

## The perceptual limitations of troubleshooting hearing-aids based on patients' descriptions

Benjamin Caswell-Midwinter & William M. Whitmer

To cite this article: Benjamin Caswell-Midwinter & William M. Whitmer (2021) The perceptual limitations of troubleshooting hearing-aids based on patients' descriptions, International Journal of Audiology, 60:6, 427-437, DOI: [10.1080/14992027.2020.1839679](https://doi.org/10.1080/14992027.2020.1839679)

To link to this article: <https://doi.org/10.1080/14992027.2020.1839679>



© 2020 The Authors. Published by Informa UK Limited, trading as Taylor & Francis Group on behalf of British Society of Audiology, International Society of Audiology, and Nordic Audiological Society.



Published online: 11 Nov 2020.



Submit your article to this journal [↗](#)



Article views: 1023



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)

## The perceptual limitations of troubleshooting hearing-aids based on patients' descriptions

Benjamin Caswell-Midwinter<sup>a,b,c</sup>  and William M. Whitmer<sup>a,b</sup> 

<sup>a</sup>Hearing Sciences – Scottish Section, Division of Clinical Neuroscience, School of Medicine, University of Nottingham, Glasgow, UK; <sup>b</sup>School of Medicine, Dentistry, and Nursing, College of Medical, Veterinary and Life Science, University of Glasgow, Glasgow, UK; <sup>c</sup>Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear, Harvard Medical School, Boston, MA, USA

### ABSTRACT

**Objectives:** Hearing-aid frequency-gain responses are routinely adjusted by clinicians to patient preferences and descriptions. This study measured the minimum gain adjustments required to elicit preferences, and the assignment of descriptors to gain adjustments, to perceptually evaluate description-based troubleshooting.

**Design:** Participants judged whether short sentences with  $\pm 0$ –12 dB gain adjustments in one of three frequency bands were “better”, “worse” or “no different” from the same sentence at their individual real-ear or prescribed gain. If judged “better” or “worse”, participants were then asked to assign one of the six common sound-quality descriptors to their preference.

**Study sample:** Thirty-two adults (aged 51–75 years) all with hearing-aid experience.

**Results:** Median preference thresholds, the minimum gain adjustments to elicit “better” or “worse” judgments, ranged from 4 to 12 dB, increasing with frequency. There was some between-participant agreement in preferences: participants generally preferred greater low-frequency gain. Within-participant reliability for preferences was moderate. There was, however, little between-participant agreement in descriptor selection for gain adjustments. Furthermore, within-participant reliability for descriptor selection was lacking.

**Conclusions:** The scale of gain adjustments necessary to elicit preferences, along with the low agreement and reliability in descriptors for these adjustments questions the efficiency and efficacy of current description-based troubleshooting, especially with short speech stimuli.

### ARTICLE HISTORY

Received 25 May 2020  
Revised 27 August 2020  
Accepted 13 October 2020

### KEYWORDS

Hearing-aid fitting; fine-tuning; gain; descriptors

### Introduction

Patient feedback is regularly used to fine-tune the electroacoustical parameters of hearing devices in the clinic (Anderson, Arehart, and Souza 2018; Jenstad, Van Tasell, and Ewert 2003; Kuk and Ludvigsen 1999; Thielemans et al. 2017). Despite being an everyday practice, there is evidence to suggest it is of little benefit (Cunningham, Williams, and Goldsmith 2001; Saunders, Lewis, and Forsline 2009). With a vast space of parameters and parameter combinations available in modern hearing devices, many adjustments may not even be noticeable for the patient to make informed comparisons. Previous research has provided evidence on noticeable adjustments of compression (Gilbert et al. 2008; Nabelek 1984; Sabin, Gallun, and Souza 2013) and speech-to-noise ratio (McShefferty, Whitmer, and Akeroyd 2015), to attempt to reduce the parameter space to one that is perceptually relevant. There has however, been little direct evidence on noticeable adjustments of the frequency-gain response.

Gain, the fundamental hearing-aid parameter for restoring audibility, is routinely adjusted towards prescribed targets in real-ear verification and away from targets to personalise fittings using patient feedback (Anderson, Arehart, and Souza 2018; Jenstad, Van Tasell, and Ewert 2003; Kuk and Ludvigsen 1999;

Thielemans et al. 2017). Gain can also be adjusted by patients themselves in self-fitting devices (Keidser and Convery 2016; Nelson et al. 2018; Sabin et al. 2020). Given this, the authors previously measured discrimination thresholds (just-noticeable differences: JNDs) for gain adjustments. The JNDs measured with speech-shaped noises were approximately 3 dB for increments of octave-band width (0.5–4 kHz) and 1.5 dB for broadband increments (Caswell-Midwinter and Whitmer 2019a), providing a psychophysical baseline for gain adjustments in clinical practice. Compared to steady-state noise, the JNDs measured with male, single-talker sentences were larger, the more so the narrower the bandwidth being adjusted; 6–10 dB for octave-band increments, 4–7 dB for wideband increments, and 2 dB for broadband increments (Caswell-Midwinter and Whitmer 2019b). The scale of these JNDs is partly explained by the sparseness of energy in any one frequency band at any given time in sentences. These JNDs indicate the limitations of using short sentences as the stimulus for adjusting gain in response to patient feedback.

In the abovementioned studies, participants made psychophysical same-different judgments on the objective acoustic equivalence of stimuli, and discrimination thresholds were measured. Judgments on a gain adjustment less than a discrimination

**CONTACT** Benjamin Caswell-Midwinter  [benjamin\\_caswellmidwinter@meei.harvard.edu](mailto:benjamin_caswellmidwinter@meei.harvard.edu)  Otolaryngology – Head and Neck Surgery, Massachusetts Eye and Ear, 243 Charles St, Boston 02114, MA, USA

© 2020 The Authors. Published by Informa UK Limited, trading as Taylor & Francis Group on behalf of British Society of Audiology, International Society of Audiology, and Nordic Audiological Society.

This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

threshold will be inefficient and will result in meaningless patient feedback, as the adjustment will not be perceptually distinguishable from the previous setting. However, when comparing adjustments in the clinic, patients make preference judgments on the subjective nature of stimuli. Preference judgments are more complex than psychophysical judgments given that they are subjective, and have real-world consequence. While an adjustment must be of at least 1 JND to elicit a preference for one setting over another, it is unclear whether a JND adjustment itself is sufficient to elicit a preference, and if not, what magnitude does so.

Previous research contrasting discrimination and subjective judgments on acoustical parameters has reported that speech-to-noise ratio adjustments needed to elicit a preference need to be greater than the corresponding JND adjustments (McShefferty, Whitmer, and Akeroyd 2016). However, there has been no research directly comparing these adjustments for gain. Keidser, Dillon, and Convery (2008) had hearing-impaired participants make forced-choice preference judgments between different gain settings and then rate the degree of difference. The root-mean-square dB difference between gains increased with difference ratings, and 83% of participants perceived the gains compared as different. However, only 25% had reliable preferences, suggesting that noticeably different gains do not necessarily result in a stable preference. The comparison to the current application is limited, as discrimination ability was not measured; participants rated stimuli on a subjective scale which cannot be acoustically defined (i.e. “somewhat different”). Furthermore, standard gains and the test stimuli presented varied (e.g. monologue in car, soft dialog in library), and therefore the difference ratings, collapsed across gains and stimuli, cannot be precisely interpreted.

Supplemental research on gain preferences has reported 3–5 dB broadband gain adjustments in speech in noise are needed to elicit changes in sound quality, intelligibility, and pleasantness ratings (Byrne and Dillon 1986; Dirks, Ahlstrom, and Noffsinger 1993; Jenstad et al. 2007), which is greater than the 2 dB broadband JNDs measured for speech in quiet (Caswell-Midwinter and Whitmer 2019b; Whitmer and Akeroyd 2011). This suggests an increase in magnitude between just-noticeable and just-preferable adjustments. However, the comparison is limited as the JNDs were measured in quiet, favourable listening conditions, while the minimum adjustments for gain preferences, inferred from subjective judgments, were measured in noise. Furthermore, the adjustments for gain preferences were only reported as broadband adjustments.

The first primary objective of the current study was to estimate the minimum gain adjustments required to elicit preferences from hearing-impaired listeners, and to do so in a manner making it possible to closely compare these thresholds for preference with thresholds for discrimination (Caswell-Midwinter and Whitmer 2019b). Secondary objectives here were to (a) report the underlying preferences and examine (b) whether there was agreement between participants, and (c) whether participants’ own preferences were reliable.

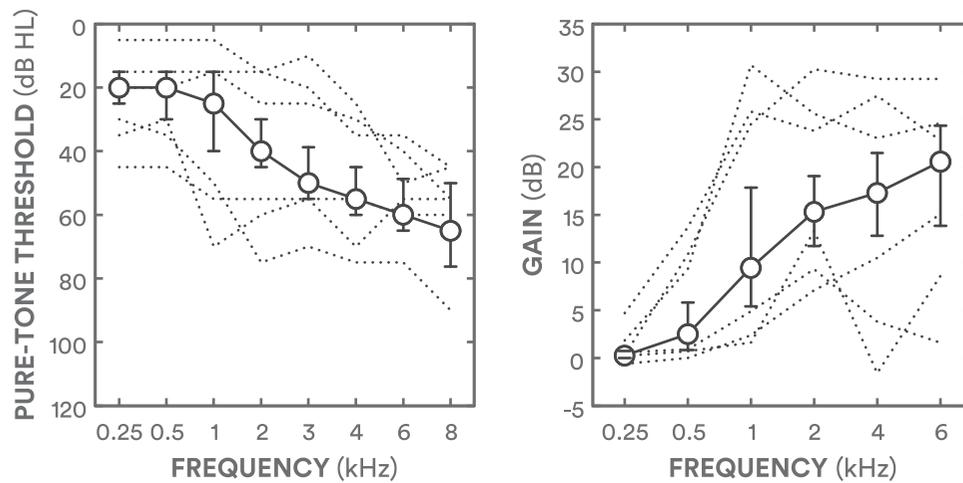
Patients’ descriptions of the acoustics of their devices are also often translated into gain adjustments by clinicians, particularly in troubleshooting. This approach assumes that listeners assign descriptors to gains, in level and frequency, in a consistent manner. However, the evidence on this approach is neither substantive nor conclusive. Presenting speech, music and everyday sounds, Gabrielsson and Sjögren (1979) and Gabrielsson et al. (1990) reported high between-participant

agreement and within-participant reliability for the mapping of descriptors to gains. However, a limited number of frequency-gain responses were judged in these studies, possibly insufficient to reflect individual differences. Additionally, the gains were highly distinct in level and frequency, lacking relevance to current fine-tuning. In Gabrielsson et al. (1990), four gains were presented: one flat, and three others with 9 dB low-, mid- or high-frequency increments.

Across-listener adjustment of gains to descriptors has also been supported by clinicians in Jenstad, Van Tasell, and Ewert’s (2003) troubleshooting system. Given the difficulties of precisely measuring descriptor judgments on interacting parameters in real devices (from a variety of manufacturers), Jenstad et al. aimed to provide an expert-based starting point for troubleshooting. They established a vocabulary of common problem descriptors and solutions by surveying American audiologists. Factor analyses grouped similar descriptors under electroacoustical components (of the suggested cause), for example, “in a barrel, tunnel, well” and “hollow” under “+ LF gain” (excessive low-frequency gain). The factor analysis explained 90.8% of the variation in data. Troubleshooting adjustments were ordered according to clinician preference. For example, for the descriptor “not clear”, the solutions were “increase high-frequency gain”, “decrease high-frequency compression ratio” and “increase overall gain”. While there was high agreement, it was established with a limited selection of unspecific adjustments. This expert system has been influential, forming the basis for automated fitting assistants in many hearing-aid manufacturers’ fitting software systems (personal communications, January 2017; Curran and Galster 2013). The expert system method has also been replicated in other languages (Thielemans et al. 2017). Despite this application, it has not been further developed, and there has been little perceptual evaluation of it, beyond a study by Sabin et al. (2011).

Presenting short sentences, Sabin et al. (2011) had hearing-impaired participants rate two pairs of descriptors, reported to be associated in Jenstad, Van Tasell, and Ewert (2003), in relation to gain adjustments from NAL-R standards. While across-participant descriptor judgments broadly agreed with Jenstad et al.’s findings, there was some between-participant disagreement. For example, one participant reported “hollow” to correspond to a positive sloping spectral tilt, opposite to another participant. Within-participant variation was also reported between associated descriptors. Another participant, who mapped “tinny” to a 1 kHz peak mapped the associated descriptor “sharp” to a 3 kHz peak. It was also reported that many participants’ own ratings were inconsistent. Furthermore, large variation in some weighting functions indicated that certain descriptors were not meaningful to some participants. This study suggests that the across-listener translation of descriptors into gains may not be wholly valid given the individual variation in meaning and electroacoustic mapping.

The second primary objective of the current study was to measure the assignment of common descriptors to gain adjustments by hearing-impaired listeners and assess agreement. Secondary objectives here were (a) to compare this assignment of descriptors to gain adjustments to those suggested by clinicians in Jenstad, Van Tasell, and Ewert (2003), and (b) to also examine whether participants’ own descriptor judgments were reliable. Adjustments from gain standards were made to short sentences, and participants assigned a descriptor (from a closed set) to each of their “better” or “worse” judgments.



**Figure 1.** Median audiogram and gains. The left panel shows the median pure-tone thresholds for participants' tested ears. The right panel shows the median gains for participants' tested ears. Error bars show interquartile ranges (25–75%). The dashed lines in the left panel show the better-ear thresholds of participants with the three lowest and three highest BE4FA thresholds. The dashed lines in the right panel show three lowest and three highest gains (averaged across frequencies).

## Methods

### Participants

Thirty-two hearing-impaired participants (13 females) were recruited from the sample of participants who completed the JND task in Caswell-Midwinter and Whitmer (2019b). The median age of participants was 68.5 years (range 51–75 years). All participants had experience wearing hearing-aids. The median hearing-aid experience was 9.5 years (range 2–35 years). At the time of testing, 22 participants had worn hearing-aids bilaterally, and 10 participants had worn a hearing-aid unilaterally. Seventeen participants reported wearing hearing-aids all of the time, 10 reported wearing hearing-aids some of the time, and five reported no longer wearing hearing-aids.

Stimuli were presented monaurally to better ears (BEs), determined by the lower four-frequency pure-tone average (4FA) of their ears. BE4FAs were calculated by averaging pure-tone thresholds at 0.5, 1, 2, and 4 kHz. Figure 1 shows the median better-ear audiogram across all participants. The median BE4FA was 35 dB HL (range 13–59 dB HL). All participants had sensorineural hearing loss: none had conductive elements to their hearing loss (based on differences between air and bone conduction thresholds exceeding 20 dB when averaged over three out of five frequencies at 0.5, 1, 2, 3, and 4 kHz; British Academy of Audiology 2016).

Ethical approval for the study was given by the University of Glasgow research ethics system committee (application number 200160138). All participants provided written informed consent.

### Stimuli

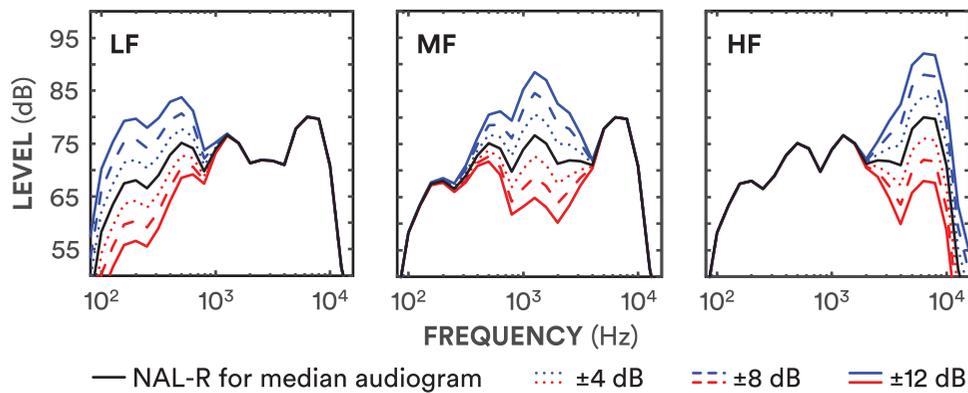
Sentences from the Bamford-Kowal-Bench corpus, spoken by a native British English speaker (Bench, Kowal, and Bamford 1979), were presented monaurally in quiet to better-hearing ears with circumaural headphones (AKG K702, Vienna, Austria). Each sentence is declarative, typically consisting of five words, for example, "they had a lovely day". The spectrum of the corpus was limited with a steep low-pass filter at 10 kHz.

For each comparison, a single sentence was randomly selected from a sample of 336 to be the standard and adjusted stimuli (i.e. comparisons were made using the same sentence). All stimuli were presented with gains. The real-ear insertions gains of 20 participants' devices (worn on their better-hearing

ears) were measured using the Siemens Unity Probe Microphone Hearing Instrument Analyser (Munich, Germany). The authors aimed to measure insertion gains for all participants, however, NAL-R gains (Byrne and Dillon 1986) were calculated for 12 participants' better-hearing ear audiograms. These participants either did not bring their devices to the laboratory or no longer had their devices. NAL-R or real-ear insertion gains were applied to the spectrum of all stimuli in a 0.25 kHz low-pass band, four octave bands centred at 0.5, 1, 2 and 4 kHz, and a 6 kHz high-pass band.

Standard stimuli were sentences plus real-ear or prescribed gains. Adjusted stimuli were sentences plus real-ear or prescribed gains, plus an incrementing or decrementing gain adjustment of 4, 8 or 12 dB at one of three frequency bands (see Figure 2). Identical, control adjustments of 0 dB were also implemented for each frequency band. The 18 experimental adjustments and three control adjustments totalled 21 adjustments. The three frequency bands consisted of a low-frequency band combining 0.25 (low-pass) and 0.5 kHz (octave) bands (LF), a mid-frequency band combining 1 and 2 kHz octave bands (MF), and a high-frequency band combining the 4 (octave) and 6 kHz (high-pass) bands (HF). Stimuli were generated as in Caswell-Midwinter and Whitmer (2019b) by convolving sentences with a 140-tap finite impulse response filter developed for NAL-R equalisation by Kates and Arehart (2010). Filters were designed using the fir2 function in MATLAB (version 9.0.0, The Mathworks, Inc., Natick, MA). While the passbands were as planned, the transition bands for LF and MF bands were slightly broader than intended due to an insufficient filter order for the sampling frequency (see Caswell-Midwinter and Whitmer 2019b for further details).

Stimuli were calibrated (using a Bruel & Kjaer Artificial Ear 4152 and Sound Level Meter 2260, Naerum, Denmark) so that without gain, the overall long-term A-weighted presentation level was 60 dB. The adjustments were also confirmed with the sound level metre. Audibility of the sentences was checked with the participant after practice comparisons by one of the authors. Each presentation was separated by a silent inter-stimulus interval of 375 ms. There were two possible stimulus combinations for each comparison: standard-adjustment and adjustment-standard. Stimulus combinations were counterbalanced and presented randomly.



**Figure 2.** Spectra for adjustments. Each panel demonstrates the filter output (averaged across sentences) for each band-specific 4–12 dB (see legend for line specification) increment (blue) and decrement (red) from prescribed NAL-R gains for the median audiogram (solid black line).

## Procedure

The experiment was conducted in a single hour session. Unmasked pure-tone thresholds were measured with each participant for a prior study (Caswell-Midwinter and Whitmer 2019b) within 6 months of the current study. Real-ear insertion gains were first measured for participants who wore hearing-aids for their better-ears. If participants did not have a hearing-aid for their better-ear (or they did not wear one to the appointment), NAL-R gains were calculated. Participants were seated in an audiometric booth and instructed to imagine they were selecting hearing-aid settings in a clinic for real-world use. Participants made responses on a touch-screen monitor.

An unforced-choice paired-comparisons design (Punch, Rakerd, and Amlani 2001) was used. Participants were asked on each comparison to listen to each presentation and decide “how did the second sentence sound compared to the first?” Participants first made a preference judgment, selecting either “better”, “worse” or “no different.” If “no different” was selected, then the next comparison began. If “better” or “worse” was selected, participants were then prompted with the same question to make a descriptor judgment, selecting one of seven comparative descriptor options. If participants selected “better”, the seven options were: “less unclear”, “less muffled”, “less hollow”, “less in a barrel, tunnel, well”, “less sharp”, “less tinny” and “none of the above”. If participants selected “worse”, the seven options were: “more unclear”, “more muffled”, “more hollow”, “more in a barrel, tunnel, well”, “more sharp”, “more tinny” and “none of the above”. Participants were instructed to select the most representative descriptor. These descriptors were commonly reported by clinicians, and associated pairs (“unclear” and “muffled”, “hollow” and “in a barrel, tunnel, well”, and “sharp” and “tinny”) were reported to have similar meanings and electroacoustical profiles by the clinicians in Jenstad, Van Tasell, and Ewert (2003). The placement of the descriptors on the monitor was initially randomised and then held constant for the study (although “none of the above” was always placed at the bottom).

There were 18 experimental adjustments (both increments (+) and decrements (-) of 4 dB, 8 dB and 12 dB for each of the LF, MF and HF bands) and three control adjustments (0 dB for all three bands). Participants compared each adjustment to their standard 10 times, totalling 210 comparisons. Comparisons were concatenated and presented randomly over two blocks (excluding practice comparisons). Participants completed two blocks of 105 comparisons, each lasting approximately 20 min. Twenty additional practice comparisons were embedded at the start of the first block. There were two possible stimulus combinations, which were counterbalanced: standard-adjustment or adjustment-standard.

## Analyses

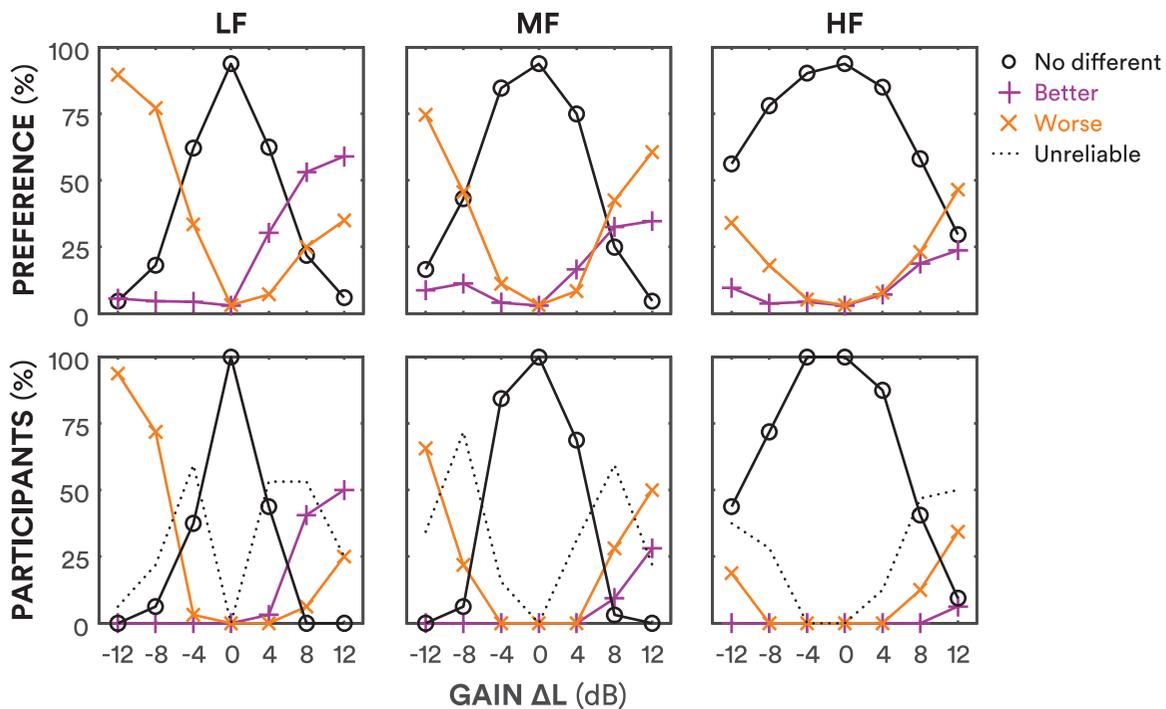
### Preferences

Each participant made a total of 210 comparisons, 10 for each adjustment (including control adjustments). With two stimulus combinations and judgments made in reference to the second stimulus compared to the first, judgments were coded as whether the adjustment was judged “better” or “worse” than the standard. Adjustments were coded as “better” if the participant judged the adjustment to be “better”, or the standard to be “worse”. Adjustments were coded as “worse” if the participant judged the adjustment to be “worse”, or the standard to be “better”. Judgments on control adjustments were collapsed across frequencies.

Between-participant preference agreement was calculated as Fleiss’  $\kappa$  (Fleiss 1971). This calculation used a matrix of adjustments (six rows) and responses (three columns) in which each cell listed the number of times that “better”, “worse” or “no different” was most frequently judged by participants at that adjustment. Adjustments were collapsed in this analysis across increments (+4 dB to +12 dB) and decrements (-4 dB to -12 dB) at each frequency band. As Fleiss’  $\kappa$  assumes rater independence across conditions, Krippendorff’s  $\alpha$  (Hayes and Krippendorff 2007) was also calculated, but produced identical results so is not reported. Within-participant preference reliability was calculated for each adjustment. Participants compared each adjustment to their standard 10 times, and seven or more identical “better”, “worse” or “no different” judgments determined a reliable judgment for that adjustment, based on binomial probability theory (Kuk and Lau 1995). An average percentage of reliable responses for adjustments (from 18 experimental adjustments) is reported. Both agreement and reliability calculations excluded control adjustment data.

### Preference thresholds

The minimum adjustment to elicit a preference, a “better” or “worse” judgment from a “no different” judgment, was estimated for each frequency band and adjustment direction. The authors refer to these adjustments as preference thresholds. Logistic functions were fit to summed “better” and “worse” percentages for separate gain increments and decrements for each frequency band. Preference thresholds were estimated at the adjustment which summed “better” and “worse” values equalled 55% (“no different” values equalled 45%), which approximately corresponds to  $d' = 1$  for an unbiased differencing observer in a same-different task (Macmillan and Creelman 2005). Median



**Figure 3.** Total preferences across participants, and participants' reliable preferences. The top panel shows preferences across participants for each adjustment. Purple lines with pluses show the percentage of “better” judgments, the orange lines with crosses show the percentage of “worse” judgments, and the black lines with circles show the percentage of “no different” judgments. Judgments for control adjustments were averaged and collapsed across frequencies. The bottom panel shows the percentage of participants with reliable responses for each adjustment. The purple lines with pluses show the percentage of participants with reliable “better” judgments. The orange lines with crosses show the percentage of participants with reliable “worse” judgments. The black lines with circles show the percentage of participants with reliable “no different” judgments. The dotted lines show the percentage of participants with unreliable judgments.

thresholds are reported, as visual inspection and Shapiro–Wilk tests indicated that LF increment and decrement thresholds, and HF decrement thresholds were not normally distributed ( $W=0.91, 0.81$  and  $0.92$ , respectively;  $p < 0.01$  for all). Three thresholds were excluded because of poor fits resulting in extreme values.

### Descriptors

As with preferences, descriptor judgments were coded as whether the adjustment was “better” or “worse” than the standard, and the percentage of descriptor judgments were measured separately for “better” (i.e. “less tinny”) and “worse” (i.e. “more tinny”) preferences in reference to each adjustment. For example, if the participant judged the standard to be “less tinny” than an adjustment, then this would be coded as a “more tinny” judgment for the adjustment.

Between-participant descriptor agreement was calculated as Fleiss'  $\kappa$  (Fleiss 1971), as with preferences. This calculation used a matrix of adjustments (six rows) and descriptors (13 columns; 12 “better” and “worse” descriptors and a single collapsed null descriptor) in which each cell listed the number of times that descriptor was most frequently assigned by participants to that adjustment. As with between-participant preference reliability calculations, increments and decrements were collapsed, and control adjustment data were excluded. Within-participant descriptor reliability was calculated similarly to within-participant preference reliability. The number of descriptor judgments made for each adjustment varied between participants as descriptor judgments were not made for “no different” judgments. Therefore, the total number of descriptor judgments made for each adjustment was calculated first, and then the number of identical descriptor judgments required to be classified as reliable

from that total was calculated (e.g. a descriptor judgment was deemed reliable if made four times or more in a total of nine judgments, where the remaining judgment was “no different”). This calculation accounted for the greater number of descriptor alternatives available compared to preference judgments. Within-participant descriptor reliability was only calculated for adjustments in which descriptor judgments were made six times or more (“no different” was selected four times or less in 10 comparisons).

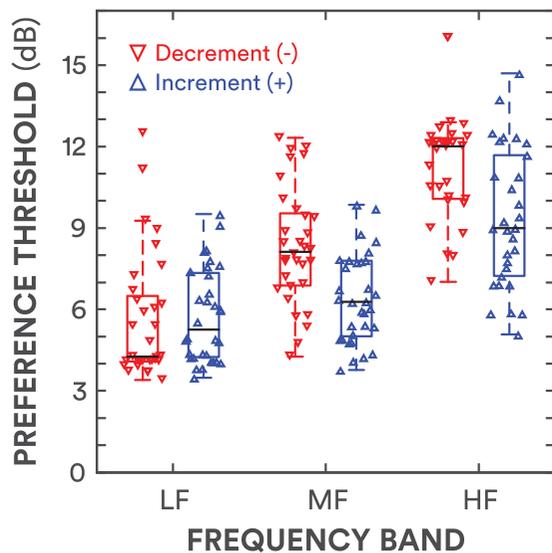
## Results

### Preferences

The line plots in the top panel of Figure 3 show preference data for every judgment made across all participants. The line plots show the “better”, “worse” and “no different” judgments for adjustments to the three frequencies. The 12 dB LF increment was the adjustment with the most “better” judgments, forming 59% of all judgments for that adjustment. The 12 dB LF decrement was the adjustment with the most “worse” judgments, forming 90% of all judgments for that adjustment. HF adjustments resulted in the most “no different” judgments; 56% of judgments for the 12 dB HF decrement were “no different”. Conversely, LF adjustments resulted in the fewest “no different” judgments; 37% of judgments for  $\pm 4$  dB LF adjustments were judged either “better” or “worse”.

### Preference agreement and reliability

The line plots in the bottom panel of Figure 3 show the percentage of participants with a reliable response for each adjustment. The 12 dB LF increment was the adjustment most reliably



**Figure 4.** Box plots of preference thresholds. The red box plots show the decrement thresholds, while the blue box plots show the increment thresholds. The black lines refer to the median thresholds. Whiskers extend to the most extreme thresholds that are within  $1.5 \times$  the interquartile range. The upward- and downward-pointing triangles show the individual thresholds for increments and decrements, respectively.

preferred to the standard by participants; 50% of participants reliably judged it to be “better” than the standard. However, 25% of participants also reliably judged it to be “worse” than their standard, demonstrating individual variation. The 12 dB LF was the most agreed-upon adjustment (in terms of a “better” or “worse” judgment); 94% of participants reliably judged it to be “worse” than their standard. Excluding control adjustments (which were reliably judged to be “no different” by all participants), the 4 dB HF decrement was judged to be most similar to the standard; all participants reliably judged it to be “no different” to the standard. The least reliable judgments were made for 8 dB MF decrements; 72% of participants made unreliable judgments for this adjustment.

Fleiss’  $\kappa$  for between-participant preference agreement was 0.25, suggesting some agreement. On average, participants’ own preference judgments were reliable for 65% ( $SD = 13\%$ ) of adjustments. Within-participant preference reliability was not correlated with age, BE4FA or hearing-aid experience ( $r = 0.06, 0.19$  and  $0.02$ , respectively;  $p > 0.05$  for all).

### Preference thresholds

The box plots in Figure 4 show preference thresholds for increments and decrements in gain. The increment thresholds and interquartile ranges (IQRs) for the LF, MF and HF bands were 5.3 [4.2–7.3] dB, 6.3 [5.0–7.8] dB and 9.0 [7.2–11.7] dB, respectively. The decrement thresholds and IQRs for the LF, MF and HF bands were 4.3 [4.1–6.5] dB, 8.1 [6.9–9.5] dB and 12.0 [10.1–12.3] dB, respectively.

### Role of centre frequency and direction

Wilcoxon signed-rank tests revealed effects of centre frequency. While LF increment thresholds were not significantly different from MF increment thresholds ( $Z = -1.46$ ;  $p > 0.05$ ), both LF and MF increment thresholds were smaller than HF increment thresholds ( $Z = -4.49$  and  $-4.33$ , respectively; both  $p < 0.001$ ). LF decrement thresholds were smaller than the MF and HF

decrement thresholds ( $Z = -4.20$  and  $-4.80$ , respectively;  $p < 0.01$  and  $0.001$ , respectively), and MF decrement thresholds were smaller than HF decrement thresholds ( $Z = -4.23$ ;  $p < 0.05$ ). There were also effects of gain-adjustment direction on preference thresholds: MF and HF increment thresholds were smaller than corresponding MF and HF decrement thresholds ( $Z = -4.62$  and  $-3.69$  respectively;  $p < 0.01$  for both). However, there was no significant difference between LF increment and decrement thresholds ( $Z = 0.74$ ;  $p \gg 0.05$ ). Increment and decrement thresholds were positively correlated ( $r = 0.61, 0.66$  and  $0.41$  for LF, MF and HF bands, respectively;  $p < 0.001, 0.001$  and  $0.05$ , respectively).

### Role of audibility

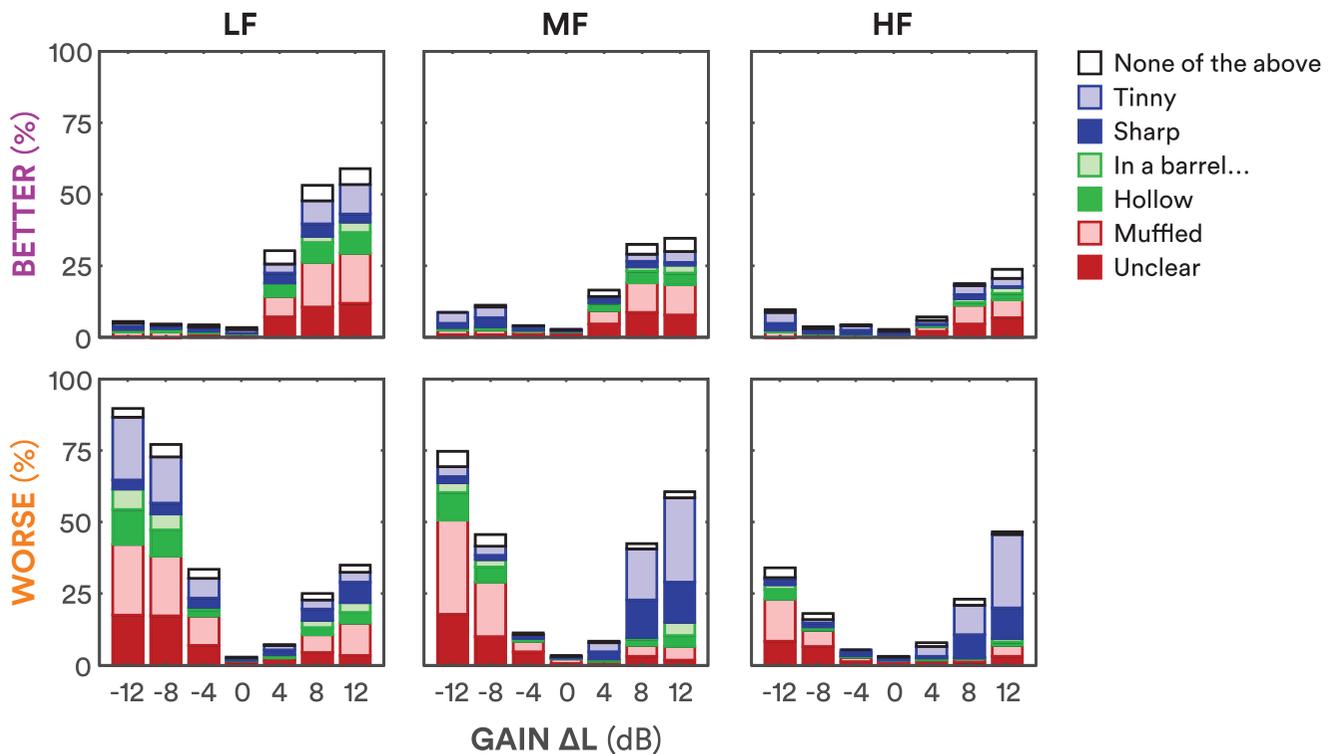
Correlations between participants’ preference thresholds and their pure-tone thresholds at corresponding frequencies were tested. LF (0.25 and 0.5 kHz) pure-tone average thresholds were not correlated with corresponding LF increment thresholds ( $r = 0.03$ ;  $p > 0.05$ ). MF (1 and 2 kHz) and HF (4 and 6 kHz) pure-tone average thresholds were positively correlated with corresponding MF and HF increment thresholds ( $r = 0.36$  and  $0.41$  for MF and HF, respectively;  $p < 0.05$  for both). LF and MF pure-tone average thresholds were not correlated with corresponding LF and MF decrement thresholds ( $r = 0.21$  and  $0.24$  for LF and MF, respectively;  $p > 0.05$  for both). HF pure-tone average thresholds were positively correlated with HF decrement thresholds ( $r = 0.52$ ;  $p < 0.01$ ). Sensation level was approximated using pure-tone thresholds and presentation level to examine the role of audibility in these correlations. Participants’ MF and HF average sensation levels were negatively correlated with their MF and HF increment thresholds ( $r = -0.43$  and  $-0.46$  for MF and HF, respectively;  $p < 0.05$  for both). HF average sensation levels were also correlated with HF decrement thresholds ( $r = -0.42$ , respectively;  $p < 0.05$ ).

### Relationship to gain JNDs

Comparisons between participants’ increment preference thresholds and their previously measured increment JNDs (Caswell-Midwinter and Whitmer 2019b) were made. Whereas the median LF, MF and HF gain increment JNDs were 3.7, 3.8 and 6.8 dB, respectively, the median gain preference thresholds were 5.3, 6.3 and 9.0 dB, respectively: preference thresholds were approximately 2 dB greater than JNDs of the same frequency band ( $Z = 4.25, 4.38$  and  $3.41$  for LF, MF and HF bands, respectively;  $p < 0.001, 0.001$  and  $0.01$ , respectively). There were no significant correlations between participants’ current LF, MF and HF preference thresholds and corresponding JNDs measured in Caswell-Midwinter and Whitmer (2019b;  $r = 0.18, 0.16$  and  $0.32$ , respectively; all  $p > 0.05$ ).

### Descriptors

The top panel of stacked bar plots in Figure 5 shows the descriptors assigned to adjustments judged to be “better” than the standard (or the standard was judged to be “worse” than the adjustment). There was large variation in descriptor selection, with many being assigned across frequency bands, levels and directions. “Less unclear” and “less muffled” were the most selected “better” descriptors, being assigned to 21% and 27% of all “better” judgments, respectively. These descriptors were mostly assigned to gain increments, irrespective of frequency:



**Figure 5.** Total descriptor judgments across participants. The top panel shows the descriptors assigned across participants when an adjustment was judged to be “better” than the standard, or the standard was judged to be “worse” than an adjustment (e.g. “better: the adjustment was less muffled than the standard”). The heights of the bars correspond to the purple “better” lines in the top panel of Figure 3. The bottom panel shows the descriptors assigned when an adjustment was judged to be “worse” than the standard, or when the standard was judged to be “better” than an adjustment (e.g. “worse: the adjustment was more sharp than the standard”). The heights of the bars correspond to the orange “worse” lines in the top panel of Figure 3. The bars are stacked in the following order (from bottom to top, alternating dark and light shading): “unclear” and “muffled” in red; “hollow” and “in a barrel, tunnel, well” in green; “sharp” and “tinny” in blue; “none of the above” in white.

“less unclear” and “less muffled” accounted for approximately 52% of LF, MF and HF increment “better” judgments. “Less hollow” and “less in a barrel, tunnel, well” were assigned to 11% and 5% of “better” judgments. These descriptors were primarily assigned to LF increments, although not definitively, accounting for 17% of LF increment “better” judgments. “Less sharp” and “less tinny” were assigned to 9% and 16% of “better” judgments, respectively. These descriptors were mostly assigned to LF increments, accounting for 22% of LF increment “better” judgments. “None of the above” was assigned to 11% of “better” judgments.

The bottom panel of stacked bar plots in Figure 5 shows the descriptors assigned to adjustments judged to be “worse” than the standard (or the standard was judged to be “better” than the adjustment). “More unclear” and “more muffled” were the most selected “worse” descriptors, being assigned to 17% and 26% of all “worse” judgments, respectively. These descriptors were assigned to adjustments universally, across increments and decrements of all frequencies, although mostly for decrements: “more unclear” and “more muffled” accounted for approximately 49% of LF decrement “worse” judgments, 68% of MF decrement “worse” judgments, and 63% of HF decrement “worse” judgments. “More hollow” and “more in a barrel, tunnel, well” were also assigned unclearly, to 9% and 6% of “worse” judgments, respectively. These descriptors were primarily assigned to LF and MF decrements, although they were also assigned to LF and MF increments. “More hollow” and “more in a barrel, tunnel, well” accounted for approximately 15% and 5% of LF and MF decrement “worse” judgments, respectively. “More sharp” and “more tinny” were assigned to 12% and 24% of “worse” judgments. These descriptors were mostly assigned to positive sloping spectral tilts: LF decrements, and MF and HF increments. “More

sharp” and “more tinny” accounted for approximately 28% of LF decrement “worse” judgments, and 70% of MF and HF increment “worse” judgments. “None of the above” was assigned to 6% of “worse” judgments.

### Descriptor agreement and reliability

Fleiss’  $\kappa$  for between-participant descriptor agreement was 0.01, suggesting little to no agreement. There was a small increase in  $\kappa$  to 0.02 when collapsing associated descriptors (e.g. “sharp” and “tinny”; cf. Jenstad, Van Tasell, and Ewert 2003), or folding across “less...” and “more...” (i.e. “better” and “worse”) descriptors. On average, participants’ own descriptor judgments were reliable for 46% ( $SD = 23\%$ ) of adjustments. Within-participant descriptor reliability was not correlated with age, BE4FA or hearing-aid experience ( $r = 0.00, 0.20$  and  $0.14$ , respectively; all  $p > 0.05$ ). Within-participant descriptor and preference reliability was not significantly correlated ( $r = 0.08$ ;  $p > 0.05$ ).

## Discussion

### Preferences

In the measurement of preference thresholds, 32 hearing-impaired listeners made preference judgments between standard and adjusted gains. The number of “better” and “worse” judgments increased with magnitude of adjustment, with the inverse true for “no different” judgments. LF adjustments elicited the most “better” or “worse” judgments, while HF adjustments elicited the most “no different” judgments. “No different”

judgments were common for 4 dB adjustments, irrespective of frequency band.

Most participants consistently judged gain decrements to be “worse” than their standards. The 12 dB LF increment was most preferred by participants. Overall preference for increased LF gain for both speech and music has previously been reported, as has individual variation (Keidser and Convery 2018; Nelson et al. 2018; Preminger et al. 2000; Vaisberg et al. 2018): several participants reliably judged LF increments to be “worse” than their standards.

There was some between-participant preference agreement: LF and MF decrements were judged to be “worse” than standards by most participants. However, beyond this, preferences varied. Within-participant preference reliability was moderate. On average, participants made reliable “better”, “worse” or “no different” responses for 65% of adjustments. This can be considered moderate in context as almost a third of adjustments were  $\pm 4$  dB, designed to be just-noticeable. This reliability is in line with previous research (Dreschler et al. 2008; Nelson et al. 2018).

### Preference thresholds

The first primary objective of the current study was to estimate thresholds for preferences, and compare these to previously measured JNDs. In Caswell-Midwinter and Whitmer (2019b), hearing-impaired participants discriminated gain increments in male, single-talker sentences. In the current study, 32 of the same participants made subjective judgments on gain adjustments made to the same stimuli, to closely compare psychophysical (same-different) and subjective (better-worse-no different) judgments. Preference thresholds for gain increments were approximately 2 dB greater than the JNDs for gain increments. This indicates that while a gain adjustment must be at least noticeable to elicit a preference, a noticeable gain adjustment will not necessarily be sufficient in magnitude to elicit a preference.

Participants’ JNDs previously measured in Caswell-Midwinter and Whitmer (2019b) and current preference thresholds did not correlate significantly. This demonstrates the increased noise within decision making on subjective attributes. The magnitude of preference thresholds compared to JNDs, and the lack of correlation between the two is in line with previous research. McShefferty, Whitmer, and Akeroyd (2016) reported that speech-to-noise ratio adjustments required to elicit a “better” or “worse” ratings were greater than JND adjustments, and that these measures did not correlate. Preference thresholds for increments were generally lower than for decrements. This effect of direction is concordant with previous discrimination thresholds (Ellermeier 1996; Moore, Oldfield, and Dooley 1989).

### Descriptors

The second primary objective of the current study was to measure the assignment of descriptors to gain adjustments, and measure whether there was agreement between participants. There was little evidence of agreement here. Furthermore, participants’ own descriptor judgments were markedly less reliable than their preference judgments. Several descriptors were assigned with no clear pattern across different frequencies and adjustment directions. While not reported in detail, Daugaard, Jørgensen, and Elmelund (2011) described large variation in naturalness ratings from hearing-aid users, suggesting that even fundamental descriptors which may be used as references can be inconsistently rated between listeners.

A secondary objective was to compare the descriptors assigned here to those suggested by clinicians in Jenstad, Van Tasell, and Ewert (2003). “Unclear” and “muffled” were reported by those clinicians to be primarily associated with insufficient high-frequency gain, with the primary solution being to increase high-frequency gain. In the current study, the most assigned descriptor for HF decrements was “more muffled”, followed by “more unclear”, and the most used “better” descriptor for HF increments was “less muffled”, followed by “less unclear”, suggesting agreement. However, these descriptors were used across adjustments; “more unclear” and “more muffled” were also the most used descriptors for LF and MF decrements.

“Hollow” and “in a barrel, tunnel, well” were reported in the expert system to be associated with excessive low-frequency gain, with the electroacoustical solution being to decrease low-frequency gain. While there was some use of “more hollow” and “more in a barrel, tunnel, well” for LF increments, these descriptors were mostly used for LF decrements, disagreeing with the expert system on adjustment direction. However, these interpretations are limited given that these descriptors were the least used, and that use was highly variable across adjustments. “In a barrel, tunnel, well” descriptors were the least used in this study. A third of those responses were from a single participant whose age (74) and average air-bone gap (6 dB HL) were outside the interquartile range but not outliers. “In a barrel, tunnel, well” was not a descriptor reported by Dutch clinicians (Thielemans et al. 2017), indicating the potential variations in descriptors between different regions.

“Sharp” and “tinny” were reported in the expert system to be associated with excessive high-frequency gain, with the primary solution being to decrease high-frequency gain, and the third solution to increase low-frequency gain. In this study, “more sharp” and “more tinny” were used similarly and largely based on spectral tilts; “less sharp” and “less tinny” were assigned to LF increments, while “more sharp” and “more tinny” were assigned to MF and HF increments. These results are concordant with the expert system.

Presenting female, single-talker sentences, Sabin et al. (2011) had 10 hearing-impaired participants rate descriptors to adjustments from NAL-R standards. As in the current study, “sharp” and “tinny” were mapped to spectral tilts. However, “hollow” and “in a barrel, tunnel, well” were generally mapped to negative spectral tilts, while these descriptors were mostly assigned to LF decrements in the current study. As in the current study, Sabin et al. reported some between-participant disagreement and within-participant variation in descriptor mapping. However, this variation was not as substantial as that in the current study, maybe due to methodological differences. In Sabin et al.’s study, participants could rate multiple descriptors to each gain adjustment. In the current study, participants were limited to assigning a single descriptor, which may have exaggerated individual differences. Furthermore, stimuli presented in Sabin et al. were repeatable, which would facilitate reliability.

### Study limitations

Stimuli in the current study were presented monaurally over headphones. While gain adjustments in the clinic may be done bilaterally with linked devices, appropriate binaural gains may differ dramatically from monaural gains across individuals (cf. Oetting et al. 2018). Hence, using monaural REIG or gain prescription as a standard facilitates interpretation of preferences and descriptors at a group level, as well as comparison to the discrimination thresholds for the same participants from our previous study. Additionally, the current study did not account

for the physical properties of hearing-aids, such as ear moulds and venting. These properties, which can be adjusted by clinicians, interact with the electroacoustics of real-world devices and influence sound quality. The troubleshooting solutions presented in Jenstad, Van Tasell, and Ewert (2003) also included adjustments to such physical properties. For example, the primary solution for “in a barrel, tunnel, well” is to increase vent size. “In a barrel, tunnel, well” was seldom assigned in this study. It may be that descriptors become more meaningful when physical properties are adjusted. However, Sabin et al. (2011), who also only adjusted gain, reported that this descriptor was generally mapped to a negative spectral tilt.

The A-weighted presentation level for stimuli was 60 dB without gain in this study, approximately standard in-quiet conversation level (Olsen 1998). Also presenting short sentences, Moore, Alcantara, and Glasberg (1998) used descriptor continuums to adaptively fine-tune gain, with “boomy-to-tinny” rated for sentences presented at 85 dB SPL, and “muffled-to-shrill” rated for sentences presented at 60 dB SPL. “Muffled” was an option in the current study for which there was no clearly associated adjustment. While the perceptual pilot study on which the Moore et al. anchors were based was not detailed, it may be that some descriptors are more relevant to speech at louder or softer levels.

The current study used a trial-by-trial psychophysical approach for measuring preference and descriptor judgments, as has been done in gain (Caswell-Midwinter and Whitmer 2019a, 2019b), compression (Gilbert et al. 2008; Nabelek 1984; Sabin et al. 2013) and speech-to-noise ratio (McShefferty, Whitmer, and Akeroyd 2015, 2016) discrimination studies. The lack of descriptor reliability and large preference thresholds may be related to the short duration stimuli. Although patients typically make quick comparisons on adjustments in the clinic, it may be that “no different” responses decrease and descriptor reliability increases with stimulus duration. A previous study with hearing-aid users has shown moderate within-participant reliability in ratings of 50- and 60-s passages of speech and music (Narendran and Humes 2003). However, those results cannot be generalised to comparisons of specific parameter adjustments. While psychophysical research with brief, basic stimuli has shown that thresholds decrease with stimulus duration (Dai and Green 1993; Florentine, Fastl, and Buus 1988; Shrivastav, Humes, and Kewley-Port 2006), it is unclear how duration affects subjective judgments made on real-world stimuli. Long-term exposure to various – or perhaps particular – sounds in one’s environment may also lead to greater descriptor reliability.

The preference query here came without further instruction; the participant was not informed of a basis upon which to make their preferences, such as comfort or intelligibility. This was done to estimate the smallest adjustment that would elicit a preference, regardless of the basis for that preference. It is not therefore possible to further interpret preferences as to the basis for such judgments (e.g. the reason(s) for the increased low-frequency gain preference). Differences in the underlying criteria for these preferences, within and across participants, may be a factor in the lack of high preference agreement or reliability. Approximately 8% of all descriptor judgments were “none of the above,” indicating that the closed set of descriptors were either not appropriate or meaningful to some participants for some adjustments. It may be that these descriptors were not relevant to the dimension of certain preferences. An open-set descriptors procedure may allow greater insight into the underlying criteria of preferences. However, it could be expected that agreement would be poorer in an

unrestricted open-set procedure which facilitates further individual differences (cf. Daugaard, Jørgensen, and Elmelund 2011).

### **Clinical implications and future directions**

The preference thresholds and discrimination thresholds for gain (Caswell-Midwinter and Whitmer 2019b), and the previous discrimination thresholds for compression (Gilbert et al. 2008; Nabelek 1984) and speech-to-noise ratio (McShefferty, Whitmer, and Akeroyd 2015, 2016) suggest that hearing-impaired listeners are not very sensitive to small electroacoustical adjustments made in a hearing-aid, at least when they are only experienced for a short time. Previous studies have also reported that patient feedback is susceptible to placebo effects when acoustically identical devices are compared, even when they are worn for substantial periods of daily use (Bentler et al. 2003; Dawes, Hopkins, and Munro 2013; Dawes, Powell, and Munro 2011; Naylor et al. 2015). This insensitivity to adjustments and susceptibility to placebo effects shows the limitations of adjusting acoustical parameters in response to patient feedback. More research is required to develop troubleshooting protocols which consider and overcome these issues. The preference thresholds here, and the previous JNDs (Caswell-Midwinter and Whitmer 2019b), suggest that with broad frequency bands the discriminability and preferences of spectral-tilt adjustments in self-fittings (e.g. Boothroyd and Mackersie 2017; Sabin et al. 2020) are based on the lower-frequency spectral tilt.

The current lack of between-participant descriptor agreement questions the validity of listener-independent rules for translating descriptors into gain adjustments. Given the individual variation in descriptor meaning, troubleshooting in the clinic with descriptors may fail (at least acoustically) on the clinician’s interpretation of the patient’s descriptor, or on the adjustment solution. Additionally, evidence here suggests that listeners’ own descriptor judgments are not necessarily reliable either. These uncertainties may be exacerbated in clinics where patients typically only make one, or at the most, very few adjustment comparisons. Clinicians should exhibit caution when using descriptors to adjust gain, particularly when using short sentences as the test stimuli. If adjustments are made to patient feedback, clinicians should consider fine-tuning in two ways. First, adjustments should first be large enough in magnitude to be noticeable, so patients can make an informed comparison; this magnitude will be dependent on the centre frequency and bandwidth of the adjustment, as well as the stimuli used in the comparison. Second, patients should be prompted for more than a single descriptor; clinicians should probe each patient’s own internal preference system independent of external anchors before making any discriminable adjustments.

Fine-tuning and troubleshooting should provide benefits in the patient’s real-world listening conditions. But adjustments in practice are verified to meet patient satisfaction in the clinic, which may poorly represent the conditions in which patients are having difficulties. The lack of ecologically valid test conditions (e.g. a quiet clinic, adjusting to the live voice of the clinician) should be considered in regards to the patient’s auditory ecology, and it may be useful to counsel the patient themselves on this disparity. Further counselling on expectations and acclimatisation will also be crucial, particularly when the patient has not had real-world experience with their device(s). Digital tools, such as smartphone-based ecological momentary assessment (Galvez et al. 2012) and photo sharing (Saunders 2019) could better inform a real-world basis to troubleshooting than recall or single,

unreliable descriptors. Such tools could help the clinician determine the most appropriate form of management, whether that is counselling and/or making adjustments.

Self-adjustments, found to be quick and/or reliable (Nelson et al. 2018; Mackersie, Boothroyd, and Lithgow 2019), may facilitate troubleshooting. This method can alleviate the need for the patient to describe their percept with language, and the need for the clinician to interpret this language into adjustments. It can also allow the patient to troubleshoot problems in-situ as they arise, alleviating the need for them to memorise and recall problems weeks to months later in the clinic, which can be inaccurate and unreliable (Bradburn, Rips, and Shevell 1987). Several studies have shown that, on average, self-fitting methods are as beneficial as clinical fittings for listeners with mild-to-moderate hearing loss (Humes et al. 2017; Nelson et al. 2018; Mackersie, Boothroyd, and Lithgow 2019; Sabin et al. 2020). However, individual patients may adjust their devices in ways that compromise speech recognition (Boymans and Dreschler 2012; Nelson et al. 2018). Further research is needed to bolster the efficacy and efficiency of self-guided alternatives to descriptor troubleshooting.

## Conclusions

The preference thresholds measured here were greater than previously measured discrimination thresholds, indicating that just-noticeable gain adjustments are not necessarily sufficient in magnitude to be subjectively meaningful. The effect of centre frequency on thresholds suggest that gain preferences for spectral tilts are likely to be based on lower frequencies. In addition to this, the magnitude of the preference thresholds demonstrate the inefficiency of using short sentences as the stimulus for adjusting gain in response to patient feedback. There was little evidence of agreement between participants in the assignment of descriptors to gains. This suggest that translating descriptors into adjustments in a consistent manner across listeners is not wholly valid. In addition to the lack of between-participant agreement, participants' own descriptor judgments were much less reliable than their preferences. Overall, these findings demonstrate limitations of troubleshooting electroacoustical hearing-aid parameters to patients' descriptions. Further research is required to develop and evaluate strategies for informing and implementing beneficial fine-tuning and troubleshooting.

## Acknowledgements

The authors thank Prof. Brian Moore and Dr. Rachel Smith for their feedback on previous versions of this work, and Prof. Graham Naylor for helpful comments on this manuscript. The authors also thank Dr. Gitte Keidser and the three anonymous reviewers for their helpful comments.

## Disclosure statement

No potential conflict of interest was reported by the author(s).

## Funding

This work was supported by funding from the Medical Research Council [grant numbers 1601056 and MR/S003576/1]; and the Chief Scientist Office of the Scottish Government.

## ORCID

Benjamin Caswell-Midwinter  <http://orcid.org/0000-0002-3386-3860>

William M. Whitmer  <http://orcid.org/0000-0001-8618-6851>

## References

- Anderson, M. C., K. H. Arehart, and P. E. Souza. 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 (2): 118–124. doi:10.3766/jaaa.16107.
- Bench, J., A. Kowal, and J. Bamford. 1979. "The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children." *British Journal of Audiology* 13 (3): 108–112. doi:10.3109/03005367909078884.
- Bentler, R. A., D. P. Neibuhr, T. A. Johnson, and G. A. Flamme. 2003. "Impact of Digital Labelling on Outcome Measures." *Ear and Hearing* 24: 215–224. doi:10.1097/01.AUD.0000069228.46916.92.
- Boothroyd, A., and C. Mackersie. 2017. "A "Goldilocks" Approach to Hearing-Aid Self-Fitting: User Interactions." *American Journal of Audiology* 26 (3S): 430–435. doi:10.1044/2017\_AJA-16-0125.
- Boymans, M., and W. A. Dreschler. 2012. "Audiologist-Driven versus Patient-Driven Fine Tuning of Hearing Instruments." *Trends in Amplification* 16 (1): 49–58. doi:10.1177/1084713811424884.
- Bradburn, N. M., L. J. Rips, and S. K. Shevell. 1987. "Answering Autobiographical Questions: The Impact of Memory and Inference on Surveys." *Science (New York, N.Y.)* 236 (4798): 157–161. doi:10.1126/science.3563494.
- British Academy of Audiology. 2016. *Guidance for Audiologists: Onward Referral of Adults with Hearing Difficulty Directly Referred to Audiology Services*. [https://www.baaudiology.org/app/uploads/2019/07/BAA\\_Guidance\\_for\\_Onward\\_Referral\\_of\\_Adults\\_with\\_Hearing\\_Difficulty\\_Directly\\_Referred\\_to\\_Audiology\\_2016\\_-\\_minor\\_amendments.pdf](https://www.baaudiology.org/app/uploads/2019/07/BAA_Guidance_for_Onward_Referral_of_Adults_with_Hearing_Difficulty_Directly_Referred_to_Audiology_2016_-_minor_amendments.pdf)
- Byrne, D., and H. Dillon. 1986. "The National Acoustic Laboratories' (NAL) New Procedure for Selecting the Gain and Frequency Response of a Hearing Aid." *Ear and Hearing* 7 (4): 257–265. doi:10.1097/00003446-198608000-00007.
- Caswell-Midwinter, B., and W. M. Whitmer. 2019a. "Discrimination of Gain Increments in Speech-Shaped Noises." *Trends in Hearing* 23: 2331216518820220. doi:10.1177/2331216518820220.
- Caswell-Midwinter, B., and W. M. Whitmer. 2019b. "Discrimination of Gain Increments in Speech." *Trends in Hearing* 23: 2331216519886684. doi:10.1177/2331216519886684.
- Cunningham, D. R., K. J. Williams, and L. J. Goldsmith. 2001. "Effects of Providing and Withholding Postfitting Fine-Tuning Adjustments on Outcome Measures in Novice Hearing Aid Users: A Pilot Study." *American Journal of Audiology* 10 (1): 13–23. doi:10.1044/1059-0889(2001/001).
- Curran, J. R., and J. A. Galster. 2013. "The Master Hearing Aid." *Trends in Amplification* 17 (2): 108–134. doi:10.1177/1084713813486851.
- Dai, H., and D. M. Green. 1993. "Discrimination of Spectral Shape as a Function of Stimulus Duration." *The Journal of the Acoustical Society of America* 93 (2): 957–963. doi:10.1121/1.405456.
- Daugaard, C., S. L. Jørgensen, and L. Elmelund. 2011. "Benefits of Common Vocabulary in Hearing Aid Fitting." In *Speech Perception and Auditory Disorders*, edited by T. Dau, M. L. Jepsen, T. Poulsen, and J. C. Dalsgaard, 432–440. Ballerup, Denmark: Danavox Jubilee Fndn.
- Dawes, P., R. Hopkins, and K. J. Munro. 2013. "Placebo Effects in Hearing-Aid Trials Are Reliable." *International Journal of Audiology* 52 (7): 472–477. doi:10.3109/14992027.2013.783718.
- Dawes, P., S. Powell, and K. J. Munro. 2011. "The Placebo Effect and the Influence of Participant Expectation on Hearing Aid Trials." *Ear and Hearing* 32 (6): 767–774. doi:10.1097/aud.0b013e3182251a0e.
- Dirks, D. D., J. Ahlstrom, and P. D. Noffsinger. 1993. "Preferred Frequency Response for Two- and Three-Channel Amplification Systems." *Journal of Rehabilitation Research and Development* 30 (3): 305–317.
- Dreschler, W. A., G. Keidser, E. Convery, and H. Dillon. 2008. "Client-Based Adjustments of Hearing Aid Gain: The Effect of Different Control Configurations." *Ear and Hearing* 29 (2): 214–227. doi:10.1097/AUD.0b013e31816453a6.
- Ellermeier, W. 1996. "Detectability of Increments and Decrements in Spectral Profiles." *The Journal of the Acoustical Society of America* 99 (5): 3119–3125. doi:10.1121/1.414797.
- Fleiss, J. 1971. "Measuring Nominal Scale Agreement among Many Raters." *Psychological Bulletin* 76 (5): 378–382. doi:10.1037/h0031619.

- Florentine, M., H. Fastl, and S. R. Buus. 1988. "Temporal Integration in Normal Hearing, Cochlear Impairment, and Impairment Simulated by Masking." *The Journal of the Acoustical Society of America* 84 (1): 195–203. doi:10.1121/1.396964.
- Gabrielsson, A., B. Hagerman, T. Bech-Kristensen, and G. Lundberg. 1990. "Perceived Sound Quality of Reproductions with Different Frequency Responses and Sound Levels." *The Journal of the Acoustical Society of America* 88 (3): 1359–1366. doi:10.1121/1.399713.
- Gabrielsson, A., and H. Sjögren. 1979. "Perceived Sound Quality of Hearing Aids." *Scandinavian Audiology* 8 (3): 159–169. doi:10.3109/01050397909076317.
- Galvez, G., M. B. Turbin, E. J. Thielman, J. A. Istvan, J. A. Andrews, and J. A. Henry. 2012. "Feasibility of Ecological Momentary Assessment of Hearing Difficulties Encountered by Hearing Aid Users." *Ear and Hearing* 33 (4): 497–507. doi:10.1097/AUD.0b013e3182498c41.
- Gilbert, G., M. A. Akeroyd, and S. Gatehouse. 2008. "Discrimination of Release Time Constants in Hearing-Aid Compressors." *International Journal of Audiology* 47 (4): 189–198. doi:10.1080/14992020701829722.
- Hayes, A. F., and K. Krippendorff. 2007. "Answering the Call for a Standard Reliability Measure for Coding Data." *Communication Methods and Measures* 1 (1): 77–89. doi:10.1080/19312450709336664.
- Humes, L., S. Rogers, T. Quigley, A. Main, D. Kinney, and C. Herring. 2017. "The Effects of Service-Delivery Model and Purchase Price on Hearing-Aid Outcomes in Older Adults: A Randomized Double-Blind Placebo-Controlled Clinical Trial." *American Journal of Audiology* 26 (1): 53–79. doi:10.1044/2017\_AJA-16-0111.
- Jenstad, L. M., D. J. Van Tasell, and C. Ewert. 2003. "Hearing Aid Troubleshooting Based on Patients' Descriptions." *Journal of the American Academy of Audiology* 14 (7): 347–360.
- Jenstad, L. M., M. P. Bagatto, R. C. Seewald, S. D. Scollie, L. E. Cornelisse, and R. Scicluna. 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. doi:10.1097/AUD.0b013e318157670a.
- Kates, J. M., and K. H. Arehart. 2010. "The Hearing-Aid Speech Quality Index (HASQI)." *Journal of the Audio Engineering Society* 58: 363–381.
- Keidser, G., and E. Convery. 2016. "Self-Fitting Hearing Aids: Status Quo and Future Predictions." *Trends in Hearing* 20: 233121651664328. doi:10.1177/2331216516643284.
- Keidser, G., and E. Convery. 2018. "Outcomes with a Self-Fitting Hearing Aid." *Trends in Hearing* 22 (11): 233121651876895–12. doi:10.1177/2331216518768958.
- Keidser, G., H. Dillon, and E. Convery. 2008. "The Effect of the Base Line Response on Self-Adjustments of Hearing Aid Gain." *The Journal of the Acoustical Society of America* 124 (3): 1668–1681. doi:10.1121/1.2951500.
- Kuk, F. K., and C. Lau. 1995. "The Application of Binomial Probability Theory to Paired Comparison Responses." *American Journal of Audiology* 4 (1): 37–42. doi:10.1044/1059-0889.0401.37.
- Kuk, F. K., and C. Ludvigsen. 1999. "Variables Affecting the Use of Prescriptive Formulae to Fit Modern Nonlinear Hearing Aids." *Journal of the American Academy of Audiology* 10: 453–465.
- Mackersie, C. L., A. Boothroyd, and A. Lithgow. 2019. "A 'Goldilocks' Approach to Hearing Aid Self-Fitting: Ear-Canal Output and Speech Intelligibility Index." *Ear and Hearing* 40 (1): 107–115. doi:10.1097/AUD.0000000000000617.
- Macmillan, N. A., and C. D. Creelman. 2005. *Detection Theory A User's Guide*. Mahwah, NJ: Lawrence Erlbaum Associates.
- McShefferty, D., W. M. Whitmer, and M. A. Akeroyd. 2015. "The Just-Noticeable Difference in Speech-to-Noise Ratio." *Trends in Hearing* 19: 233121651557231. doi:10.1177/2331216515572316.
- McShefferty, D., W. M. Whitmer, and M. A. Akeroyd. 2016. "The Just-Meaningful Difference in Speech-to-Noise Ratio." *Trends in Hearing* 20: 233121651562657. doi:10.1177/2331216515626570.
- Moore, B. C., J. I. Alcantara, and B. R. Glasberg. 1998. "Development and Evaluation of a Procedure for Fitting Multi-Channel Compression Hearing Aids." *British Journal of Audiology* 32 (3): 177–195. doi:10.3109/03005364000000062.
- Moore, B. C., S. R. Oldfield, and G. J. Dooley. 1989. "Detection and Discrimination of Spectral Peaks and Notches at 1 and 8 kHz." *The Journal of the Acoustical Society of America* 85 (2): 820–836. doi:10.1121/1.397554.
- Nabelek, I. V. 1984. "Discriminability of the Quality of Amplitude-Compressed Speech." *Journal of Speech and Hearing Research* 27: 571–577. doi:10.1044/jsr.2704.571.
- Narendran, M. M., and L. E. Humes. 2003. "Reliability and Validity of Judgments of Sound Quality in Elderly Hearing Aid Wearers." *Ear and Hearing* 24 (1): 4–11. doi:10.1097/01.AUD.0000051745.69182.14.
- Naylor, G., M. Öberg, G. Wänström, and T. Lunner. 2015. "Exploring the Effects of the Narrative Embodied in the Hearing Aid Fitting Process on Treatment Outcomes." *Ear and Hearing* 36 (5): 517–526. doi:10.1097/AUD.0000000000000157.
- Nelson, P. B., T. T. Perry, M. Gregan, and D. VanTasell. 2018. "Self-Adjusted Amplification Parameters Produce Large between-Subject Variability and Preserve Speech Intelligibility." *Trends in Hearing* 22: 2331216518798264. doi:10.1177/2331216518798264.
- Oetting, D., V. Hohmann, J. E. Appell, B. Kollmeier, and S. D. Ewert. 2018. "Restoring Perceived Loudness for Listeners with Hearing Loss." *Ear and Hearing* 39 (4): 664–678. doi:10.1097/AUD.0000000000000521.
- Olsen, W. A. 1998. "Average Speech Levels and Spectra in Various Speaking/Listening Conditions: A Summary of the Pearson, Bennett, and Fidell (1977) Report." *American Journal of Audiology* 7 (2): 21–25. doi:10.1044/1059-0889(1998)012.
- Preminger, J. E., A. C. Neuman, M. H. Bakke, D. Walters, and H. Levitt. 2000. "An Examination of the Practicality of the Simplex Procedure." *Ear and Hearing* 21 (3): 177–193. doi:10.1097/00003446-200006000-00001.
- Punch, J. L., B. Rakerd, and A. M. Amlani. 2001. "Paired-Comparison Hearing Aid Preferences: Evaluation of an Unforced-Choice Paradigm." *Journal of the American Academy of Audiology* 12 (4): 190–201.
- Sabin, A. T., D. J. Van Tasell, B. Rabinowitz, and S. Dhar. 2020. "Validation of a Self-Fitting Method for over-the-Counter Hearing Aids." *Trends in Hearing* 24: 2331216519900589. doi:10.1177/2331216519900589.
- Sabin, A. T., F. J. Gallun, and P. E. Souza. 2013. "Acoustical Correlates of Performance on a Dynamic Range Compression Discrimination Task." *The Journal of the Acoustical Society of America* 134 (3): 2136–2147. doi:10.1121/1.4816410.
- Sabin, A. T., L. Hardies, N. Marrone, et al. 2011. "Weighting Function-Based Mapping of Descriptors to Frequency-Gain Curves in Listeners with Hearing Loss." *Ear and Hearing* 32: 399–409. doi:10.1097/AUD.0b013e318202b7ca.
- Saunders, G. H. 2019. "Photo-Sharing as an Audiological Rehabilitation Tool." *The Hearing Journal* 72 (9): 16–17. doi:10.1097/01.HJ.0000582436.09398.32.
- Saunders, G. H., M. S. Lewis, and A. Forsline. 2009. "Expectations, Prefitting Counseling, and Hearing Aid Outcome." *Journal of the American Academy of Audiology* 20 (5): 320–334. doi:10.3766/jaaa.20.5.6.
- Shrivastav, M. N., L. E. Humes, and D. Kewley-Port. 2006. "Individual Differences in Auditory Discrimination of Spectral Shape and Speech-Identification Performance among Elderly Listeners." *The Journal of the Acoustical Society of America* 119 (2): 1131–1142. doi:10.1121/1.2151794.
- Thielemans, T., D. Pans, M. Chenault, and L. Anteunis. 2017. "Hearing Aid Fine-Tuning Based on Dutch Descriptions." *International Journal of Audiology* 56 (7): 507–515. doi:10.1080/14992027.2017.1288302.
- Vaisberg, J., S. Beaulac, D. Glista, M. Van Eeckhoutte, E. Macpherson, and S. Scollie. 2018. "Preferred Hearing Aid Gain Settings for Music-Listening Using a 3D Modified Simplex Procedure Implemented with the Open Source Master Hearing Aid Platform." Oral presentation at the International Hearing Aid Research Conference, Lake Tahoe, CA.
- Whitmer, W. M., and M. A. Akeroyd. 2011. "Level Discrimination of Speech Sounds by Hearing-Impaired Individuals with and without Hearing Amplification." *Ear and Hearing* 32 (3): 391–398. doi:10.1097/AUD.0b013e318202b620.