound quality evaluations of prescriptive formulae have appeared in the literature. Comparisons of NAL-NL2 and CAM2 (a variation of CAMEQ2-HF) revealed individual differences in sound quality preferences for either formula (Johnson, 2013; Moore & Sęk, 2013, 2016) which were attributed, in part, to factors including greater high-frequency gain in CAM2, stimulus input level, noise type, and hearing profile. Other studies have identified a range of fittings that deviate from prescribed fittings but maintain acceptable intelligibility and quality. Jenstad et al. (2007) measured hearing-impaired listeners' speech quality judgments and consonant identification scores using a range of hearing aid fittings, including targets prescribed using the DSL[i/o] formula (Cornelisse et al., 1995). A range of fittings varying by up to 10 dB in low- and high-frequency bands relative to the DSL[i/o]-prescribed fitting were associated with near-optimal quality judgments and speech identification scores. Polonenko et al. (2010) demonstrated that preferred listening levels were similar to DSLv5-adult targets, which are lower than the DSL[i/o] targets used by Jenstad et al. (2007). More recently, Van Eeckhoutte et al. (2020) studied whether preferred listening levels vary with contemporary full bandwidth fittings versus a narrowband fitting and determined few differences in preference, while speech recognition improved significantly in the full-band condition.

Similarly, Dirks et al. (1993) evaluated hearing-impaired listeners' preferred frequency-gain responses for two- and three-channel amplification systems relative to linear NAL targets. Listeners compared NAL-processed discourse in noise varying in low-, mid-, and high-frequency gain and were instructed to make overall preference judgments based on their own internal weighting of intelligibility and quality attributes. Listener fittings were within 6 dB of NAL targets, except for more relative low-frequency gain. Other studies have found similar results using other fitting formulas (Kuk & Pape, 1992; van Buuren et al., 1995). Together, these studies characterize the impact of prescriptive formulae on the sound quality of speech.

If speech-based amplification is the foundation of hearing aid technology, then sound quality implications of prescribed amplification should be evaluated for music listening, because modern hearing assistance technology is becoming more and more integrated with wireless streaming and other consumer audio technologies.

Hearing aid users report dissatisfaction with hearing aid processed music (Greasley et al., 2020; Leek et al., 2008; Looi et al., 2019; Madsen & Moore, 2014; Vaisberg et al., 2019), and this may relate to the variable nature of the acoustics of music.Speech is acoustically predictable because it originates from the vocal tract and has well-understood spectral content and levels (Hillenbrand et al., 1995; Olsen, 1998). One study demonstrated similar long-term spectra across 12 languages (Byrne et al., 1994). Music, however, originates from a variety of instruments differing in shape, size, and composition, which creates a larger, less predictable range of spectra and level fluctuations (Chasin & Hockley, 2014). For instance, music genres such as rock and rap tend to exhibit smaller dynamic ranges than classical genres such as opera and orchestra (Kirchberger & Russo, 2016), and music tracks containing more percussion instruments tend to exhibit more low- and high-frequency energy than those without (Elowsson & Friberg, 2017).

Dissatisfaction with hearing aid processed music may also relate to listening purpose and intelligibility. When listening to speech, intelligibility is highly relevant. Since music often includes lyrics, it may also be important to optimize lyric intelligibility. However, many listeners do not to attend to music lyrics (Condit-Schultz & Huron, 2015), and the sound quality of music is also driven by instrumental components of song. Therefore, lack of intelligibility in music may not affect listening experience in ways it would for speech communication. This may mean that sound quality, rather than intelligibility, optimization may be the primary goal for hearing aid-amplified music.

Many researchers have investigated the impact of different hearing aid settings for music listening. Preferences for CAM2 over NAL-NL2 observed by Moore and Sęk (2013, 2016) were obtained, in part, using music stimuli. Further, Moore et al. (2016) investigated the impact of modified hearing aid fittings relative to NAL-NL2 for different acoustic scenes. In a music scene, NAL-NL2 processing was compared with processing with an additional average 10 dB of gain at 0.25 and 0.5 kHz. Listeners rated the quality of the test condition as *boomy* relative to NAL-NL2, leading to the recommendation for low-frequency gain modifications for aided music listening lower than those tested. Despite reports of *boominess* by Moore et al. (2016), additional low-frequency energy is often preferred for music (Arehart et al., 2011; Franks, 1982; Punch, 1978; Vaisberg et al., 2020), as is extended high-frequency

audibility, at least for listeners with flat hearing configurations (Moore et al., 2011; Ricketts et al., 2008). Linear processing is also typically preferred compared with compressive nonlinear processing (Davies-Venn et al., 2007; Kirchberger & Russo, 2016).

While this research advances knowledge for optimizing hearing aid music listening, the research tends to be focused on specific features, restricting music preferences to a set of discrete experimental manipulations of a technology under question. An experimental methodology that enables more listener personalization and user-centric findings is more desirable compared with a methodology comparing a few manipulations of one or two parameters. Further, broadly conceived selection criteria such as preference or overall impression, as measured by Dirks et al. (1993), allow listeners to choose individualized optima based on internal weightings of more specific objective and subjective attributes such as intelligibility, quality, or loudness.

In summary, hearing aid prescriptions are commonly used to provide frequency-gain shaping that supports speech intelligibility at a reasonable listening level. However, the suitability of prescriptive gain for non-speech stimuli is less understood. This study sought to identify preferred amplification settings for speech and music with adjustment in three bands, using the DSL v5.0-adult prescription as a reference. This study used the modified simplex procedure (Amlani & Schafer, 2009; Kuk & Lau, 1995, 1996; Kuk & Pape, 1992, 1993; Neuman et al., 1987; Preminger et al., 2000; Stelmachowicz et al., 1994) to determine preferred frequency-gain shaping. The simplex procedure was implemented using an overall preference criterion, where listeners judged using internal weightings of intelligibility and quality attributes, similar to the preference criterion implemented by Dirks et al. (1993). Follow-up sound quality ratings with specific sound quality descriptors were also used to understand perceptual differences between preferred and prescribed settings, and how those perceptual differences may explain listener preference judgments. Sound quality ratings were also gathered for novel stimuli shaped using the preferred shaping from similar-genre counterparts from the simplex procedure to assess if perceptions of sound quality generalize between similar stimuli belonging to the same genre.

The objectives of this study were therefore (a) to quantify listener's preferred frequency-gain shaping compared with prescribed frequency-gain shaping for speech and music using overall preference as a criterion, (b) to determine whether unique preferred frequency-gain shaping exists for speech versus music, and (c) to determine which, if any, sound quality descriptors explained listener preferences. This study used equipment that provided frequency-gain shaping in combination with multichannel dynamic range compression so that any interactions of these would be represented. Further, this study used a closed coupling to the ear so that the impact of amplification in the extended high- and low-frequency ranges could be fully investigated.

Methods

Listeners

Twenty-two adult listeners between the ages of 51 and 81 years (mean = 68.3, standard deviation [SD] = 7.3) participated in the study. On average, listeners had symmetrical mild sloping to moderate hearing loss (Figure 1). Pure-tone thresholds were measured using ER-3A insert earphones in a sound-attenuated booth at octave and interoctave frequencies from .25 to 8 kHz. Eleven listeners were hearing aid users (1–22 years of experience, mean = 8.9, SD = 6.1), and 13 listeners reported having experience playing musical instruments (1–62 years of experience, mean = 9.53, SD = 19.4). A hearing aid user was defined as one who
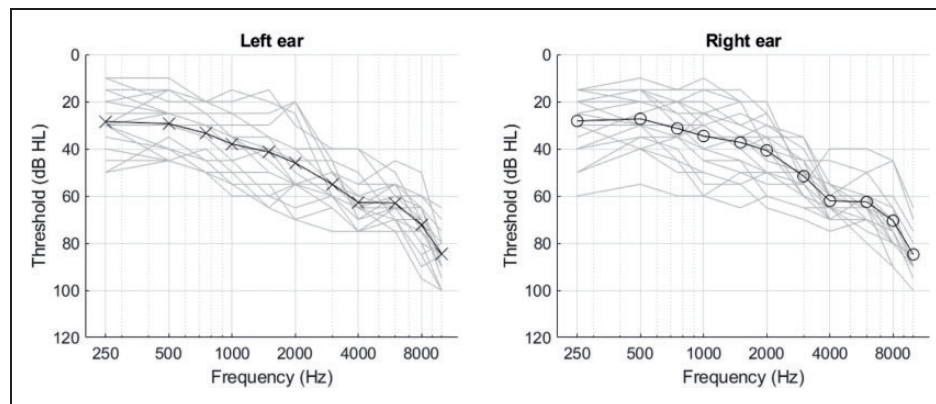

Figure 1. Mean air conduction pure-tone thresholds for listeners' left ears (left panel) and right ears (right panel). The dark lines show the group means.

owned and wore a hearing aid to mitigate impacts of hearing loss. Music experience was defined as one who engaged in music lessons, formal performance, or casual playing alone or with others. This study was approved by the Western University Health Science Research Ethics Board, and listeners were paid for their participation.

Sample size estimation for an 8-measurement within-subjects repeated-measures analysis of variance (ANOVA), assuming .95 power, .05 significance, and .7 correlation among repeated measures, determined that 15 listeners would be sufficient. Sample size estimation was conducted using G*Power Version 3.1.9.2 (Faul et al., 2007).

Procedure

Calibration and Fitting to Initial Targets. This study used an open source master hearing aid (openMHA; Herzke et al., 2017) to process and amplify test materials. The openMHA was installed on a Linux computer (Ubuntu 18.04) and connected to a low-latency Focusrite Scarlett 18i8 USB soundcard (High Wycombe, UK), which sent the signals to two Etymotic Research 4p (ER4p, Elk Grove Village, IL, USA) insert earphones. The ER4p insert earphones were coupled to listeners' ears using an occluding foam tip. A fully occluding transducer was desired so that low-frequency gain adjustments could be fully explored without needing to account for leakage from listeners' ear canals. The openMHA was calibrated by presenting the International Speech Test Signal (ISTS; Holube et al., 2010) from the ER4p into a Bruel & Kjær (B&K, Næerum, Denmark) Type 4157 occluded ear simulator mounted on a B&K Type 2250 sound level meter. The ISTS was digitally scaled to produce a long-term average of 70 dB SPL when zero gain was applied. This scaling allowed sufficient headroom for listeners to increase the overall level and modify frequency-gain shaping during the experiment before encountering digital peak-clipping or earphone distortion. Daily calibration checks were performed in a hearing aid analyzer (Audioscan Verifit2, Dorchester, ON, Canada) 0.4 cc coupler.

The openMHA implemented 21-channel multiband dynamic compression. The openMHA inputs were the digital stimuli. Next, the openMHA applied a reference input peak level of 125 dB SPL (0 dB full scale corresponds to this SPL) to determine the simulated SPL level of the waveform. The test materials were scaled to 55 dB below full scale so that the openMHA would apply level-dependent gains for a simulated average 70 dB SPL input level. Next, the waveform was processed using a fast Fourier transform filterbank, in which the signal was processed using 21 Hann window filters centered on 1/3rd octave bands from 0.1 to 10 kHz

with 50% overlap of adjacent filters. Frequency-gain shaping and dynamic range compression were prescribed for each listener and ear using DSL v5.0 gains for 55, 65, and 75 dB SPL input levels. Gains were manually inputted into the openMHA software. Fast-acting compression attack (0.02 seconds) and release times (0.1 seconds) were applied. The amplified waveform was then produced by summing the filter outputs and digital to analog conversion. Analog outputs were sent to the ER4p transducers.

Individual openMHA fittings were verified in the Verifit2. First, thresholds and wideband real-ear-to-coupler difference measurements were entered into the Verifit2, which generated targets for a speech signal with an overall input level of 70 dB SPL. The wideband real-ear-to-coupler difference was measured to capture individual ear canal resonances. Second, the openMHA output was routed to the ER4p transducers, which coupled to the Verifit2 0.4 cc couplers. The openMHA was fine-tuned such that output SPL was within an average 3.5 dB root-mean-square of audiometric targets at octave and interactive frequencies from 0.25 to 8 kHz across listeners. Specifically, average systematic deviations from targets at 0.5, 1, 2, and 4 kHz were –3.3, –3.0, –1.5, and –1.6 dB, respectively, and average absolute deviations were 3.4, 3.3, 2.4, and 2.4 dB, respectively.

Modified Simplex Procedure. This study implemented the simplex method in a three-dimensional space, permitting listeners to adjust hearing aid amplification relative to prescribed settings using preference-based gain adjustments in three frequency bands. The initial shaping used DSL v5.0 gains with a simulated hearing aid input level of 70 dB SPL. This level is an average comfortable listening input level for aided music listening (Croghan et al., 2016). The 21 channels were grouped into low-frequency (0.1–0.8 kHz), mid-frequency (1–2.5 kHz), and high-frequency (3–10 kHz) bands, in which listeners compared between gain differences of ±6 dB in each band. This step size, previously implemented in three-dimensional of the simplex procedure (Dirks et al., 1993), is considered large enough to be perceptible while small enough to provide sensitive preference evaluation. This step size is also similar to just noticeable differences for frequency-gain adjustments found in recent studies (Caswell-Midwinter & Whitmer, 2019a, 2019b).

The simplex procedure determines preferred frequency-gain shaping using a series of iterations and follows a method for a two-dimensional simplex originally described by Neuman et al. (1987). For example, a center coordinate (0,0,0) represents the prescribed initial estimate, and each step along the x, y, and z dimensions represents a ±6 dB adjustment in either the low-, mid-,

or high-frequency gain. Each iteration consists of three paired comparisons. In the first iteration, the listener chooses either the prescribed frequency-gain shape (0,0,0) or an alternative with a +6 dB low-frequency (1,0,0), –6 dB mid-frequency (0,–1,0), or –6 dB high-frequency (0,0,–1) gain adjustment. The listener can repeat each comparison once. The winner from each of the three comparisons determines the estimated preferred shape for the subsequent iteration. If the listener selects the initial estimate (0,0,0), then the listener would compare their selection (0,0,0) with even more gain in the subsequent iteration (0,0,1). If the listener selects the adjustment with less gain (0,0,–1), then the listener would compare their selection (0,0,–1) with even less gain in the subsequent iteration (0,0, –2). This procedure gets repeated for each dimension and predicts an estimated preferred shape for the subsequent iteration. For example, if during the first iteration the listener selects more gain at low and mid frequencies and less gain at high frequencies, the estimated preferred shape for the second iteration would be (1,0,–1), and this would be compared with (2,0,–1), (1,1,–1), and (1,0,–2). If the listener selects the estimated preferred shape after all three comparisons in the second iteration, a reversal would have occurred, and the paired comparisons would be reflected. Therefore, the third iteration of the aforementioned example consists of (1,0,–1), as the estimated preferred shape and gets compared with (0,0,–1), (1,–1,–1), and (1,0,0). If the listener selects the estimated preferred shape after all three comparisons in the third iteration, then the listener has indicated preference for the estimated preferred shape estimate to every possible adjustment within the simplex framework. Therefore, the run is terminated, and the estimated preferred shape defined by the center coordinate from the final iteration is taken as the listener's preferred frequency-gain shaping. If after 18 iterations (54 comparisons) the listener does not complete two reversals, then the run is terminated, and the final shape is taken as the listener's preferred shape. This stopping rule was adopted because in past three-dimensional adaptive procedures similar to simplex, a set of 54 paired comparisons was considered a reasonable amount of testing to minimize fatigue and reach an optimum (Franck et al., 2004, 2007).

Test Materials. The test materials were two sentence pairs (Institute of Electrical and Electronics Engineers [IEEE], 1969) and two music files. The sentence pairs were chosen to represent both genders of voice and consisted of a male-spoken, *Raise the sail and steer the ship northward. A cone costs five cents on Mondays* and a female-spoken *Would you please give us the facts? He arrived home every other night.* The music samples were chosen to represent genres that might interact differently with signal processing adjustments (Arehart et al., 2011; Davies-Venn et al.,

2007). These consisted of a 5.1-second sample from the contemporary/pop song *A Little Help from my Friends* by The Beatles and a 2.6-second sample from Mozart's classical string arrangement *Serenade No. 6, K. 239 Serenata notturna: III. Rondo. Allegro* by the Franz Liszt Chamber Orchestra & Sandor Frigyes. The music sample durations were long enough to include a full musical phrase (i.e., a passage having a complete musical sense of its own) but short enough so that minimal fatigue occurred during a test session. Pilot testing suggested that a 2-second stimulus duration was sufficient for listeners to confidently judge stimulus preference. The stimuli sampling rate was 44.1 kHz with 16 bits per sample. If a stimulus was in stereo format, it was summed to mono format so that the two openMHA inputs received the same audio signal prior to amplification.

Simplex Implementation. The simplex procedure was written and administered using MATLAB (version 2017b). The Windows computer was connected to a touch screen monitor inside the sound booth and to the Linux MHA via an ethernet connection. This allowed listener judgments to trigger stimulus presentations.

Before experimental testing, listeners completed a practice run using nonadaptive paired comparisons, also programmed with MATLAB and implemented via the openMHA. Their instructions were to listen to the two stimuli and choose the one they preferred. Listeners were prompted to choose the version of the stimulus that they would prefer to listen to throughout the day. The practice run conditions were defined using the simplex parameters and consisted of six predefined stimulus pairs comparing prescribed frequency-gain shaping (0,0,0) with highly modified versions: a simulated low-pass filter (1,–4,–4) in which the low-frequency band gain was increased by 6 dB and the mid- and high-frequency band gains were decreased by 24 dB; or a simulated high-pass filter (–4,–4,1) in which the high-frequency band gain was increased by 6 dB and the mid- and low-frequency band gains were decreased by 24 dB. All listeners preferred the prescribed shaping over the filtered stimuli, as expected.

During experimental testing, each listener completed two simplex runs for each of the four stimuli, totaling eight simplex runs. The direction of adjustment (increasing or decreasing gains) during the initial simplex iteration and the ordering of parameter comparisons within each simplex iteration were randomized. The presentation order of the stimuli was also randomized, except that simplex runs for the same stimulus did not occur in adjacent trials. Each listener's preferred frequency-gain shaping was determined by calculating the average of the final x-, y-, and z- coordinates across the two simplex runs for the same stimulus.

Sound Quality Ratings. Listeners rated several dimensions of sound quality for each stimulus, using two versions of each stimulus: one processed using their prescribed frequency-gain shaping and one processed using their preferred frequency-gain shaping as determined using the modified simplex procedure. The sound quality dimensions were adapted from Gabrielsson et al. (1988) and Davies-Venn et al. (2007) and consisted of *Overall Impression, Loudness, Fullness,* and *Sharpness.* Ratings of *Intelligibility* were also obtained, but for speech only. Listeners gave ratings along each dimension using a continuous horizontal scroll bar with five descriptors from lowest to highest, which produced a number from 0 (lowest) to 10 (highest). The listeners were blind to the numerical rating. The descriptors, adapted from Gabrielsson et al. (1988), from lowest to highest, for *Overall Impression* were *Very Bad, Rather Bad, Midway, Rather Good,* and *Very Good.* The descriptors for *Loudness* were *Very Soft, Rather Soft, Midway, Rather Loud,* and *Very Loud.* The descriptors for *Fullness* were *Very Thin, Rather Thin, Midway, Rather Full,* and *Very Full.* The descriptors for *Sharpness* were *Very Gentle, Rather Gentle, Midway, Rather Shrill,* and *Very Shrill.* The descriptors for *Intelligibility* were *Very Unclear, Rather Unclear, Midway, Rather Clear,* and *Very Clear.*

Test Materials. The test materials for the sound quality ratings consisted of the speech and music passages from the simplex procedure, referred to here as *experimental* stimuli, as well as new speech and music passages belonging to the same categories, referred to here as *generalization* stimuli.

The generalization sentence pairs were the male-spoken, *The ramp led up to the wide highway. Beat the dust from the rug onto the lawn,* and female-spoken, *They could laugh, although they were sad. Farmers came in to thresh the oat crop,* IEEE sentences. The talker for each gender differed between the experimental and generalization stimuli. The generalization music passages were downloaded from iTunes and included a 6.4-second clip of *New Orleans is Sinking* by The Tragically Hip for the pop genre and a 7.3-second clip of Beethoven's *String Quartet No. 4 in C Minor, Op. 18: III. Menuetto: Allegretto* by the Emperor String Quartet for the classical string genre.

Music stimulus durations for the sound quality ratings were approximately twice as long and all stimuli were looped compared with stimulus durations for the simplex procedure, which were shorter and presented only up to two times. This was done for two reasons. First, whereas a fixed number of sound quality ratings were obtained from each listener during the sound quality rating procedure, the number of paired comparisons presented to listeners during the simplex procedure was highly variable. It was decided to reduce stimulus duration during the simplex procedure to complete experimental testing within the allotted time frame. Second, while listeners indicated preferences during the simplex procedure using a preference criterion, listeners were required to shift their attention to different sound quality attributes for the sound quality ratings. Therefore, given the fixed number of stimulus presentations in the sound quality rating task, listeners were permitted to take as much time as needed to complete ratings. Unlike the simplex procedure, listeners were not able to repeat a stimulus.

Sound Quality Rating Implementation. The sound quality rating procedure was written and administered in MATLAB and used the same hardware setup as the simplex procedure. Each stimulus was processed using each listener's prescribed and preferred shaping. This yielded a total of 16 stimuli to be rated (4 categories × 2 experimental/generalization × 2 prescribed/preferred shaping). A total of 4 descriptors for each music stimulus and 5 descriptors for each speech stimulus meant that 144 ratings were completed. Speech and music stimuli were presented in separate blocks. Within each block, listeners rated all sound quality dimensions for a single condition (stimulus × [prescribed or preferred]) before conducting ratings for another condition. Block order, condition order, and sound quality dimension order within each condition were randomized between listeners.

Analysis

The data were analyzed as follows. First, simplex task performance and reliability were interpreted using descriptive statistics and cumulative distributions to ensure that listeners completed the simplex task correctly and consistently. Next, the simplex results were analyzed using (a) a series of *t* tests to determine if preferred frequency-gain shaping was significantly different from prescribed shaping within each stimulus/frequency band combination and (b) a repeated-measures ANOVA to determine if preferred frequency-gain shaping varied between stimulus and frequency band factors. Finally, sound quality results were analyzed using (a) a series of *t* tests to determine if sound quality ratings for preferred frequency-gain shapes were significantly different from sound quality ratings for prescribed frequency-gain shapes, (b) a linear mixed-effects model to determine which sound quality descriptors were predictive of overall impression ratings, and (c) rank correlation coefficients between ratings for experimental and generalization stimulus pairs.

Reliability

For the simplex procedure, the number of iterations per test–retest and per stimulus, number of complete tests per stimulus, number of time-outs across listeners, and reliability between test and retest were calculated. Reliability was assessed by measuring the distance in steps between listeners' preferred shaping coordinates in test and retest for the same stimulus either within each dimension or across all dimensions, as was done by Kuk and Pape (1992). Within each dimension, the final coordinate from the first run was subtracted from the final coordinate from the second run. Across all dimensions, the three-dimensional distance was calculated between the final coordinate from the first run and the final coordinate from the second run by measuring the square root of the sum of squares of the differences across all three dimensions. This observed reliability was then compared with a simulated simplex of random preferences to test whether listener responses were random or systematic. Cumulative distribution curves were computed for 1,000 pairs of randomly selected preferred shaping coordinates as was done by Franck et al. (2004, 2007).

Preferred Shaping

Preferred gain adjustments were measured by subtracting each listener's prescribed gain response from their average preferred gain response for each stimulus. The dB differences were verified by measuring the simulated real ear output for each listener's prescribed and preferred frequency-gain shaping from the ER4p insert earphones coupled to the Verifit2 0.4 cc coupler. The shaping for each stimulus was measured using the ISTS; this allowed the shaping differences to be compared across stimulus types.

Differences between preferred and prescribed shaping were quantified as the difference between the preferred and prescribed spectra at octave and interoctave frequencies (0.25, 0.5, 0.75, 1, 2, 3, 4, 6, 8 kHz). Differences were averaged within the low- (0.1–0.8 kHz), mid- (1–2.5 kHz), and high- (3–10 kHz) frequency bands. To determine if preferred shaping was significantly different from prescribed shaping, 12 t tests were conducted—one for each stimulus. The critical alpha level was corrected with Bonferroni's procedure to .004. To determine if preferred shaping differed between stimuli, a 4×3 repeated-measures ANOVA was used to test the effect of stimulus × frequency band on observed gain differences. The differences were assessed for normality by visual inspection of histograms, and Greenhouse–Geisser corrections to degrees of freedom were applied to adjust for departures from sphericity. Post hoc contrasts were performed using the Holm correction. Statistical analyses were completed using RStudio (Version 1.0.132; R Core Team, 2017) and the ez package (Lawrence, 2016).

Sound Quality Ratings of Prescribed Versus Preferred Shapes

Sound quality ratings were analyzed with three objectives: (a) to determine if preferred and prescribed shapes led to different perceived sound quality, (b) to determine which sound quality descriptor rating differences were most predictive of Overall Impression rating differences, and (c) to determine whether stimulus-specific sound quality ratings generalized to other stimuli belonging to a similar genre for music or to the same gender for speech.

Sound quality ratings were analyzed using paired t tests between stimuli shaped using preferred and prescribed settings. t tests were computed for each sound quality descriptor within each stimulus, which consisted of Overall Impression, Loudness, Fullness, and Sharpness for all stimuli and Intelligibility for speech alone (totaling 18 t tests). The alpha level was corrected with Bonferroni's procedure to .0028. Analyses were not computed across stimuli due to stimulus-specific spectral shaping.

To determine which sound quality rating scales were most predictive of Overall Impression ratings, differences between sound quality ratings were calculated for each individual by subtracting the sound quality rating for the prescribed stimuli from that for the preferred stimuli. Multiple linear mixed models were used to test if rating differences for each sound quality scale were predictive of Overall Impression rating differences. One model was used to analyze differences pooled across all speech stimuli (male and female, preferred and prescribed shaping) with Fullness, Loudness, Sharpness, and Intelligibility differences as fixed effects variables and Overall Impression differences as the outcome variable. Another model was used to analyze differences pooled across all music stimuli with Fullness, Loudness, and Sharpness as predictor variables and Overall Impression differences as the outcome variable. Models were fitted across pooled stimuli because they were used to understand the relationship between sound quality descriptors rather than the difference in ratings between preferred and prescribed shapes. Separate models were fit for speech and music due to the likelihood of listeners having different listening goals for each. The assumption of normally distributed residuals was assessed by visual inspection of histograms (Field et al., 2012, p. 870). The assumption of no multicollinearity was assessed by computing a correlation matrix between all predictor variables and verifying that no Pearson product-moment correlation coefficient between any two predictor variables was greater than $r = .8$. Correlation coefficients

below 0.8 are generally permissible (Field et al., 2012, p. 276). Statistical analyses were completed using RStudio (Version 1.0.132; R Core Team, 2017) and the lme4 package (Bates et al., 2015).

The generalizability of sound quality ratings was inferred by measuring the Spearman rank correlation coefficient between the sound quality ratings for the experimental stimuli and the sound quality ratings for the paired generalization stimuli. Separate correlation coefficients were calculated for each sound quality rating scale pooling all speech stimulus pairs or pooling all music stimulus pairs. Each correlation coefficient was measured across listeners.

Results

Simplex Task Performance and Reliability

Across stimuli, listeners completed an average of 9.4 iterations ($SD = 1.1$) per simplex run, corresponding to an average of 29.4 stimulus pairs, requiring an average of 4 minutes and 22 seconds ($SD = 30$ seconds) per simplex run. The minimum number of iterations per simplex run was 2 (6 stimulus pairs), and the maximum number of iterations was 18 (54 stimulus pairs). A run with two iterations implied that the listener's preferred shaping was no different from their prescribed shaping, and one with 18 iterations implied that the run timed out and the final shaping was considered their preferred shaping. All listeners finished a simplex run before a time-out at least once per stimulus.[1] On average, 1.05 time-outs ($SD = 1.13$) occurred per listener. It was noted that 40.9% of listeners did not time-out, 27.3% of listeners timed out once, 22.7% of listeners timed out twice, 1 listener (4.5%) timed out three times, and 1 listener (4.5%) timed out four times. The minimum amount of time spent was 34 seconds and the maximum amount of time spent was 14 minutes per simplex run. Note that differences in time spent per simplex run may result from differences between stimulus durations and/or number of stimulus pair repetitions. The number of iterations and percentage of time-outs for each stimulus/test–retest combination are listed in Table 1.

Figure 2 shows cumulative distribution curves from the simplex procedure. The figure illustrates how many gain-adjustment steps listeners deviated by between the two simplex runs for a single stimulus along low-, mid-, and high-frequency gain dimensions (top left, top right, and bottom left panels, respectively), as well as the three-dimensional root-mean-square gain-adjustment steps (bottom right panel). The figure also illustrates what percentage of individuals deviated by a given number or fewer step sizes. The cumulative distribution curves of listeners' test–retest differences, within and across dimensions, fell above and to the left of the random

Table 1. Average Number of Iterations (SD = Standard Deviation) Required to Complete a Simplex Run.

Stimulus	Test/retest	No. of iterations (SD)	Time-outs (%)
Male speech	Test	9.3 (1.2)	11.4%
Male speech	Retest	9.6 (1.1)	9.1%
Female speech	Test	10.5 (1.1)	9.1%
Female speech	Retest	8.1 (0.9)	2.3%
Pop music	Test	9.4 (1.0)	4.5%
Pop music	Retest	8.2 (1.1)	2.3%
Classical music	Test	8.8 (0.9)	2.3%
Classical music	Retest	11.0 (1.2)	11.4%

Note. Time-outs refer to the percentage of simplex time-outs for each stimulus × time/retest. A time-out was an instance where the simplex procedure stopped due to listeners not selecting a preferred setting after 18 iterations.

distribution curves. This suggests that listeners indicated preferences more reliably and consistently than randomly selecting a winner for each paired comparison. Seven percent of listeners selected the same preferred shaping coordinates across all stimuli and all dimensions (bottom right). Between test and retest, 36% of listeners selected the same low-frequency adjustment (top left), 41% of listeners selected the same mid-frequency adjustment (top right), and 27% of listeners selected the same high-frequency adjustment (bottom left). For the low-frequency dimension, 93% of listeners were within two step sizes between simplex runs. For the mid-frequency dimension, 83% of listeners were within two step sizes. For the high-frequency dimension, 71% of listeners were within two step sizes. Across all dimensions, 33% of listeners were within two three-dimensional step sizes, and 91% of listeners were within seven step sizes.

Preferred Versus Prescribed Shaping

The average differences from prescribed shaping are illustrated using box-and-whisker plots in Figure 3. The gain for the low-frequency band was increased significantly from prescribed to preferred shaping for all four stimuli; pop music, $t(43) = 8.0$, $p < .0001$, classical music, $t(43) = 5.3$, $p < .0001$, female speech, $t(43) = 4.0$, $p = .0002$, and male speech, $t(43) = 3.7$, $p = .0007$, and gain for the high-frequency band was decreased significantly from prescribed to preferred shaping for pop music, $t(43) = -3.1$, $p = 0.003$, and male speech, $t(43) = -3.2$, $p = .003$.

A repeated-measures ANOVA on preferred versus prescribed gain differences revealed significant main effects of stimulus, $F(2.18, 45.87) = 5.57$, $p < .01$, $\eta^2 = 0.03$, and frequency band, $F(1.42, 29.83) = 27.94$, $p < .0001$, $\eta^2 = 0.21$. There was also a significant interaction of stimulus and frequency band, $F(3.65, 76.64) =$
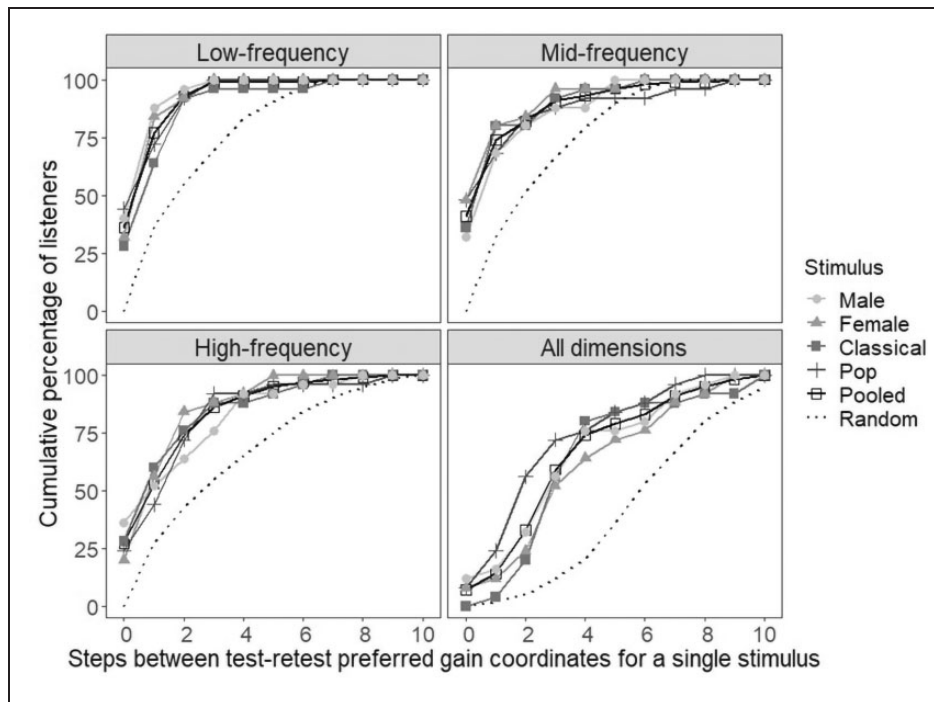
Figure 2. Cumulative distributions showing the percentage of listeners who deviated up to a given number of steps between test–retest preferred gain for each stimulus for the low-frequency dimension (top left panel), mid-frequency dimension (top right panel), high-frequency dimension (bottom left panel), and root-mean-square distance across all three dimensions (bottom right panel). The dotted curves show the cumulative distributions for randomly selected simplex paired comparisons preferred shaping coordinates over 1000 test–retest simulations.
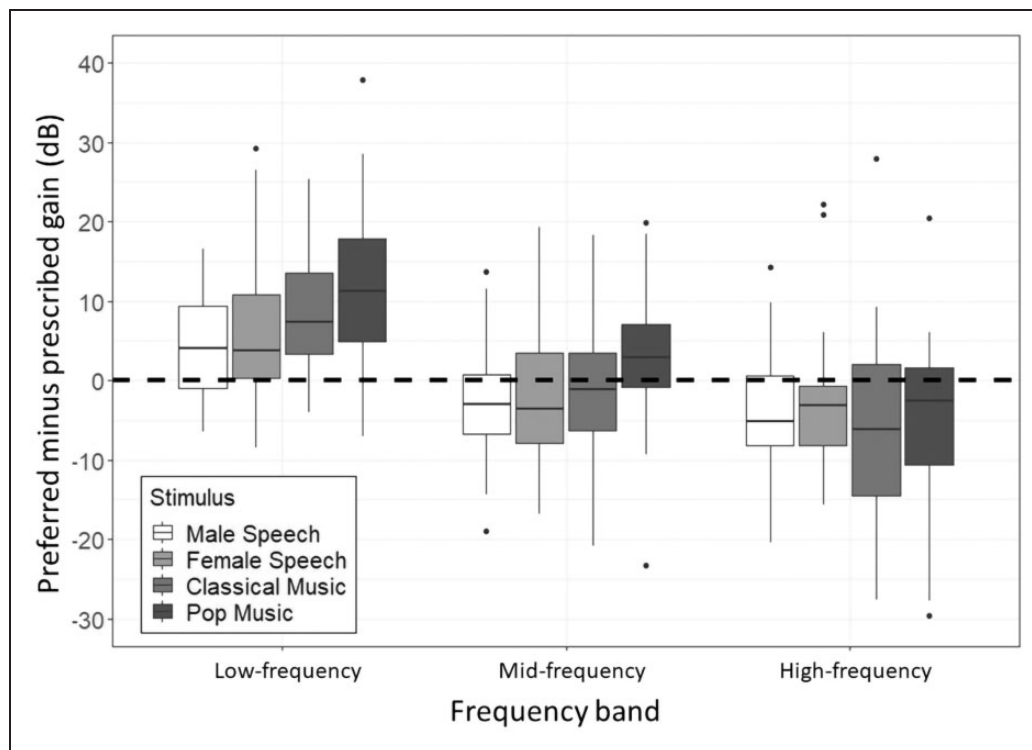
Figure 3. Box-and-whisker plots of differences from prescribed gains in the low- (0.1–0.8 kHz), mid- (1–2.5 kHz) and high- (3–10 kHz) frequency bands for a 70-dB SPL input level. The boxes represent the interquartile ranges of differences, with the lines through the boxes representing the median differences. The lines outside the boxes represent the 91st (top) and 9th (bottom) percentiles of differences, with the dots representing outlier differences. The dashed line represents no difference.

4.32, $p<.01$, $\eta^2=0.03$. With the exception of frequency band, the effect sizes were small. The main effect of stimulus was driven by pop music and male speech. Across listeners and frequency bands, the gain for pop music was adjusted to be on average 4.5 dB higher than for the male speech stimulus. The remaining differences were all 2.7 dB or less and did not reach statistical significance. The main effect of frequency band was driven by differences between the low-frequency band and the other bands. The low-frequency gain was adjusted to be 8.3 dB higher than mid-frequency gain and 11.6 dB higher than high-frequency gain. Post hoc comparisons within each frequency band revealed that gains for pop music were adjusted to be higher than for other stimuli. The low-frequency gain for pop music was increased by 5.3 dB more than for classical music, 6.4 dB more than for female speech, and 8.5 dB more than for male speech. The mid-frequency gain for pop music was increased by 6.2 dB more than for male speech. The remaining contrasts were nonsignificant.

Sound Quality Ratings

Figure 4 shows box-and-whisker plots of the difference in sound quality ratings between prescribed and preferred shaping. Ratings of Overall Impression, $t(84)=$

-3.4, $p<.001$, Loudness, $t(80)=-5.6$, $p<.0001$, and Fullness, $t(72)=-5.9$, $p<.0001$, were significantly higher for the preferred settings than for prescribed settings only for pop music.

The speech sound quality model (Table 2) revealed that changes in Loudness and Intelligibility ratings were significantly predictive of changes in Overall Impression ratings. Based on the magnitudes of the model coefficients, changes in Loudness ratings were most strongly predictive ($\beta=-0.52$, $p<.0001$) and were negatively associated with changes in Overall Impression ratings. Changes in Intelligibility ratings were second-most strongly predictive ($\beta=0.23$, $p<05$) and positively associated with changes in speech Overall Impression ratings. The music sound quality model (Table 3) revealed that changes in Fullness and Sharpness ratings were significantly predictive of changes in Overall Impression ratings. Based on the magnitudes of the model coefficients, changes in Sharpness ratings were most strongly predictive ($\beta=-0.49$, $p<.0001$) and negatively associated with change in music Overall Impression ratings. Changes in Fullness ratings were second-most strongly predictive ($\beta=0.16$, $p<.05$) and positively associated with changes in music Overall Impression ratings.
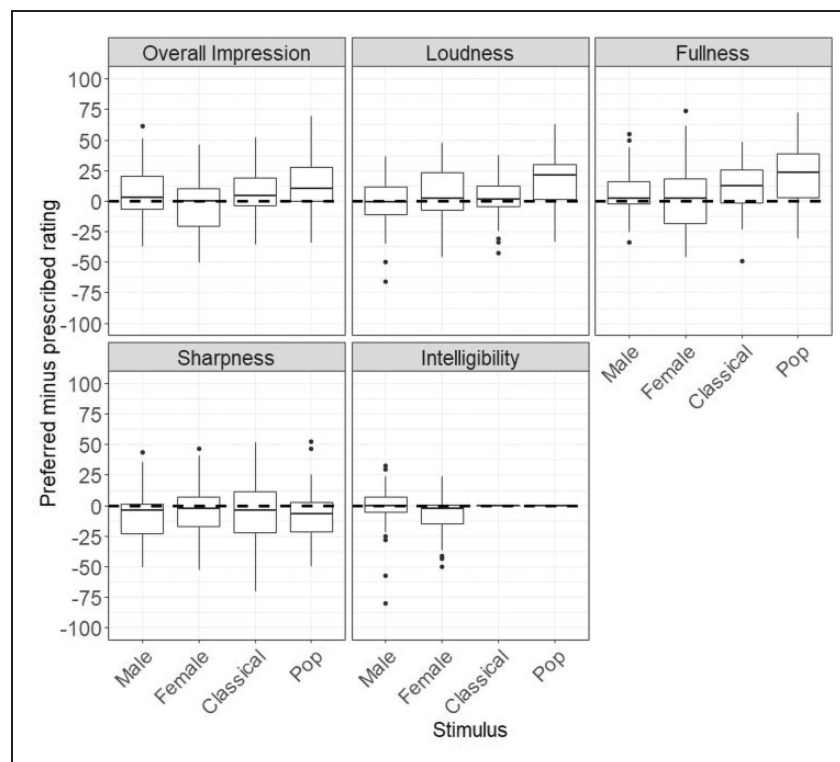

Figure 4. Sound quality differences between ratings for stimuli processed using preferred shaping and prescribed shaping. Intelligibility ratings for classical and pop music were not gathered. The boxes represent the interquartile ranges of differences, with the lines through the boxes representing the median differences. The lines outside the boxes represent the 91st (top) and 9th (bottom) percentiles of differences, with the dots representing outlier differences. The dashed line represents no difference.

Table 2. Linear Mixed Model Results for Speech Sound Quality Ratings.

Fixed effects variable	β estimate	SE	t value	df	p value
Intercept	3.28	2.2	1.5	27.7	.15
Loudness	−0.52	0.12	−4.3	79.0	<.0001
Fullness	0.20	0.10	1.9	84.4	.07
Sharpness	−0.016	0.12	−0.133	83.1	.89
Intelligibility	0.23	0.12	2.0	87.6	<.05

Table 3. Linear Mixed Model Results for Music Sound Quality Ratings.

Fixed effects variable	β estimate	SE	t value	df	p value
Intercept	3.29	2.7	1.2	35.9	.22
Loudness	0.09	0.09	1.0	87.8	.37
Fullness	0.16	0.08	2.2	83.2	<.05
Sharpness	−0.49	0.09	−5.8	65.3	<.0001

Sound quality ratings were moderately correlated across listeners between experimental and generalization stimuli. All correlations were statistically significant ($p < .0001$). For speech stimuli, the correlation coefficients were $\rho_{spearman} = 0.48$ for Overall Impression, $\rho_{spearman} = 0.62$ for Loudness, $\rho_{spearman} = 0.56$ for Fullness, $\rho_{spearman} = 0.49$ for Sharpness, and $\rho_{spearman} = 0.61$ for Intelligibility. For music stimuli, the correlation coefficients were $\rho_{spearman} = 0.49$ for Overall Impression, $\rho_{spearman} = 0.61$ for Loudness, $\rho_{spearman} = .62$ for Fullness, and $\rho_{spearman} = .64$ for Sharpness.

Discussion

This study investigated whether listener preferences differed from prescribed amplification for music and speech stimuli, the impact of preferred settings on sound quality, and which sound quality descriptors drove listener preferences. On average, listeners increased low-frequency gain by about 10 and 5 dB for music and speech, respectively, and decreased high-frequency gain by about 4 dB regardless of stimulus type, relative to the DSLv5-adult prescription. Mid-frequency gain adjustments were minimal. Only the preferred settings for pop music produced significantly greater overall impression, loudness and fullness ratings than the prescribed settings. However, across music stimuli, increases in fullness ratings and decreases in sharpness ratings were significantly associated with increases in overall impression ratings. For speech stimuli, decreases in loudness ratings and increases in intelligibility ratings were significantly associated with increases in overall impression ratings.

Observed Differences From Prescribed Shaping

On average, listeners preferred increased low-frequency gain and decreased high-frequency gain for music and, to a lesser extent, for speech relative to prescribed amplification. These findings support those of Madsen and Moore (2014), in which hearing aid users reported a lack of bass or a shrill/harsh sound quality. The current findings are also consistent with past evaluations of listener frequency-gain preferences. Listeners have preferred amplified music with more low-frequency energy than music with less low-frequency energy (Franks, 1982; Punch, 1978; Vaisberg et al., 2020), and this complements recommendations for an extended low-frequency responses for hearing aid music programs (Moore, 2016). Speech-based studies have also found that listeners prefer more low-frequency gain and less high-frequency gain relative to prescribed NAL-based fittings (Caswell-Midwinter & Whitmer, 2020; Kuk & Pape, 1992, 1993; Nelson et al., 2018; Preminger et al., 2000).

Drivers of Preference Judgments

The observed differences from prescribed shaping may have been unique for each stimulus. This may be due to the differences in the stimuli themselves and/or to the different roles of listening criteria such as quality or intelligibility for speech and music. This interpretation is supported by the results of the sound quality rating procedure.

For music, increases in Overall Impression ratings between prescribed and preferred stimuli were associated with increases of Fullness ratings and decreases of Sharpness ratings. Increased ratings of Fullness correspond to more energy in the low-frequency region (Gabrielsson & Sjögren, 1979), and Fullness ratings have been most strongly associated with ratings of Overall Impression in previous hearing aid music investigations (Davies-Venn et al., 2007; Gabrielsson et al., 1988), corroborating this study's findings. Similarly, decreased ratings of Sharpness correspond to less energy in the high-frequency region. This may explain why sharpness rating differences predicted preferences for decreased high-frequency gain.

For speech stimuli, changes in Overall Impression ratings between prescribed and preferred stimuli were most strongly associated with increases of Intelligibility ratings and decreases of Loudness ratings. Intelligibility ratings have previously been shown to be significantly associated with Overall Impression ratings for speech (Davies-Venn et al., 2007). The relationship between Intelligibility and Overall Impression ratings supports the interpretation that speech understanding is a primary objective of amplification when speech is present. It may

also explain why descriptors such as Fullness and Sharpness were less predictive for speech. Preminger and Van Tasell (1995) studied the quality-intelligibility relationship and found that listeners would attend to sound quality attributes so long as intelligibility was optimal. However, if intelligibility declined, then sound quality ratings would be predicted by intelligibility ratings and would decline in a similar way. In the current study, stimulus pairs during the first simplex iteration consisted of samples processed using prescribed amplification. Therefore, speech intelligibility would have been close to optimal, and listeners likely would have judged stimulus pairs by attending to other sound quality descriptors. However, if after several iterations the preferred shaping departed from the prescribed shaping in a way that made speech less intelligible, then listeners may have adjusted their internal criterion and attended to intelligibility more heavily. The fact that some listeners rated Overall Impression and Intelligibility for female speech higher for prescribed shaping than for preferred shaping supports the interpretation that listeners may have shifted their internal preference criterion once intelligibility was no longer ideal. However, while Intelligibility ratings were predictive of Overall Impression ratings, absolute Intelligibility ratings were comparable between prescribed and prescribed gain. Future research is needed with stimuli in which intelligibility is systematically adjusted between conditions.

The finding that Loudness ratings were negatively associated with Overall Impression ratings contrasts with previous research. Loudness ratings have historically been among the sound quality descriptors least and nonsignificantly associated with ratings of overall impression (Davies-Venn et al., 2007; Gabrielsson et al., 1988). In the current study, speech and music were presented at the same level prior to amplification. However, in the study by Davies-Venn et al. (2007), sound quality ratings were gathered for a variety of experimental parameters, one of which was the level of speech. They found that intelligibility ratings were more strongly associated with overall impression ratings for soft speech than for loud speech. In the current study, increases in low-frequency gain may have inadvertently led to the speech stimuli being louder than preferred, which may have led to poorer Overall Impression ratings for the preferred condition than for the prescribed condition.

Stimulus Dependencies

Differences in acoustic content between stimuli may have partially driven stimulus-specific gain adjustments. A given parameter adjustment may have had different perceptual effects across stimuli due to differences in acoustic content. The Beatles pop sample in this study

included drums, whereas the classical sample did not, therefore containing more high-frequency content than the classical sample (Elowsson & Friberg, 2017). Good sound quality depends on an appropriate balance between high- and low-frequency energy (Moore & Tan, 2003), so it is possible that listeners increased the low-frequency gain partly because the Beatles stimulus contained more high-frequency energy. The classical sample contained less low-frequency energy than the pop sample and so a low-frequency gain increase of similar magnitude to that for the pop sample would have been less noticeable. Similarly, Davies-Venn et al. (2007) found that listeners rated popular music as sharper than classical music because the popular music contained more high-frequency energy. Arehart et al. (2011) found that acoustic characteristics from different music genres can affect how the genres will interact with hearing aid processing features. For example, they reported that a large compression ratio, which did not impair sound quality for a continuous vocal signal, did impair sound quality for a jazz stimulus with greater high frequency content, faster rhythm, and wider dynamic range. Together, this evidence suggests that acoustic differences between stimuli can lead to different perceptual consequences for the same parameter adjustments. Further research should investigate whether specific stimulus characteristics can predict the degree to which specific gains will be adjusted.

While the current study evaluated preferred gain settings for speech and music, it did not evaluate preferred gain settings for speech in noise. It is possible that preferred gain settings from this study do not generalize to speech in noise. For instance, the DSL v5.0 method recommends different frequency-gain settings for quiet speech than for noisy speech (Scollie et al., 2005). Similarly, many hearing aid manufacturers apply different gain settings for speech-in-noise programs than speech-in-quiet programs. The low-frequency gain that was preferred for music in this study would likely be inappropriate for speech in noise. Noise typically contains significant low-frequency components, and more low-frequency gain may increase upward spread of masking which would impair speech intelligibility. Future research should explore listener preferences for speech in noise using the simplex procedure.

Generalizability of Preferred Shaping

The findings from the simplex procedure suggest that listener satisfaction may be augmented if listener-specific, stimulus-dependent frequency-gain shaping is applied. However, in practice, it would be cumbersome to determine a unique set of parameters for every new stimulus. We assessed the sound quality of novel stimuli belonging to the same genre as the experimental stimuli

to determine if preferred shaping determined using experimental stimuli would generalize to other similar stimuli. The findings revealed moderate correlations between ratings of experimental and novel stimuli across sound quality descriptors for speech and music, suggesting that preference-based frequency-gain shaping generalizes to novel stimuli; listeners may prefer speech and music shaped using frequency-gain characteristics determined using similar stimuli, relative to prescribed frequency-gain shaping. This is consistent with test–retest sound quality correlations for aided speech and music in the literature. D'Onofrio et al. (2019) found that hearing-impaired listeners rated the sound quality of aided speech and music with moderate reliability between test–retest conditions. Narendran and Humes (2003) also investigated the test–retest reliability of different sound quality ratings for aided speech and music with hearing-impaired listeners and found that correlations for descriptors most similar to those used in this study (clarity, fullness, loudness, and total impression for aided speech and music) were moderate. These findings suggest that listeners experience similar sound quality when listening to novel stimuli using genre-specific frequency-gain shaping to that for listening to identical stimuli for a second time. However, it should be recognized that we explored generalizability between groups of two short stimuli. Further research should investigate generalizability for more realistic listening situations, such as continuous discourse and/or entire musical passages.

Individual Variability

The frequency-gain shaping data presented here reflect preference-driven gain adjustments averaged across individuals. Despite the finding that across stimuli listeners preferred a low-frequency gain increase relative to prescribed frequency-gain shaping, 12% of listeners preferred a decrease in low-frequency gain, while 40% and 32% of listeners preferred mid- and high-frequency gain increases, respectively, despite mean group-level decreases in preferred gain relative to prescribed gain. Further, the magnitude of adjustment varied between listeners. Across stimuli, the 25th and 75th percentiles for low-frequency gain adjustments were 1 dB and 13 dB, respectively. The 25th and 75th percentiles were –6 dB and 4 dB for mid-frequency adjustments and –11 dB and 2 dB for high-frequency adjustments. These findings are consistent with that of Caswell-Midwinter and Whitmer (2020), in which listeners reliably increased the low-frequency gain, whereas other gain adjustments were less consistent in direction and magnitude. This variability may be attributed to individual factors, such as nature of hearing loss, hearing aid experience, age, and cognition. For example, the impact of hearing aid processing on speech intelligibility is associated with age, working memory, and degree of hearing loss (Arehart et al., 2013, 2015; Souza et al., 2015, 2019). For classical music, listeners with greater degrees of hearing loss prefer linear amplification over wide-dynamic range compression (Croghan et al., 2014). While some research has shown that factors such as age, gender, hearing loss, and hearing aid experience do not explain individual variability in self-adjusted gain settings (Perry et al., 2019), further research is needed to explain why this may be the case, as well as to evaluate how other individual factors could affect preferred frequency-gain shaping.

Acoustic Considerations

This study made use of a fully occluding transducer that completely sealed listeners' ear canals. However, occluding hearing aid fittings can be problematic in practice due to the trapping of low-frequency energy in the ear canal. This leads to subjective reports of the occlusion effect in which one's own voice sounds *boomy* (Kuk et al., 2005) and is a common complaint among hearing aid (Ricketts et al., 2019). Therefore, many hearing aids are coupled to the ear using an open fitting, in which the *ear canal is . . . open for directly receiving ambient sounds* (Winkler et al., 2016, p. 4). Open-fit hearing aids are typically prescribed for hearing aid users with milder losses and near-normal thresholds at low frequencies. Relative to closed fits, open fits are usually preferred for speech quality and own-voice perception (Winkler et al., 2016).

This study's findings likely do not generalize to open-fit hearing aids. D'Onofrio et al. (2019) evaluated whether hearing-impaired listeners preferred self-adjusted gain compared with prescribed gain for speech and music. Listeners wore open-fit receiver-in-the-canal hearing aids. Low-frequency gain adjustments were all less than 2 dB from prescribed settings, significantly contrasting with the low-frequency adjustments observed in the current study. Future research should consider whether the sound quality benefits of using an increased bass response in an occluded fit outweigh potential own-voice discomfort from the occlusion effect and whether enough low-frequency amplification can be achieved in a vented fitting using traditional acoustic hearing aids. It should be noted that low-frequency amplification (down to 125 Hz) can be achieved using an open-fitting, wideband direct drive hearing aid (Arbogast et al., 2019) and that listeners prefer stimuli containing more low-frequency energy than stimuli with less low-frequency energy while wearing the direct drive hearing aid (Vaisberg et al., 2020).

Finally, this study used studio-compressed music recordings at a fixed level of 70 dB SPL. Previous studies have found that higher input levels can affect the impact

of hearing aid compression on music sound quality (Davies-Venn et al., 2007; Moore et al., 2011). Higher input levels may also influence preferred gain adjustments. For example, a given low-frequency boost at a high input level may lead to increased upward spread of masking, which could negatively affect sound quality, thus leading listeners to prefer less low-frequency gain at higher input levels. In addition, highly compressed studio music results in smaller crest factors which enables listeners to listen at higher overall levels (Croghan et al., 2016). In contrast, live music has much larger crest factors (Chasin & Hockley, 2014), which produces music peaks that may cause hearing aid output limiting or peak-clipping, both of which can be detrimental to music listening (Davies-Venn et al., 2007). Therefore, further research should seek to understand the relationship between listening levels, crest factors, and hearing aid circuity and how that relationship interacts with listener gain preferences.

Summary and Conclusions

This study evaluated the degree to which hearing-impaired listeners made preference-based adjustments to hearing aid amplification relative to prescribed settings, whether amplification adjustments were stimulus dependent, and whether any sound quality descriptors explained listener preferences. Using a three-dimensional simplex procedure, listeners selected preferred frequency-gain shaping parameters via preference judgments between stimulus pairs varying in gain that deviated from prescribed amplification in low-, mid-, and high-frequency bands. Listeners increased the low-frequency gain and decreased the high-frequency gain by a smaller magnitude. Mid-frequency gain adjustments were not significantly different than prescribed gain. Low-frequency gain was increased by the greatest magnitude for pop music, followed by classical music, and female and male speech. High-frequency gain was decreased by a similar magnitude for pop music and male speech. The gain adjustments were largest for music, and preferences for music were mainly driven by changes in Fullness and Sharpness. Gain adjustments were smaller for speech, and preferences were mainly driven by changes in Intelligibility and Loudness. Perceived intelligibility was an important driver of frequency-gain preferences for speech, for which the gain adjustments relative to prescribed settings were smaller for speech versus music. Therefore, prescribed amplification for mild to moderate hearing loss would generally be appropriate for speech intelligibility. However, it should not be treated as suitable for other types of stimuli. Alternative frequency-gain settings should be considered to improve listener satisfaction for amplified music.

Author Note

All research activities were conducted during author J. M. V.'s previous affiliation at the National Centre for Audiology, Western University, London, Ontario, Canada. Bose Corporation was not involved in this research.

Declaration of Conflicting Interests

Funding

ORCID iD

Jonathan M. Vaisberg [ID] https://orcid.org/0000-0001-9500-1097

Note

1. Data were initially gathered from 26 listeners. However, 4 of the 26 listeners timed out twice for the same stimulus for at least one of the four stimuli. For these listeners, it was not clear if they had understood the task, as two time-outs per stimulus may have implied guessing. Therefore, their data were excluded from the analysis.

References

Abrams, H. B., & Kihm, J. (2015). An introduction to MarkeTrak IX: A new baseline for the hearing aid market. *The Hearing Review*, *22*(6), 16.

American Speech-Language-Hearing Association Ad Hoc Committee on Hearing Aid Selection and Fitting. (1998).

Guidelines for hearing aid fitting for adults. *American Journal of Audiology*, 7(1), 5–13.

Amlani, A. M., & Schafer, E. C. (2009). Application of paired-comparison methods to hearing aids. *Trends in Amplification*, 13(4), 241–259. https://doi.org/10.1177/1084713809352908

Arbogast, T. L., Moore, B. C. J., Puria, S., Dundas, D., Brimacombe, J., Edwards, B., & Levy, S. C. (2019). Achieved gain and subjective outcomes for a wide-bandwidth contact hearing aid fitted using CAM2. *Ear and Hearing*, 40(3), 741–756. https://doi.org/10.1097/AUD.0000000000000661

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. https://doi.org/10.3109/14992027.2010.539273

Arehart, K. H., Souza, P., Baca, R., & Kates, J. K. (2013). Working memory, age and hearing loss: Susceptibility to hearing aid distortion. *Ear & Hearing*, 34(3), 251–260. https://doi.org/10.1097/AUD.0b013e318271aa5e

Arehart, K., Souza, P., Kates, J., Lunner, T., & Pedersen, M. S. (2015). Relationship between signal fidelity, hearing loss and working memory for digital noise suppression. *Ear & Hearing*, 36(5), 505–516. https://doi.org/10.1097/AUD.0000000000000173

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1), 1–48. https://doi.org/10.1121/1.410152

British Society of Hearing Aid Audiologists. (2012). *Guidance on professional practice for Hearing Aid Audiologists*. https://www.bshaa.com/write/MediaUploads/BSHAA Publications/bshaa_guidance_14%5B1%5D.pdf

Byrne, D., Dillin, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M. N., Nasser, N. H. A., El Kohly, W. A. H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., ... Lodvigsen, C. (1994). An international comparison of long-term average speech spectra. *The Journal of the Acoustical Society of America*, 96(4), 2108. https://doi.org/10.1121/1.410152

Caswell-Midwinter, B., & Whitmer, W. M. (2019a). Discrimination of gain increments in speech-shaped noises. *Trends in Hearing*, 23, 1–12. https://doi.org/10.1177/2331216518820220

Caswell-Midwinter, B., & Whitmer, W. M. (2019b). Discrimination of gain increments in speech. *Trends in Hearing*, 23, 1–9. https://doi.org/10.1177/2331216519886684

Caswell-Midwinter, B., & Whitmer, W. M. (2020). The perceptual limitations of troubleshooting hearing-aids based on patients' descriptions. *International Journal of Audiology*. Advance online publication. https://doi.org/10.1080/14992027.2020.1839679

Chasin, M., & Hockley, N. S. (2014). Some characteristics of amplified music through hearing aids. *Hearing Research*, 308, 2–12. https://doi.org/10.1121/1.412980

Condit-Schultz, N., & Huron, D. (2015). Catching the lyrics: Intelligibility in twelve song genres. *Music Perception: An Interdisciplinary Journal*, 32(5), 470–483. https://doi.org/10.1525/MP.2015.32.5.470

Cornelisse, L. E., Seewald, R. C., & Jamieson, D. G. (1995). The input/output formula: A theoretical approach to the fitting of personal amplification devices. *Journal of the Acoustical Society of America*, 97(3), 1854–1864. https://doi.org/10.1121/1.412980

Cox, R. M., Matesich, J. S., & Moore, J. N. (1988). Distribution of short-term rms levels in conversational speech. *Journal of the Acoustical Society of America*, 84(3), 1100–1104. https://doi.org/10.1121/1.396697

Croghan, N. B. H., Arehart, K. H., & Kates, J. M. (2014). Music preferences with hearing aids: Effects of signal properties, compression settings, and listener characteristics. *Ear and Hearing*, 35(5), e170–e184.

Croghan, N. B. H., Swanberg, A. M., Anderson, M. C., & Arehart, K. H. (2016). Chosen listening levels for music with and without the use of hearing aids. *American Journal of Audiology*, 25, 161–166. https://doi.org/10.1044/2016_AJA-15-0078

Davies-Venn, E., Souza, P., & Fabry, D. (2007). Speech and music quality ratings for linear and nonlinear hearing aid circuitry. *Journal of the American Academy of Audiology*, 18(8), 688–699. https://doi.org/10.3766/jaaa.18.8.6

Dirks, D. D., Ahlstrom, J., & Noffsinger, P. D. (1993). Preferred frequency response for two- and three-channel amplification systems. *Journal of Rehabilitation Research and Development*, 30(3), 305–317.

D'Onofrio, K. L., Gifford, R. H., & Ricketts, T. A. (2019). Musician and nonmusician hearing aid setting preferences for music and speech stimuli. *American Journal of Audiology*, 28(2), 333–347. https://doi.org/10.1044/2019_AJA-18-0125

Elowsson, A., & Friberg, A. (2017, May). Long-term average spectrum in popular music and its relation to the level of the percussion [Paper presentation]. 142nd Audio Engineering Society (AES) Convention, Berlin, 12 pp.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, 39(2), 175–191. https://doi.org/10.3758/BF03193146

Field, A., Miles, J., & Field, Z. (2012). *Discovering statistics Using R*. Sage Publications.

Franck, B. A. M., Boymans, M., & Dreschler, W. A. (2007). Interactive fitting of multiple algorithms implemented in the same digital hearing aid. *International Journal of Audiology*, 46(7), 388–397. https://doi.org/10.1080/14992020701261421

Franck, B. A. M., Dreschler, W. A., & Lyzenga, J. (2004). Methodological aspects of an adaptive multidirectional pattern search to optimize speech perception using three hearing-aid algorithms. *Journal of the Acoustical Society of America*, 116(6), 3620–3628. https://doi.org/10.1121/1.1808220

Franks, J. R. (1982). Judgments of hearing aid processed music. *Ear and Hearing*, 3(1), 18–23. https://doi.org/10.1097/00003446-198201000-00004

Gabrielsson, A., Schenkman, B. N., & Hagerman, B. (1988). The effects of different frequency responses on sound

quality judgements and speech intelligibility. *Journal of Speech and Hearing Research*, *31*, 166–177. https://doi.org/10.1044/jshr.3102.166

Gabrielsson, A., & Sjögren, H. (1979). Perceived sound quality of hearing aids. *Scandinavian Audiology*, *8*(3), 155–169. https://doi.org/10.3109/01050397909076317

Greasley, A., Crook, H., & Fulford, R. (2020). Music listening and hearing aids: Perspectives from audiologists and their patients. *International Journal of Audiology*, *59*(9), 694–706. https://doi.org/10.1080/14992027.2020.1762126

Herzke, T., Kayser, H., Loshaj, F., Grimm, G., & Hohmann, V. (2017). Open signal processing software platform for hearing aid research (openMHA). In *Proceedings of the Linux Audio Conference* (pp. 35–42). Universite Jean Monnet.

Hillenbrand, J. M., Getty, L. A., Clark, M. J., & Wheeler, K. (1995). Acoustic characteristics of American English vowels. *Journal of the Acoustical Society of America*, *97*(5), 3099–3111. https://doi.org/10.1121/1.411872

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. https://doi.org/10.3109/14992027.2010.506889

Humes, L. E. (2003). Modeling and predicting hearing aid outcome. *Trends in Amplification*, *7*(2), 41–75. https://doi.org/10.1177/108471380300700202

Institute of Electrical and Electronics Engineers [IEEE]. (1969). IEEE recommended practice for speech quality measurements (IEEE Standards Publication No. 297-1969). *IEEE Transactions on Audio and Electroacoustics*, *AU-17*(3), 225–246.

Jenstad, L. M., Bagatto, M. P., Seewald, R. C., Scollie, S. D., Cornelisse, L. E., & Scicluna, R. (2007). Evaluation of the desired sensation level [input/output] algorithm for adults with hearing loss: The acceptable range for amplified conversational speech. *Ear and Hearing*, *28*(6), 793–811. https://doi.org/10.1097/AUD.0b013e318157670a

Johnson, E. E. (2013). An initial-fit comparison of two generic hearing aid prescriptive methods (NAL-NL2 and CAM2) to individuals having mild to moderately severe high-frequency hearing loss. *Journal of the American Academy of Audiology*, *24*(2), 138–150. https://doi.org/10.3766/jaaa.24.2.7

Keidser, G., Dillon, H. R., Flax, M., Ching, T., & Brewer, S. (2011). The NAL-NL2 prescription procedure. *Audiology Research*, *1*(e24), 88–90. https://doi.org/10.4081/audiores.2011.e24

Kirchberger, M., & Russo, F. A. (2016). Dynamic range across music genres and the perception of dynamic compression in hearing-impaired listeners. *Trends in Hearing*, *20*, 1–16. https://doi.org/10.1177/2331216516630549

Kuk, F., Keenan, D., & Lau, C.-C. (2005). Vent configurations on subjective and objective occlusion effect. *Journal of the American Academy of Audiology*, *16*(9), 747–762. https://doi.org/10.3766/jaaa.16.9.11

Kuk, F. K., & Lau, C. C. (1995). Effect of initial setting on convergence to optimal hearing aid setting using a simplex method. *British Journal of Audiology*, *29*(5), 263–269. https://doi.org/10.3109/03005369509076741

Kuk, F. K., & Lau, C. (1996). Comparison of preferred frequency gain settings obtained with category rating and modified simplex procedure. *Journal of the American Academy of Audiology*, *7*, 322–331. https://doi.org/10.1097/AUD.0000000000000617

Kuk, F. K., & Pape, N. M. C. (1992). The reliability of the modified simplex procedure in hearing aid frequency response selection. *Journal of Speech and Hearing Research*, *35*(2), 418–429. https://doi.org/10.1044/jshr.3502.418

Kuk, F. K., & Pape, N. M. C. (1993). Relative satisfaction for frequency responses selected with a simplex procedure in different listening conditions. *Journal of Speech and Hearing Research*, *36*(1), 168–177. https://doi.org/10.1044/jshr.3601.168

Lawrence, M. A. (2016). *ez: Easy analysis and visualization of factorial experiments. R package version 4.4-0*. https://cran.r-project.org/package = ez

Leek, M. R., Molis, M. R., Kubli, L. R., & Tufts, J. B. (2008). Enjoyment of music by elderly hearing-impaired listeners. *Journal of the American Academy of Audiology*, *19*(6), 519–526. https://doi.org/10.3766/jaaa.19.6.7

Looi, V., Rutledge, K., & Prvan, T. (2019). Music appreciation of adult hearing aid users and the impact of different levels of hearing loss. *Ear and Hearing*, *40*(3), 529–544. https://doi.org/10.1097/AUD.0000000000000632

Madsen, S. M. K., & Moore, B. C. J. (2014). Music and hearing aids. *Trends in Hearing*, *18*, 1–29. https://doi.org/10.1177/2331216514558271

Moore, B. C. J. (2016). Effects of sound-induced hearing loss and hearing aids on the perception of music. *Journal of the Audio Engineering Society*, *64*(3), 112–123. https://doi.org/10.17743/jaes.2015.0081

Moore, B. C. J., Baer, T., Ives, D. T., Marriage, J., & Salorio-Corbetto, M. (2016). Effects of modified hearing aid fittings on loudness and tone quality for different acoustic scenes. *Ear and Hearing*, *37*(4), 483–491. https://doi.org/10.1097/AUD.0000000000000285

Moore, B. C. J., Füllgrabe, C., & Stone, M. A. (2011). Determination of preferred parameters for multichannel compression using individually fitted simulated hearing aids and paired comparisons. *Ear and Hearing*, *32*(5), 556–568. https://doi.org/10.1097/AUD.0b013e31820b5f4c

Moore, B. C. J., Glasberg, B. R., & Stone, M. A. (2010). Development of a new method for deriving initial fittings for hearing aids with multi-channel compression: CAMEQ2-HF. *International Journal of Audiology*, *49*(3), 216–227. https://doi.org/10.3109/14992020903296746

Moore, B. C. J., & Sęk, A. (2013). Comparison of the CAM2 and NAL-NL2 hearing aid fitting methods. *Ear and Hearing*, *34*(1), 83–95. https://doi.org/10.1097/AUD.0b013e3182650adf

Moore, B. C. J., & Sęk, A. (2016). Comparison of the CAM2A and NAL-NL2 hearing-aid fitting methods for participants with a wide range of hearing losses. *International Journal of Audiology*, *55*, 93–100. https://doi.org/10.3109/14992027.2015.1095360

Moore, B. C. J., & Tan, C.-T. (2003). Perceived naturalness of spectrally distorted speech and music. *The Journal of the Acoustical Society of America*, *114*(1), 408–419.

Narendran, M. M., & Humes, L. E. (2003). Reliability and validity of judgments of sound quality in elderly hearing aid wearers. *Ear and Hearing*, *24*(1), 4–11. https://doi.org/10.1097/01.AUD.0000051745.69182.14

Nelson, P. B., Perry, T. T., Gregan, M., & Van Tasell, D. (2018). Self-adjusted amplification parameters produce large between-subject variability and preserve speech intelligibility. *Trends in Hearing*, *22*, 1–13. https://doi.org/10.1177/2331216518798264

Neuman, A. C., Levitt, H., Mills, R., & Schwander, T. (1987). An evaluation of three adaptive hearing aid selection strategies. *The Journal of the Acoustical Society of America*, *82*(6), 1967–1976. https://doi.org/10.1121/1.395641

Olsen, W. O. (1998). Average speech levels and spectra in various speaking/listening conditions: A summary of the Pearson, Bennett, & Fidell (1977) report. *American Journal of Audiology*, *7*(2), 21–25. https://doi.org/10.1044/1059-0889(1998/012)

Perry, T. T., Nelson, P. B., & Van Tasell, D. J. (2019). Listener factors explain little variability in self-adjusted hearing aid gain. *Trends in Hearing*, *23*, 1–17. https://doi.org/10.1177/2331216519837124

Polonenko, M. J., Scollie, S. D., Moodie, S., Seewald, R. C., Laurnagaray, D., Shantz, J., & Richards, A. (2010). Fit to targets, preferred listening levels, and self-reported outcomes for the DSL v5.0 a hearing aid prescription for adults. *International Journal of Audiology*, *49*(8), 550–560. https://doi.org/10.3109/14992021003713122

Preminger, J. E., Neuman, A. C., Blakke, M. H., Deirdre, W., & Levitt, H. (2000). An examination of the practicality of the simplex procedure. *Ear and Hearing*, *21*(3), 177–193. https://doi.org/10.1097/00003446-200006000-00001

Preminger, J. E., & Van Tasell, D. J. (1995). Quantifying the relation between speech quality and speech intelligibility. *Journal of Speech and Hearing Research*, *38*, 714–725. https://doi.org/10.1044/jshr.3803.714

Punch, J. L. (1978). Quality judgments of hearing aid-processed speech and music by normal and otopathologic listeners. *Journal of the American Audiology Society*, *3*(4), 179–188.

Ricketts, T., Bentler, R. A., & Mueller, G. H. (2019). Chapter 9: Ear impressions, earmolds, and associated plumbing. In B. A. Stach (Ed.), *Essentials of modern hearing aids: Selection, fitting, and verification* (pp. 273–306). Plural Publishing.

Ricketts, T. A., Dittberner, A. B., & Johnson, E. E. (2008). High-frequency amplification and sound quality in listeners with normal through moderate hearing loss. *Journal of Speech, Language, and Hearing Research*, *51*(1), 160–172. https://doi.org/10.1044/1092-4388(2008/012)

Scollie, S., Seewald, R., Cornelisse, L., Moodie, S., Bagatto, M., Laurnagaray, D., ... Pumford, J. (2005). The desired sensation level multistage input/output algorithm. *Trends in Amplification*, *9*(4), 159–197. https://doi.org/10.1177/108471380500900403

Souza, P., Arehart, K., Schoof, T., Anderson, M., Strori, D., & Balmert, L. (2019). Understanding variability in individual response to hearing aid signal processing in wearable hearing aids. *Ear and Hearing*, *40*(6), 1280–1292. https://doi.org/10.1097/AUD.0000000000000717

Souza, P. E., Arehart, K. H., Shen, J., Anderson, M., & Kates, J. M. (2015). Working memory and intelligibility of hearing-aid processed speech. *Frontiers in Psychology*, *6*(May), 526. https://doi.org/10.3389/fpsyg.2015.00526

Stelmachowicz, P. G., Lewis, D. E., & Carney, E. (1994). Preferred hearing aid frequency responses in simulated listening environments. *Journal of Speech and Hearing Research*, *37*(3), 712–719. https://doi.org/10.1044/jshr.3703.712

R Core Team. (2017). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.r-project.org/

Vaisberg, J. M., Folkeard, P., Levy, S., Dundas, D., Agrawal, S., & Scollie, S. (2020). Sound quality ratings of amplified speech and music using a direct drive hearing aid. *Otology & Neurotology,* *42*(2), 227–234. https://doi.org/10.1097/mao.0000000000002915

Vaisberg, J. M., Martindale, A. T., Folkeard, P., & Benedict, C. (2019). A qualitative study of the effects of hearing loss and hearing aid use on music perception in performing musicians. *Journal of the American Academy of Audiology*, *30*(10), 856–870. https://doi.org/10.3766/jaaa.17019

Valente, M., Abrams, H., Benson, D., Chisolm, T., Citron, D., Hampton, D., ... Sweetow, R. (2006). Guidelines for the audiologic management of adult hearing impairment. *Audiology Today*, *18*(5), 1–44. http://audiology-web.s3.amazonaws.com/migrated/haguidelines.pdf_53994876e92e42.70908344.pdf

van Buuren, R. A., Festen, J. M., & Plomp, R. (1995). Evaluation of a wide range of amplitude-frequency responses for the hearing impaired. *Journal of Speech and Hearing Research*, *38*(1), 211–221. https://doi.org/10.1044/jshr.3801.211

Van Eeckhoutte, M., Folkeard, P., Glista, D., & Scollie, S. (2020). Speech recognition, loudness, and preference with extended bandwidth hearing aids for adult hearing aid users. *International Journal of Audiology*, *59*(10), 780–791. https://doi.org/10.1080/14992027.2020.1750718

Winkler, A., Latzel, M., & Holube, I. (2016). Open versus closed hearing-aid fittings: A literature review of both fitting approaches. *Trends in Hearing*, *20*, 1–13. https://doi.org/10.1177/2331216516631741