2023

Modified Multiple Stimulus With Hidden Reference and Anchors–Gabrielsson Total Impression Sound Quality Rating Comparisons for Speech in Quiet, Noise, and Reverberation

Mohamed Rahme

Paula Folkeard

Steve Beaulac

Susan Scollie

Vijay Parsa

Modified MUSHRA-Gabrielsson Total Impression Sound Quality Rating Comparisons for Speech in Quiet, Noise, and Reverberation.

Mohamed Rahme[1,2], Paula Folkeard[1], Steve Beaulac[1], Susan Scollie [1,3] and Vijay Parsa[1,3]

Authors' Affiliations:

[1]National Centre for Audiology, Western University, London, Ontario, Canada

[2]Faculty of Health Sciences, Western University, London, Ontario, Canada

[3]School of Communication Sciences & Disorders, Western University, London, Ontario, Canada

Corresponding author:

Mohamed Rahme

Room 1219, Elborn College Western University London, ON, Canada N6G 1H1. Phone: 519-661-2111. Email: mrahme@uwo.ca

Acknowledgements

The authors would like to thank Muneeb Alam for his assistance with data analysis.

Funding

The Ontario Research Foundation (RE08-072) provided support for this project.

Conflicts of Interest

No conflicts of interest are reported.

Keywords: MUSHRA, Gabrielsson, sound quality modeling, hearing aids.

ABSTRACT

Purpose: The purpose of the study was to obtain, analyze, and compare subjective sound quality data for the same test stimuli using modified multi-stimulus MUSHRA (Multiple Stimulus with Hidden Reference and Anchors) based procedures (*viz.* MUSHRA with custom anchors and MUSHRA without anchor), and the single-stimulus Gabrielsson's total impression rating procedure. The secondary objective was to evaluate the test-retest reliability of the sound quality procedures

Methods: Twenty normally hearing young adults were recruited in this study. Participants completed sound quality ratings on two different hearing aid recording datasets – Dataset A contained speech recordings from four different hearing aids under a variety of noisy and processing conditions, and dataset B contained speech recordings from a single hearing aid under a combination of different noisy, reverberant, and signal processing conditions. Recordings in both datasets were rated for their quality using the total impression rating procedure. In addition, quality ratings of dataset A recordings were obtained using a MUSHRA with custom anchors, while the ratings of dataset B recordings were collected using a MUSHRA with no anchors.

Results: Statistical analyses revealed a high test-retest reliability of quality ratings for the same stimuli that were rated multiple times. In addition, high interrater reliability was observed with all three rating procedures. Further analyses indicated: (i) a high correlation between the total impression rating and the two modified MUSHRA ratings, and (ii) a similar pattern in the variability of the data obtained by the total impression rating and MUSHRA with custom anchors on dataset A, and the total impression rating and the MUSHRA without anchor on dataset B.

Conclusions: Both sound quality procedures, *viz.* the MUSHRA based procedures and the total impression rating scale, obtained similar quality ratings of varied hearing aid speech recordings with high reliability.

Introduction

Sound quality has been consistently ranked as a top priority for hearing aid users and is found to be the top correlate of consumer satisfaction and preference in both premium and entry-level hearing aids (Kochkin 2010, Saleh et al., 2021; Picou, 2020). To address the quality of sound external to the device, hearing aid manufacturers use digital signal processing strategies such as noise reduction algorithms and directional microphones to help reduce reverberation and background noise. These features improve the signal to noise ratio (SNR), increase speech intelligibility and make listening more comfortable in complex listening environments (Banerjee 2011; Cord et al. 2004). However, digital signal processing can generate internal noise, limit available bandwidth, and add distortion to the output signal (Kates & Arehart 2010). These signal degradations can impact sound quality (Arehart et al., 2007 & 2011). To date, hearing aids seem to still underperform in a few listening environments, most prominently in hearing in a classroom, following a conversation in noise, and talking on a phone (Abrams & Kihm, 2015; Picou 2020).

Subjective evaluation of sound quality can provide an insight on hearing aid performance as perceived by users for which sound quality continues to play a critical role in hearing aid adoption (Abrams & Kihm, 2015). Therefore, providing optimal sound quality through hearing aid processing is of high interest, and specific methods have been developed for assessing sound quality of hearing aids. The use of magnitude estimation procedures in the context of sound quality and hearing aid literature is well documented (Barry & Kidd 1981; Studebaker & Sherbecoe 1988). An example of such a subjective sound quality procedure is the Gabrielsson scale (Gabrielsson et al., 1988). The Gabrielsson rating scale was initially developed to assess the performance of different types of loudspeakers (Gabrielsson & Sjorgen 1979) and have been

used to assess performance of earphones (Gabrielsson et al., 1990), earmolds (Lundberg et al., 1992) and hearing aids (Gabrielsson et al., 1988; Lundberg et al., 1992; Kondo 2012; Narendran & Humes, 2003).

During the Gabrielsson rating task, participants are asked to listen to and then rate on a 10-point scale the quality of sound signal presented to them. This procedure uses dimension-specific descriptors such as: brightness, clarity, fidelity, fullness, loudness, nearness, softness, spaciousness and total impression. Narendran & Humes (2003) reported that the Gabrielsson task had moderate test-retest reliability across all dimensions for both unaided (median r value= 0.58) and aided (median r value = 0.51) listening conditions. For the purpose of this study, a single dimension "Total Impression" descriptor (taking into consideration the clarity, loudness, background noise, and quality of the recording) was used to evaluate the sound quality of the recordings. This procedure will be referred to as the "total impression rating scale" in this paper.

Although using a single rating scale is a fast, direct and subjective evaluation of sound quality, it also has limitations. These caveats may include subjective bias due to context effects, ceiling and floor effects (Studebaker & Sherbecoe 1988). An alternative procedure that is commonly used in the sound quality literature is the Multiple Stimulus with Hidden Reference and Anchors (MUSRHA, ITU- R, 2003). The MUSHRA procedure includes a reference (clean signal), two anchors (distorted signal), and the test stimuli (ITU- R, 2003). Participants begin the task with listening to the reference and then are asked to rate all the other sentences in the group against the reference, in random order. A copy of the reference is included in the randomized list and is referred to as the "hidden" reference. Ratings are provided in a text descriptor along the scale and a percentage score (out of 100%) is then calculated by the software. Ratings of the hidden reference and anchors are included to ensure that the listener uses the full rating scale,

and to help prevent ceiling and floor effects. The ratings obtained for the test stimuli may be compared to evaluate various signal processing algorithms.

The MUSHRA procedure has been previously applied in assessing the speech and music sound quality of a wide range of signal processing types including, but not limited to, frequency lowering, bandwidth, and noise reduction (Franz & Bitzer, 2010; Glista et al., 2019; Huber et al., 2018; Scollie et al., 2016; Vaisberg et al., 2021). Additionally, MUSHRA's application has been extended to evaluate the sound quality of both speech and music in cochlear implant devices (Caldwell, et al., 2017; Roy et al., 2012).

As mentioned previously, the reference signals in standard MUSHRA procedure are clean stimuli with full bandwidth that may be recorded either in quiet or in a positive SNR, while the anchors are traditionally low-pass filtered versions of the original recordings (ITU- R, 2003). However, for assessing the hearing aid sound quality, alternative poor quality or distorted anchor stimuli, such as those containing distortion (e.g., peak or center clipping), reverberation, or background noise, may be more appropriate. For example, Simonsen & Legarth (2010) utilized a recording from an older hearing aid as anchor, in benchmarking the sound quality newer generation hearing aids. Muralimanohar et al. (2013) and Scollie et al. (2016) used noisy speech recordings at low SNRs (-5 dB and -10 dB respectively). This modified MUSHRA based procedure will be referred to for the remainder of this study as MUSHRA with custom anchors.

Including references and anchors within the MUSHRA task is part of the standardized procedure (ITU-R, 2003). However, other modification of the standard MUSHRA could be omitting the references and/or anchors. This alternative MUSHRA based procedure may be chosen when sensible reference or anchors cannot be feasibly recorded, or if the underlying constructs associated with "high" or "low" sound quality are not well-understood. For example,

Salehi et al. (2018) compared the impact of signal processing features in two different hearing aids on perceived sound quality by listeners with hearing differences by using a modified MUSHRA based procedure that did not have a reference or anchors. This modified MUSHRA procedure will be referred as MUSHRA with no anchor for the remainder of this study.

Although MUSHRA has been commonly administered as a subjective sound quality procedure, it requires a computerized interface (Parsa et al., 2013), can be an effortful, relatively costly and time-consuming procedure (Völker et al., 2016). Völker and colleagues argued that the barriers that impact the duration of completing ratings on a traditional MUSHRA dataset can include the factors of age, technological commitment, and hearing status. Furthermore, the sound quality procedures administered in this study, MUSHRA with custom anchors, MUSHRA without anchors, and the total impression rating scale (Gabrielsson et al., 1988) differ in administration and set-up where one method could be advantageous over the other depending on the research question (Studebaker & Sherbecoe 1988; & Völker et al., 2016). The MUSHRA based procedures have the listener compare multiple stimuli on a trial and provide one sound quality rating per stimulus while during the Gabrielsson rating scale procedure, the listener evaluates the sound quality of a single stimulus on a trial and provide ratings along multiple dimensions of sound quality. The rationale for choosing the single descriptor "total impression rating scale" over the rest of the Gabrielsson's dimensions was to provide a single sound quality rating that is comparable to the MUSHRA based procedures. Narendran & Humes (2003) have shown previously that the single descriptor total impression rating scale was sufficient to judge the sound quality of speech processed in noise and in quiet in an unaided condition.

In summary, hearing aid sound quality varies with hearing aid settings and environments and is an important outcome measure that is used in assessing new signal processors. Two sound

quality procedures have been used in recent studies: comparisons of multiple stimuli (e.g., MUSHRA based procedures) and single-stimulus ratings (e.g., total impression rating scale). How these two common procedures relate to one another is not known, it is therefore difficult to compare ratings made with the two procedures. Therefore, the current study is devised to provide initial comparisons between the modified MUSHRA based procedures and the single stimulus rating procedure in a sample of young adults with hearing thresholds within normal range. In particular, the purpose of this study was to evaluate the reliability of, and agreement between multi-stimulus MUSHRA based procedures and the single-stimulus rating procedure for the same test stimuli with a normal hearing young adult participant cohort. The test stimuli were selected across a range of common factors used in hearing aid studies, including signal processing, signal to noise ratio, and reverberation.

Materials and methods

Participants

Twenty participants (seven males and thirteen females) were recruited and completed the study protocol. Participant ages ranged from 20-31 years with a mean age of 24.6 (SD 2.95 years). Male ages ranged from 20-24 years (mean 22.7, SD 1.38 years) and female ages ranged from 23-31 (mean 25.72, SD 3.08 years). Participants were native speakers of English, and all participants were tested to confirm that they had audiometric thresholds at 25 dB HL or better between 250 and 8000 Hz, and normal tympanograms prior to starting the study. The experimental procedures were approved by the Western University Human Research Ethics Board (REB 111665). Participants were compensated for their time.

Test Stimuli

The recordings for this study were made at a constant sound level and frequency content to avoid any potential impact on the dimension-specific ratings. Two sets of stimuli were used in this study (Table 1). These two sets were selected as exemplars of two types of sound quality studies: (1) studies that compare sound quality of a signal processor type *across brands*; and (2) studies that compare variations in signal processing *within one hearing aid*. Both stimulus sets were rated using both a modified MUSHRA procedure and single-stimulus ratings, and both were presented in sound field from zero degrees azimuth at 60 dBA, which was a comfortable level for our normal hearing individuals. Participants were not instructed to change the volume of the presentation level. Both stimulus sets used a range of listening conditions, varying the signal to noise ratio, background noise type, and/or reverberation characteristics of the room to generate stimuli that varied in sound quality. This type of environmental manipulation may be done in hearing aid studies when it is of interest to assess signal processing in a range of acoustic settings. Details of the stimuli are provided for each stimulus set below. The stimuli used in this study came from two previously completed hearing aid projects.

MUSHRA with custom anchors stimulus dataset (dataset A)

Stimulus set for the MUSHRA with custom anchors procedure included recordings from four different hearing aid brands, generated from a previous study of noise reduction signal processing (Scollie et al., 2016). Briefly, each hearing aid was programmed to a flat moderate hearing loss using DSLv5 child targets. Sentence pairs (IEEE sentences) were recorded through each hearing aid, in three types of background noise [quiet, speech shaped noise (SSN), and babble] at two different SNRs, [0 dB and 5 dB]. Each hearing aid was recorded twice, with the brand-specific digital noise reduction (DNR) signal processing enabled and disabled. For this study, we chose anchors that were the same sentence pairs, mixed with noise at -10 dB SNR

using both types of noise, and a reference from one hearing aid recorded in quiet with digital noise reduction off. Two additional stimuli were also included by mixing speech and noise at a +10 dB SNR, to better represent low noise listening conditions in the stimulus set.

MUSHRA without anchors stimulus dataset (dataset B)

Stimulus set for MUSHRA without anchor procedure used recordings from one, 20 channel digital behind-the-ear style hearing aid. This aid used a hybrid noise management system that combined multiple microphone options, digital noise reduction, and level dependent speech enhancement to create a set of overall noise management profiles ranging from off, through mild, medium, and strong settings of noise management (Table 2). The hearing aid was programmed to a flat moderate hearing loss to DSL v5.0 adult targets and verified using real ear measures. The aids were then placed on a B&K Head and Torso Simulator (HATS) to record hearing aid output. Hearing In Noise Test (HINT) sentences were presented from a desktop computer through a Tucker Davis system and presented through Anthony Gallo Acoustics Nucleus speakers. The HATS was placed in the participant's position in low- and high-reverberation environments. The low-reverberation room was a large double walled sound booth (Industrial Acoustic Company) with a measured reverberation time (RT60) of 100ms. The high reverberation room was a purpose-built reverberation chamber with the dimensions 6.1m x 4.0m x 2.6m. One corner floor to ceiling curtain (135 cm wide) made of micro suede and quilted micro suede and filled with polyester batting and one corner foam pad (122 x 10 x 60cm) were placed in opposite corners of the room. These baffles reduced the reverberation time of the room to 900 ms. The sentence and noise were presented through a desktop computer with levels controlled by a Sound Web and presented through Tannoy Di5 DC speakers.

Recordings from these two rooms resulted in stimuli from four digital signal processing settings, two SNR levels (0 dB and 5 dB) and at a several noise conditions [quiet, noise only (SSN and babble), reverberation only, SSN + reverberation and babble + reverberation. The reference was recorded in a quiet environment with minimum sound processing. No additional anchors were included, because the low-SNR high-reverberant conditions were sufficient to represent low sound quality in this stimulus set.

Sound Quality Ratings

The MUSHRA based procedures were completed on one day and the total impression rating procedure was completed on a separate day. The time period between visits varied between one to seven days with an average of 2.42 days between visits (SD 1.77 days). The starting rating task (i.e., MUSHRA based or total impression rating procedure) was counterbalanced between participants. The startup selection of MUSHRA with custom anchors (dataset A) or MUSHRA without anchors (dataset B) was randomized using an online tool (random.org). The presentation order of the sound files within each rating experiment was then randomized within the custom software that executed the multiple comparison procedure and total impression rating procedure. The rating instructions were provided to each participant in writing and are presented in the supplemental material. Participants completed the experimental procedure on a computer monitor while seated in a double walled sound booth.

Ratings using MUSHRA with custom anchors for dataset A

The recordings were presented in sound field from zero degrees azimuth. Custom software was developed to mediate the rating data collection. The test stimuli were divided into five sets, with each set representing a noisy condition (quiet, SSN at 0 dB and 5 dB, and babble at 0 dB and 5 dB SNRs). Within each set, there were eleven test stimuli for comparison: the reference recording,

recordings from the four hearing aids with the DNR feature on and off, the +10 dB SNR recording, and the anchor (-10 dB SNR recording).The software presented the five sets twice, to gather test-retest ratings for subsequent reliability analyses. In addition, the software randomized the order of the sets, and the order the test stimuli within each set. Figure 1 represents the user interface of the custom MUSHRA software that mediated the data collection.

The participants listened to the reference, hidden anchors and test stimuli in each set as many times as they needed to make their ratings, before moving on to the next set. Text descriptors were presented on the software on the left hand side (ranging from "Excellent" to "Bad"). A percentage score appeared on top of the slider as participants moved the slider. The software then logged the percentage scores generated during the rating procedures.

Ratings using MUSHRA with no anchors for dataset B

The MUSHRA without anchors procedure was administered using the same custom MUSHRA software in a similar manner, where participants were presented with the reference, test stimuli but no anchors (Table 1). For dataset B, there were ten stimulus sets: 5 noise conditions (quiet, SSN 0 dB SNR, SSN 5 dB SNR, Babble 0 dB SNR, Babble 5 dB SNR) x 2 reverberation conditions. Within each set, there was a reference stimulus and the hearing aid recordings under the four signal processing settings (Table 1). On each set, the participants were asked to listen to the reference and then rate the other recordings in comparison to it. Similar to the MUSHRA with custom anchors procedure, the software presented the ten stimulus sets twice, with the order of the sets and order of the test stimuli within the set randomized.

Both Stimulus Sets: Total Impression Ratings

In this task, the participants were presented with one recording per screen and were asked to move a software slider (Figure 2) to make their rating of the Total Impression of the recording.

Participants were asked to take into consideration the clarity, loudness, background noise, and quality of the recording. The participants listened to the recordings as many times as they needed to before rating the total impression of them. Ratings ranged from 0 to 10 with labels on the screen that provided guidance for the raters (e.g., low score indicating poorer sound quality rating and high score indicating better sound quality rating).

Results

Reliability of the subjective ratings was analyzed first. The recordings from both the MUSHRA with custom anchors and the MUSHRA without anchor (e.g., references and anchors) that were administered multiple times were examined for their test-retest reliability across all participants. The two-sample Kolmogorov-Smirnov (K-S) test was administered on each participant's test-retest data to ascertain whether they came from the same distribution. For both datasets, the K-S test results showed for 18 of the 20 participants, the two sample K-S test did not reveal that the ratings collected during test and retest sessions were from distinct distributions. To further probe the test-retest reliability, the non-parametric equivalent of the Intraclass Correlation Coefficient (ICC) was employed (Rothery, 1979). The non-parametric statistic was used as the results from the Shapiro-Wilk's test on the test-retest data revealed non-normality. The non-parametric concordance coefficient was computed in R version 4.2.2 using the 'nopaco' package, for all the 20 participants and separately for the two datasets. The concordance coefficients ranged between 0.76 to 0.9 for dataset A and between 0.75 to 0.9 for dataset B, with all coefficients statistically significant ($p < 0.025$). As such, the test-retest ratings were averaged for all participants for both MUSHRA based procedures. It is worthwhile to point out that comparable test-retest concordance coefficients were observed for the two MUSHRA based procedures.

The histograms of the test-retest averaged rating data in MUSHRA based procedures and the raw Total Impression rating data from all participants across all test conditions in both datasets are depicted in the panels (a) and (b) of Figures 3 and 4 respectively. Deviation from normality is evident in the rating distributions, especially for the Total Impression procedure. The deviation from normality was statistically confirmed through the Shapiro-Wilk's test of normality. The non-parametric concordance coefficient was therefore once again utilized for assessing the inter-rater reliability. The concordance coefficients across MUSHRA with custom anchors and MUSHRA without anchors datasets were 0.69 and 0.70 for the Total Impression ratings, and 0.70 and 0.73 for the MUSHRA based procedure ratings respectively, all of which were statistically significant ($p < 0.001$).

Figures 3(c) and 4(c) depict the scatter plots between the Total Impression and MUSHRA based procedures both datasets. In order to evaluate how the ratings mapped from one task to the other, a regression analysis was completed using SPSS (SPSS v24; IBM Corporation Chicago, IL). Regression analysis of MUSHRA with custom anchors and MUSHRA without anchors scatter plots revealed correlation coefficients of r= +0.82 and r= +0.85 respectively. Figures 3(d) and 4(d) display the total impression and MUSHRA based ratings averaged across listeners for each test stimulus in both datasets. The correlation coefficients improved to 0.99 and 0.98 respectively, when applied to the condition-averaged data.

It is evident from panels (c) and (d) in Figures 3 and 4, there is considerable variability in both the MUSHRA based procedures and the Total Impression ratings, and evidence suggests that this variability must be considered in comparing different rating methods (Hoßfeld et al., 2011, Parizet et al., 2005). We followed the procedure outlined for Hoßfeld et al., (2011), wherein a quadratic function is fit between the averaged ratings for each test stimulus and the

square of their corresponding standard deviations. The quadratic equation is described as: $y^2 = \alpha(-x^2 + (\vartheta_{min} + \vartheta_{max})x - \vartheta_{min}\vartheta_{max})$, where *x* represents the averaged rating, *y* is the corresponding standard deviation, α is the fitting parameter, and ϑ_{min} and ϑ_{max} are the minimum and maximum rating values respectively. If the rating procedures exhibit similar distribution of ratings across individual participants, then it is expected that they would lead to comparable fitted α parameters (Hoßfeld et al., 2011). We therefore fit separate quadratic functions, parameterized by the α parameter, between the averaged Total Impression or MUSHRA based ratings and their standard deviations, separately for both databases. To derive comparable α parameters, the MUSHRA based ratings were divided by 10, so that both Total Impression and MUSHRA based ratings span between $\vartheta_{min} = 0$ and $\vartheta_{max} = 10$. The α parameter was determined through the least squares fitting function in MATLAB. Figures 5(a) and 5(b) depict the scatter plots of the averaged ratings versus the corresponding standard deviations, along with the fitted quadratic curves, for different rating procedures and across the two datasets. The resulting α parameters were: 0.161 (95% CI: 0.109 – 0.213) and 0.142 (95% CI: 0.1 – 0.185) for the Total Impression Ratings and the MUSHRA with custom anchors ratings; and 0.11 (95% CI: 0.099 – 0.120) and 0.1 (95% CI: 0.092 – 0.108) for the Total Impression ratings and the MUSHRA without anchor ratings. The comparable magnitude of the α parameter highlights the similarity between the speech quality ratings obtained by the Total Impression and MUSHRA based procedures.

Discussion

Hearing aid sound quality is an important factor that relates to overall hearing aid outcome and satisfaction (Picou 2020), preference for one hearing aid over another (Saleh et al., 2021) and can be a reason for hearing aid rejection or dissatisfaction if poor (Abrams & Kihm, 2015).

Therefore, assessment of hearing aid sound quality is sometimes included in laboratory evaluations of new signal processing (Glista et al., 2019; Huber et al., 2018; Scollie et al., 2016). There are, however, multiple ways to assess sound quality including standardized and multi-stimulus MUSHRA based (ITU- R, 2003) and single-stimulus ratings that may include multiple dimensions of sound quality, based on Gabrielsson's work (Gabrielsson et al., 1988).

This study sought to compare the data arising from these two procedures, using a within-subjects design to assess whether and how sound quality ratings differ based on the method used. The findings of this study indicated a high, positive correlation between the ratings of the two procedures. Also, the total impression ratings were positively correlated with both of the multiple comparison datasets.

Findings of this study indicated the same stimuli (e.g., multiple references and anchors) that were rated multiple times (during the MUSHRA based procedures) by each participant had a high test-retest reliability meaning that the multiple ratings of the same stimuli were ranked similarly and consistently by participants. Additionally, the multi-stimulus MUSHRA based procedures and the single stimulus rating procedure used in this study showed a high degree of correspondence and agreement in rating the same stimuli. The outcomes of this study suggest that a single-stimulus rating scale, without references and anchors, may be able to generate reliable, consistent and comparable sound quality data when compared to a standardized and commonly used multiple comparison procedures for young, normal-hearing participants. However, the degree to which this result will generalize to other signal processing conditions and other participant populations is unknown and would require further evaluation for conditions not evaluated in this study.

Recall that the participants listened to the recordings as many times as they needed to before rating them on the MUSHRA based procedure. This mixed evaluation procedure allowed the participants to compare different recordings in comparison to the reference before finalizing their choice of rating. This methodology is backed up in the literature where a previous study showed a high accuracy in evaluating the pleasantness of different sounds (Parizet et al., 2005).

A Standard deviation of Opinion Scores (SOS) analysis was as the Mean Opinion Score, which is the average of the participants' ratings, does not represent rating diversity and therefore only captures a part of the raters' perception. A combined SOS-MOS analysis (Hoßfeld et al., 2011) revealed similar patterns between the total impression ratings and the MUSHRA with custom anchors dataset, as well as the total impression ratings and the MUSHRA without anchor ratings dataset.

The MUSHRA without anchor ratings had a higher correlation with the total impression ratings compared to the MUSHRA with custom anchors dataset. Recall that the MUSHRA without anchor dataset used in this study did not have any anchors, which are exemplars of a low-quality sentence pairs. This modification to the MUSHRA procedure was used in this study to extend the comparison to additional stimulus conditions and multiple variations in MUSHRA implementations. The high correlation can be best explained by the fact that the total impression rating scale does not use anchors, just like MUSHRA without anchor dataset, and thus the sound quality ratings of the participants were not influenced by the anchor recordings.

The MUSHRA process includes multiple stages of recording the sentence pairs and processing the anchors and test stimuli, such as by peak clipping or asymmetrically distorting the signal. Therefore, setting up the MUSHRA test procedure (or modified versions of it) might require additional efforts, cost and time (Völker et al., 2016). Additionally, participants might be

subjected to fatigue due to the extended duration of the test having them listening and rating multiple references, anchors and stimuli. Unlike the single stimulus rating scale, the ratings could be administered on a paper copy (rather than on a computerized interface to accommodate wider range of lab set-ups) or digitally and can be a relatively easy and straightforward procedure.

Data collected from sound quality measures are used by hearing aid manufactures and researchers to reinforce and strengthen the objective sound quality models (Völker et al., 2016). Having sound quality data of both subjective and objective procedures in agreement and highly correlated, assists in improving the output of digital signal processing of the HAs. For instance, a previous study indicated a high correlation between the MUSHRA ratings and an objective model of sound quality (Muralimanohar et al., 2013). Findings of the current study demonstrated a high correlation between total impression and MUSHRA based procedures. However, results from MUSHRA without anchors and total impression scales are to be interpreted with caution in the context of reinforcing objective sound quality paradigms. The use of anchors is recommended to reduce any subjective biases that could impact the quality of prediction of mathematical and objective sound quality models (Zielinski et al., 2008).

A limitation of this study was recruiting a small sample (N= 20) where all participants were young adults who had normal hearing. As a result, the generalizability of these results to those who are hearing aid users is unknown. Future work is encouraged to consider recruiting participants of different ages with and without hearing loss, and to consider the role of hearing aid experience. The datasets used in this study varied in noise reduction properties, so future studies should assess the sound quality of other signal processing strategies and to investigate how the MUSHRA ratings relate to other dimensions of the Gabrielsson's single-stimulus rating

scale. Another limitation of this study was including recordings made at a constant sound level. Future studies are recommended with loudness equalization for the recorded sentence pairs.

Conclusion

In conclusion, the current study indicated a high correlation between sound quality data collected via MUSHRA with custom anchors, MUSHRA without anchor and total impression rating scales. Participants reported consistent ratings on the Gabrielsson's total impression scale compared to MUSHRA with custom anchors and MUSHRA without anchor procedures without having to compare the test stimuli to hidden references and anchors. The efficiency and simplicity of administering single stimulus rating scale may facilitate the collection of subjective sound quality data across larger number of participants and wider range of digital signal processing strategies.

Data Availability Statement

The datasets generated during and/or analyzed during the current study are not publicly available due to Western University Human Research Ethics Board restriction but are available from the corresponding author on reasonable request.

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.

Arehart K.H., Kates J.M., Anderson M.C. & Harvey L.O. (2007). Effects of noise and distortion on speech quality judgments in normal-hearing and hearing-impaired listeners. *Journal of the Acoustical Society of America, 122*, 1150 – 1164.

Arehart K.H., Kates J.M. & Anderson M.C. (2010). Effects of noise, nonlinear processing, and linear filtering on perceived speech quality. Ear & Hearing, 31, 420 – 436.

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

Banerjee, S. (2011). Hearing Aids in the Real World: Typical Automatic Behavior of Expansion, Directionality, and Noise Management, 22, 34–48.

Barry, S. J., & Kidd Jr, G. (1981). Psychophysical scaling of distorted speech. *Journal of Speech, Language, and Hearing Research, 24*(1), 44-47.

Caldwell, M. T., Jiam, N. T., & Limb, C. J. (2017). Assessment and improvement of sound quality in cochlear implant users. *Laryngoscope Investigative Otolaryngology*, *2*(3), 119–124. https://doi.org/10.1002/lio2.71

Cord, M. T., Surr, R. K., Walden, B. E., & Dyrlund, O. (2004). Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *Journal of the American Academy of Audiology, 15*(05), 353-364.

Falk, T. H., Parsa, V., Santos, J. F., Arehart, K., Hazrati, O., Huber, R., Kates, J. M., & Scollie, S. (2015). Objective Quality and Intelligibility Prediction for Users of Assistive Listening Devices: Advantages and limitations of existing tools. *IEEE Signal Processing Magazine, 32*(2), 114–124. https://doi.org/10.1109/MSP.2014.2358871

Franz, S., & Bitzer, J. (2010). Multi-channel algorithms for wind noise reduction and signal compensation in binaural hearing aids. In *Proc. Intl. Workshop Acoust. Echo Noise Control (IWAENC)*.

Gabrielsson, A., Schenkman, B. N., & Hagerman, B. (1988). The effects of frequency responses on sound quality judgments and speech intelligibility. *Journal of Speech and Hearing Research*, *31*, 166–177.

Gabrielsson, A., Hagerman, B., Bech-Kristensen, T., & Lundberg, G. (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. https://doi.org/10.1121/1.399713

Gabrielsson, A., & Sjögren, H. (1979). Perceived sound quality of sound-reproducing systems. *The Journal of the Acoustical Society of America*, *65*(4), 1019–1033. https://doi.org/10.1121/1.382579

Glista, D., Hawkins, M., Vaisberg, J. M., Pourmand, N., Parsa, V., & Scollie, S. (2019). Sound Quality Effects of an Adaptive Nonlinear Frequency Compression Processor with Normal-Hearing and Hearing-Impaired Listeners. *Journal of the American Academy of Audiology*, *30*(07), 552–563. https://doi.org/10.3766/jaaa.16179

Hoßfeld, T., Schatz, R., & Egger, S. (2011). SOS: The MOS is not enough!. In *2011 third international workshop on quality of multimedia experience* (pp. 131-136). IEEE.

Huber, R., Bisitz, T., Gerkmann, T., Kiessling, J., Meister, H., & Kollmeier, B. (2018). Comparison of single-microphone noise reduction schemes: can hearing impaired listeners tell the difference?. *International Journal of Audiology*, *57*(sup3), S55-S61.

Huber, R., & Kollmeier, B. (2006). PEMO-Q—A New Method for Objective Audio Quality Assessment Using a Model of Auditory Perception. *IEEE Transactions on Audio, Speech,*

and Language Processing, *14*(6), 1902–1911. https://doi.org/10.1109/TASL.2006.883259

Huber, R., Parsa, V., & Scollie, S. (2014). Predicting the Perceived Sound Quality of Frequency-Compressed Speech. *PLoS ONE*, *9*(11), e110260. https://doi.org/10.1371/journal.pone.0110260

IEEE. (1969). *IEEE Recommended Practice for Speech Quality Measurements* (Institute of Electrical and Electronic Engineers, New York).

ITU-R. (2003). *Recommendation BS.1534: Method for the subjec- tive assessment of intermediate quality levels of coding systems*. Geneva, Switzerland: International Telecommunications Union.

Kates J.M. & Arehart K.H. (2010). The hearing-aid speech quality index (HASQI). *Journal of the Audio Engineering Society*, *58*, 363 – 381.

Kochkin S. 2010. MarkeTrak VIII: Customer satisfaction with hearing aids is slowly increasing. *Hearing Journal*, *63*, 11 – 19.

Kondo, K. (2012). *Subjective quality measurement of speech: its evaluation, estimation and applications*. Springer Science & Business Media.

Lundberg, G. Ovegård, A. Hagerman, B., Gabrielsson, A., & Brändström, U. (1992) Perceived Sound Quality in a Hearing Aid with Vented and Closed Earmould Equalized in Frequency Response, *Scandinavian Audiology*, *21*(2), 87-92, DOI: 10.3109/01050399209045987

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

Möller, S., Chan, W.-Y., Côté, N., Falk, T. H., Raake, A., & Wältermann, M. (2011). Speech Quality Estimation: Models and Trends. *IEEE Signal Processing Magazine*, *28*(6), 18–28. https://doi.org/10.1109/MSP.2011.942469

Muralimanohar, R. K., Kronen, C., Arehart, K., Kates, J., & Pichora-Fuller, M. K. (2013). Quality of voices processed by hearing aids: Intra-talker differences. *Proceedings of Meetings on Acoustics*, *19*(1), 060112. https://doi.org/10.1121/1.4800397

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

Nilsson, M., Soli, S. D., & Sullivan, J. A. (1994). Development of the Hearing In Noise Test for the measurement of speech reception thresholds in quiet and in noise. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.408469

Parizet, E., Hamzaoui, N., & Sabatie, G. (2005). Comparison of some listening test methods: a case study. *Acta Acustica united with Acustica*, *91*(2), 356-364.

Parsa, V., Scollie, S., Glista, D., & Seelisch, A. (2013). Nonlinear Frequency Compression: Effects on Sound Quality Ratings of Speech and Music. *Trends in Amplification*, *17*(1), 54–68. https://doi.org/10.1177/1084713813480856

Roy, A. T., Jiradejvong, P., Carver, C., & Limb, C. J. (2012). Musical Sound Quality Impairments in Cochlear Implant (CI) Users as a Function of Limited High-Frequency Perception. *Trends in Amplification*, *16*(4), 191–200. https://doi.org/10.1177/1084713812465493

Saleh, H. K., Folkeard, P., Van Eeckhoutte, M., & Scollie, S. (2021). Premium versus entry-level hearing aids: Using group concept mapping to investigate the drivers of preference.

International Journal of Audiology, 1–15. https://doi.org/10.1080/14992027.2021.2009923

Salehi, H., Suelzle, D., Folkeard, P., & Parsa, V. (2018). Learning-based reference-free speech quality measures for hearing aid applications. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 26(12), 2277-2288.

Scollie, S., Levy, C., Pourmand, N., Abbasalipour, P., Bagatto, M., Richert, F., Moodie, S., Crukley, J., & Parsa, V. (2016). Fitting Noise Management Signal Processing Applying the American Academy of Audiology Pediatric Amplification Guideline: Verification Protocols. *Journal of the American Academy of Audiology*, *27*(03), 237–251. https://doi.org/10.3766/jaaa.15060

Studebaker, G. A., & Sherbecoe, R. L. (1988). Magnitude estimations of the intelligibility and quality of speech in noise. *Ear and Hearing*, *9*(5), 259-267.

Simonsen, C. S., & Legarth, S. V. (2010). A procedure for sound quality evaluation of hearing aids. Hearing Review, 17(13), 32-37.

Vaisberg, J., Folkeard, P, Levy, S, Dundas, D, Agrawal, S, & Scollie, S. (2021). Sound Quality Ratings of Amplified Speech and Music Using a Direct Drive Hearing Aid: Effects of Bandwidth, *Otology & Neurotology 42*(2), 227-234 doi: 10.1097/MAO.0000000000002915

Völker, C., Bisitz, T., Huber, R., Kollmeier, B., & Ernst, S. M. A. (2018). Modifications of the MUlti stimulus test with Hidden Reference and Anchor (MUSHRA) for use in audiology. *International Journal of Audiology*, *57*(sup3), S92–S104. https://doi.org/10.1080/14992027.2016.1220680

Zielinski, S., Rumsey, F., & Bech, S. (2008). On some biases encountered in modern audio quality listening tests-a review. *Journal of the Audio Engineering Society*, *56*(6), 427-451.

Figure caption

Figure 1- MUSHRA with custom anchors and MUSHRA without anchor user interface screenshot.

Figure 2. Gabrielsson's Total Impression user interface.

Figure 3. Comparison of sound quality rating methods for MUSHRA with custom anchors (labeled Dataset A). (a) Histogram of the raw ratings across all conditions for the MUSHRA method with custom anchors. (b) Histogram of the raw ratings across all conditions from the Total Impression method. (c) Scatter plot between the Total Impression and MUSHRA with costumed Anchors ratings (MUSHRA-CA). (d) Condition-averaged Total Impression and MUSHRA with costumed anchors ratings, where the data was sorted such that the Total Impression ratings ranged from the lowest to the highest. Error bars represent the standard deviation of the corresponding mean rating.

Figure 4. Comparison of sound quality rating methods for MUSHRA without anchors (labeled Dataset B). (a) Histogram of the raw ratings across all conditions for the MUSHRA without anchor method, where no anchors were present. (b) Histogram of the raw ratings across all conditions from the Total Impression method. (c) Scatter plot between the Total Impression and MUSHRA without anchor ratings (MUSHRA-WA). (d) Condition-averaged Total Impression and MUSHRA without anchor ratings for Dataset B, where the data was sorted such that the Total Impression ratings ranged from the lowest to the highest. Error bars represent the standard deviation of the corresponding mean rating.

Figure 5 – Scatter plots of the mean sound quality ratings versus their corresponding standard deviations, for different rating procedures and datasets. The least-squares curve fits to the scatter data are also shown. (a) Dataset A, and (b) Dataset B.

Table caption

Table 1: Summary of stimulus conditions for the two datasets.

Table 2: MUSHRA without anchor dataset - Hearing aid settings.