Citation/Reference	Denys S, Latzel M, Francart T, Wouters J A preliminary investigation into hearing aid fitting based on automated real-ear measurements integrated in the fitting software: test-retest reliability, matching accuracy and perceptual outcomes. Int. J. Audiol., 1-9, Dec 2018
Archived version	Author manuscript: the content is identical to the content of the published paper, but without the final typesetting by the publisher
Published version	Sam Denys et al 2018. J. Int. Aud.
DOI	10.1080/14992027.2018.1543958
Journal homepage	J. Int. Aud.
Author contact	sam.denys@kuleuven.be , jan.wouters@kuleuven.be

A Preliminary Investigation into Hearing Aid Fitting Based on Automated Real-Ear Measurements
integrated in the Fitting Software: Test-Retest Reliability, Matching Accuracy and Perceptual Outcomes

Sam Denys¹, Matthias Latzel², Tom Francart¹, and Jan Wouters¹

¹KU Leuven, Department Of Neurosciences, ExpORL, Leuven, Belgium

²Phonak AG, Staefa, Switzerland

Corresponding author

Sam Denys

Herestraat 49 - box 721

3000 Leuven

Belgium

sam.denys@kuleuven.be

0032 16 37 69 61

16 Keywords (5-6)

17

18 Real-Ear Measurements (REM); verification; speech intelligibility; hearing aid fitting; target mismatch

19 Acronyms & abbreviations

20

21 DSL Desired Sensation Level

22 FF FirstFit

23 HI Hearing Impaired

24 ISTS International Speech Test Signal

25 LIST Leuven Intelligibility Sentence Test

26 MF Manual Fitting

27 NAL-NL National Acoustic Laboratories-Non Linear

28 NVA Nederlandse Vereniging voor Audiologie

29 REAR Real-Ear Aided Response

30 REM Real-Ear Measurement

31 REUG Real-Ear Unaided Gain

32 SNR Signal-to-Noise Ratio

33 $SRT_{n/q}$ Speech Reception Threshold (n: in noise, q: in quiet)

34 SSQ Speech Spatial and Qualities

35 TM TargetMatch

36 Acknowledgements

37

38 The authors thank Katleen De Voecht of the UZ-MUCLA (Multidisciplinair Centrum voor Logopedie en
39 Audiologie of the UZ Leuven University Hospital, ENT-Department) and Laure Jacquemin for their help
40 with patient recruitment and data acquisition.

41 Professional meeting details

42

43 Portions of this article were presented at the International Hearing Aid Research Conference (IHCON),

44 Granlibakken, Tahoe City, California, USA, August 10-14, 2016, and at the Internationaler

45 Hörgeräteakustiker-Kongress (EUHA) Hannover, Germany, October 19-21, 2016.

46 Funding details

47

48 This study was funded by Phonak AG.

49 Disclosure statement

50

51 Matthias Latzel is a Clinical Research Manager at Phonak. He was involved in the study design. No other

52 conflicts of interest are declared for any of the authors.

53

54 Abstract

55

56 *Objective*

57 The recent integration of automated Real-Ear Measurements (REM) in the fitting software facilitates
58 the hearing aid fitting process. Such a fitting strategy, TargetMatch, was evaluated. Test-retest
59 reliability and matching accuracy were quantified, and compared to a REM-based fitting with manual
60 adjustment. Also, it was investigated whether TargetMatch leads to better perceptual outcomes
61 compared to a FirstFit (using software predictions only).

62 *Design and study sample*

63 Ten hearing impaired participants were enrolled in a counterbalanced single-blinded cross-over study
64 comparing TargetMatch and FirstFit. Aided audibility, speech intelligibility and real-life benefits were
65 assessed. Repeated measurements of both TargetMatch and REMs with manual adjustment were
66 performed.

67 *Results*

68 Compared to a REM-based fitting with manual adjustment, TargetMatch had higher test-retest
69 reliability. Also, TargetMatch outperformed the other fitting strategies in terms of matching accuracy.
70 Compared to a FirstFit, improved aided audibility and real-life benefits were found. Speech
71 intelligibility did not improve.

72 *Conclusions*

73 Preliminary data suggest that automated REMs increase the likelihood of meeting amplification targets
74 compared with a FirstFit. REMs integrated in the fitting software provide additional reliability and
75 accuracy compared to traditional REMs. Findings need to be verified in a larger and more varied
76 sample.

Introduction

A major goal of hearing aid fitting is to improve the audibility of speech (Johnson, 2013). To this end, validated fitting rules, such as the NAL-NL2 (Keidser et al, 2011) or DSL-v5 (Scollie et al, 2005), prescribe amplification targets that should be reached in the ear of an individual patient. It has been shown that fitting to target leads to improved speech intelligibility in quiet and in noise (Byrne, 1986; Alcantara et al, 2004) and, more generally, that appropriate gain provision is linked to benefits for hearing aid users, and - in turn - to user satisfaction (Kochkin et al, 2010; Abrams et al, 2012; Baumfield & Dillon, 2001). Real-Ear Measurements (REMs) are useful to verify whether the output of the hearing aid at the eardrum matches the desired prescribed target, and to be informed on how to adjust the amplification to improve this match. After more than 30 years of commercial availability, and despite being the recommended standard of care by many professional societies, US surveys indicate an underuse of REMs. While the majority of hearing healthcare professionals own verification equipment, they are not routinely verifying real-ear amplification (Mueller & Picou, 2010). The perceived benefit of performing REMs relative to clinical time demands is rather low (Kochkin et al, 2010; Abrams et al, 2012).

A large body of research demonstrates a failure to match prescribed amplification targets, using software predictions only (a so-called FirstFit, FF) (Aazh & Moore, 2007; Hawkins & Cook, 2003; Aarts & Caffee, 2005; Abrams et al, 2012; Aazh et al, 2012). In addition, boundaries for matching accuracy are not clearly defined. However the British Society of Audiology (2007) defines the acceptable margins of error to be ± 5 dB (up to 2000 Hz) and ± 8 dB (for higher frequencies). FFs are based on simulations of in situ output. These simulations might work for patients with ear canals that have close to average acoustic characteristics, but are far from optimal for many other patients. For instance, large inter-subject differences in the Real-Ear Unaided Gain or Real-Ear to Coupler Difference have been demonstrated (Valente & Goebel, 1991; Fikret-Pasa & Revit, 1992). REMs take these sources of

variability into account. Therefore, using REMs, more fittings are expected to be on target (Aazh & Moore, 2007).

Recent research indicates an improved perceived quality of service when patients were fitted based on REMs (Amlani et al, 2016). Studies investigating patients' self-perceived benefit in terms of speech intelligibility in daily life or general improved hearing ability when fitted with REMs compared to a FF are limited (Abrams et al, 2012). Furthermore, studies investigating the objective merits of REMs in terms of improved perceptual outcomes (such as aided audibility, aided speech intelligibility performance in quiet and in noise) compared to a FF are lacking.

Using currently recommended clinical procedures, REMs are considered reliable and accurate. Key to appropriate high frequency amplification measurement is the probe tube insertion depth, which interacts with standing wave effects in the ear canal (Mueller, 2001). Other factors influencing the reliability of REMs are probe tube movements between unaided and aided measurements, and the experience level of the hearing healthcare professional conducting the measurements (Jespersen & Møller, 2013).

In order to increase the clinical use of REMs, hearing aid manufacturers have been developing integrated REMs in the fitting software with automatic fitting to target (Koehler & Kulkarni, 2014; Beck & Crowe, 2017). This approach has several benefits over performing iterative measurements and manually matching targets. It minimizes the need to switch between fitting and verification software, and the implementation of automated fitting procedures that adjust multiple controls in bilateral devices simultaneously leads to clinical time-saving. Also, it makes measurement time relatively independent of the experience level of the audiologist. By efficiently guiding the hearing healthcare professional through the necessary measurements, the complexity of REM procedures is reduced. While integrated and automated REM protocols seem to agree well with manual REM-based fitting (Koehler & Kulkarni, 2014), the test-retest reliability of integrated REMs with automatic fitting has not yet been reported (Digiovanni & Pratt, 2010; Yanz et al, 2007; Koehler & Kulkarni, 2014).

In this study, a new REM-based fitting strategy integrated in Phonak fitting software with automatic fitting to target, TargetMatch (TM), was evaluated, and compared to a FF approach and a traditional non-integrated REM-based strategy with manual fitting to prescribed targets (MF). Increased reliability and accuracy are aimed for with TM, ensuring deep probe insertion in each patient and automatic fitting to prescriptive targets.

The following research questions were addressed: (1) How reliable is TM in hearing impaired (HI) participants, compared to a MF approach? (2) Are targets more accurately matched with TM compared to MF and FF? (3) Do HI participants benefit from using TM over a FF approach in terms of improved aided audibility, speech intelligibility in quiet and in noise, and self-reported benefit?

We hypothesize TM to outperform MF in terms of test-retest reliability, given the automatic progress of the measurements. We expect prescribed targets being well matched within recommended acceptable margins (BSA, 2007). Significantly higher mismatches to target (especially in the high frequencies) are expected using a FF approach (Aazh & Moore, 2007; Hawkins & Cook, 2003; Aarts & Caffee, 2005; Abrams et al, 2012; Aazh et al, 2012). Improved perceptual outcomes as well as self-reported benefit are expected when targets are more accurately matched (Byrne, 1986; Alcantara et al, 2004; Baumfield & Dillon, 2001; Abrams et al, 2012). Hence, hearing aid fittings with TM should result in better perceptual outcomes.

Method

Participants

Participants were recruited via Leuven University Hospitals. Eligible participants had unaided hearing thresholds within the fitting range of Bolero V-70-P hearing aids (Phonak), the devices used in this study. The participants' age at the start of the study ranged from 22 to 76 years (mean $\pm$ SD = 43 $\pm$ 21 years). All participants were native Dutch speakers and did not show signs of severe cognitive impairment, as indicated by a score of $\geq$ 21 on the Montreal Cognitive Assessment screening test (Nasreddine et al, 2005). The audiograms, measured at the start of the study, are shown in **Figure 1**. All participants had a bilateral sensorineural hearing loss, with an air bone gap averaged over the frequencies 500-4000 Hz of $\leq$ 15 dB HL. According to the rather strict hearing loss symmetry criteria of Margolis & Saly (2008), 6 participants had an asymmetrical hearing loss. Otoscopy showed no obstructing cerumen. All participants were active and experienced bilateral hearing aid users, with hearing aid experience ranging from 6 to 22 years. Participants signed an informed consent form. The study was approved by the Ethics Committee of University Hospitals Leuven.

Materials

Unaided hearing thresholds were measured with a Madsen Orbiter 922 (GN Otometrics) audiometer, connected to TDH-39 headphones (Telephonics) and a B71 bone conductor (RadioEar) in a sound-proof booth. Otoscopy was done with a Beta 100 otoscope (Heine). Ear impressions were taken for the production of custom-made earmolds, using standard clinical materials.

Participants were fitted with Bolero V-70-P (Phonak) behind-the-ear hearing aids, using the manufacturer's fitting software (Target, version 4.3) and an iCube II (Phonak) for wireless transfer of fitting data.

All participants were fitted with earmolds. Venting parameters were selected in relation to their unaided hearing thresholds for frequencies between 250 and 1000 Hz, according to guidelines provided by Sandlin (2000), and varied from 0 to 2 mm.

REMs with silicone probe tubes were administered with an Aurical FreeFit in combination with an Aurical 1081 loudspeaker (GN Otometrics). The REM equipment was controlled by Otosuite verification software, version 4.76 via a wireless Bluetooth connection. Participants were seated facing the loudspeaker at a distance of 1 meter and a 0° angle. Fittings were done in a sound-proof room. Fitting software and verification software modules were integrated in NOAH 4 (HIMSA).

Behavioral measurements and self-report measures

Aided audibility was assessed using the Phoneme Perception Test (Schmitt et al, 2016), a NOAH compatible module developed by Phonak. Aided detection thresholds for narrow band noises were measured, analogous to measuring a free field audiogram. Stimuli were played by the Aurical loudspeaker and routed via the Aurical AUD soundcard.

Speech intelligibility testing was also done in free field with the participant facing a Genelec 8020A loudspeaker at a distance of 1 meter. All speech stimuli were presented using the speech-in-noise module of APEX 3.1 software (Francart et al, 2008; Francart et al, 2017). For speech intelligibility in quiet, the Nederlandse Vereniging voor Audiologie (NVA) word material was used (Wouters et al., 1994).

For speech intelligibility in noise, the Leuven Intelligibility Sentences Test (LIST) was conducted (van Wieringen & Wouters, 2008). A Fireface UCX soundcard (RME) was used to present the signals.

Self-perceived benefit was assessed with the Speech, Spatial and Qualities questionnaire (SSQ12, Noble et al, 2013).

Procedures

Course of the study

Figure 2 shows a scheme of the study protocol, together with the outcome measures obtained at the different study visits.

All participants were invited for an intake visit, in which an otological history was taken, as well as an audiological assessment including otoscopy and pure tone audiometry. Ear impressions were taken.

Following the intake visit, the participants were enrolled in a repeated measures cross-over study (3 visits). TM was compared with a FF with respect to target mismatches and perceptual outcomes. More precisely, participants were randomly assigned to start the trial (visit 1) either with a FF (N=5) or with TM (N=5), both according to the NAL-NL2 fitting rule. Immediately after the fitting (within the same visit), as well as after a home trial of about 2 weeks, a series of perceptual tests was conducted (visit 2). Also, participants had to fill in the SSQ12 during the home trial.

Next, the participant switched to the other fitting strategy (visit 2) followed by a perceptual evaluation, a home trial and a perceptual re-evaluation (visit 3). Additionally, TM was compared to a MF strategy, both targeting the DSL-v5 fitting rule (Scollie et al, 2005)¹. In order to assess and compare measurement reliability, in half of the participants a test-retest with TM (to DSL-v5) was done. In the other half the MF (to DSL-v5) was retested. Test- and retest measurements were spread over 2 visits.

Fittings

Before each fitting, the acoustic parameters were entered in the software (acoustic coupling type, venting parameters, tubing, etc.), the fitting rule was chosen, and a feedback test was performed. The

¹ The DSL-v5 fitting rule was chosen to challenge the system to matching targets that prescribe higher output in low and high frequencies.

feedback test may impose a limit to the output that can be reached, but such a limit would not be affected by whether the gain is set automatically (by TM or FF) or manually (MF).

FirstFit (FF)

Amplification (for 50, 65 and 80 dB SPL input signals) was calculated by the fitting software based on average REUG data. REMs were only used to verify but not to adjust the amplification. REMs were performed to assess REARs for a 50, 65 and 80 dB SPL International Speech Test Signal (ISTS) input (Holube et al, 2010). The measured REARs were used to calculate mismatches to prescribed fitting targets. Furthermore, obtaining the REARs in the FF condition ensured participants not to be biased with respect to the fitting strategy they were currently using (Naylor et al, 2015).

Manual Fitting to Target (MF)

This fitting strategy makes use of REMs. It is a non-integrated and non-automated REM-based fitting strategy, which requires switching between verification and fitting software and manual adjustments. Individual REUGs were measured by the verification software for a pink or white noise (automatically chosen by the system, depending on ambient noise levels). Tragus markers on the REM-probes were placed at 27 mm, and inserted to the point where the tragus markers reached the *insicura intertragica*. As such, a probe tip-eardrum distance of less than 5 mm was aimed for (constant insertion depth method). In case of REUG values lower than -5 dB at 6000 Hz, the probe was moved a little further towards the eardrum. Average REUG data in the fitting software were manually replaced by these individual REUGs. Afterwards REARs were measured for a 50, 65 and 80 dB SPL ISTS input signal, i.e. the REAR50, REAR65 and REAR80, respectively. Amplification was manually adjusted based on the mismatches to prescribed targets. This process was iterated until targets were matched within 5 dB from the target.

TargetMatch (TM)

This new fitting strategy is also based on REM. The same steps as for the MF were conducted and the same test signals were used, but - compared to a traditional REM-based approach - this fitting was automated and integrated in the fitting software. A guided probe tube placement functionality was used to insert the probe. This acoustical positioning method (Storey & Dillon, 2001) shows the simulated probe tube insertion depth on-screen, and the software indicates when the probe tube was positioned 5 mm from the eardrum. Afterwards, TM moves on to measuring the REUG and - after insertion of the hearing aids - other transfer functions (such as occluded measurements, the real-ear to coupler difference and microphone location effects), thereby automatically communicating with the fitting and verification software (e.g. switching on and off the hearing aid when necessary and playing the different test signals). Finally, the software automatically adjusts the amplification to match prescriptive REAR targets.

Behavioral measurements and self-report measures

All perceptual tests were performed with bilateral aids and in free field. The order of tests was counterbalanced. All tests were performed with a so-called calm situation program of the hearing aid, ensuring all adaptive features (e.g. noise reduction strategies and directionality) were fixed across tests and participants.

Aided audibility

Aided detection thresholds for narrow band noises of 3000, 5000 and 6000 Hz were measured. A 2-up 4-down method was used.

Speech Intelligibility in Quiet

NVA lists were presented at three fixed presentation levels individually selected to bracket the level at which 50% of the speech material was correctly identified. The speech material consists of 15 lists of 12 monosyllabic consonant-vocal-consonant words, uttered by a male speaker.

Phoneme scoring was used to calculate percentage correct scores. To increase measurement precision, per level 2 lists were presented and the averaged %-correct score was calculated². Using a logistic regression, the speech reception threshold in quiet (SRTq) was estimated, using the following formula:

$$\text{percentage correct score} = \frac{100}{100 + e^{4 \cdot \text{slope} \cdot (\text{SRT} - \text{presentation level})}}$$

Speech Intelligibility in Noise

LIST-sentences were presented in speech-shaped stationary noise, using an adaptive procedure targeting 50% correct. The speech material consists of 35 lists of 10 sentences each, uttered by a female speaker.

The noise level was fixed at 65 dB A. The level of speech was varied in 2 dB steps. A keyword scoring was used. A training list was assessed, with a starting Signal-to-Noise Ratio (SNR) of 0 dB A. Next, 2 other lists were assessed to increase measurement precision. The start-SNR was set 4 dB below the speech reception threshold in noise (SRTn) reached in the training. The first sentence was presented at increasing levels (in steps of 2 dB) until correctly identified. SRTs were calculated as the averaged SNR of the last 6 sentences, including the SNR of a fictive 11th sentence (not presented). The averaged SRTn (in dBA SNR) of the 2 test lists was calculated.

Speech Spatial and Qualities (SSQ12)

This questionnaire assesses experiences with hearing aids in daily life for 3 important domains: speech intelligibility, spatial hearing and sound quality. Statements needed to be scored on a visual-analog scale, ranging from 0 to 10.

Statistical analyses

²For the first 4 participants, 2 (different) lists were presented at fixed levels of 65, 50 and 45 dB A. At these levels, scores were all well above the SRTq. Therefore, SRTq estimates could only be obtained in 6 participants.

293
 294 Per participant, REUG data were averaged across ears per 1/3 octave frequency between 250 and 8000
 295 Hz prior to the statistical analyses. For the REAR, also an average across levels (50, 65, 80 dB SPL) was
 296 calculated.

297 To investigate the difference (including sign) in matching accuracy between fitting strategies, non-
 298 parametric paired Wilcoxon signed rank tests were performed per frequency. In case of multiple
 299 comparisons, a Bonferroni correction to the p-value was applied.

300 Test-retest reliability per frequency was calculated as the root mean square of the standard deviation
 301 of repeated measurements. As such, per frequency, test-retest reliability is expressed as a single group
 302 SD estimate, calculated as the square root of the average individual variance estimates (SD_i): *test –*

303
$$retest\ reliability = \sqrt{\frac{1}{n} \sum_{i=1}^n SD_i^2}.$$

304 The target mismatch per frequency was calculated per participant as the difference (including sign)
 305 between the prescribed fitting target (expressed in REAR) and the measured REAR. Per frequency,
 306 mismatches were averaged across ears, levels and sessions prior to the statistical analysis³.

307 Non-parametric statistical tests were used because of non-normally distributed data. A significance
 308 level of $\alpha = 0.05$ was used. Depending on the hypothesis, statistical testing was one- or two-tailed.

³ For TM and MF according to the DSLv5 fitting rule, analyses were performed on test data only, since not all participants were tested twice with each fitting strategy (retest data were discarded).

Results

Test-retest reliability

Test-retest reliability estimates per frequency are shown in Table 1.

Test-retest reliability of the REUG decreases with increasing frequency. For TM, test-retest reliability is generally better than 1.5 dB. For MF, test-retest reliability was slightly lower (especially in the high frequencies), but generally still below 2 dB.

For the REAR, with TM, test-retest differences were generally below 2 dB, and relatively independent of frequency. For MF, test-retest differences were generally larger than 1.5 dB, with the greatest differences being 3.2 and 4.2 dB at 250 and 8000 Hz, respectively. Compared to TM, test-retest reliability was worse, except for frequencies between 500 and 1000 Hz. As for the REUG measures, test-retest reliability tended to decrease with increasing frequency.

Mismatch to target

Figure 3 shows the deviation from DSLv5 targets for TM and MF (upper panel), and from NAL-NL2 targets for TM and FF (lower panel). Significantly better target matching was found with TM for frequencies between 800 and 5000 Hz for the DSLv5 formula, and for all but one frequency between 315 and 5000 Hz with the NAL-NL2 formula. Median differences per frequency, test statistics and p-values can be found in Table 2. Averaged across frequencies, the mismatches with TM were 2.6 dB (SD = 1.1 dB) and 1.5 dB (SD = 1.2 dB) for DSLv5 and NAL-NL2, respectively. For MF and FF, the average target mismatches were higher, and were 5.4 dB (SD = 1.3 dB) and 4.3 dB (SD = 1.8 dB), respectively.

Perceptual outcomes of TM versus FF in HI participants

335 Because no significant differences were found with respect to the perceptual outcome measures
 336 before and after the 2-week home trial, except for aided audibility in the TM condition [FF: $z = -0.61$,
 337 $p = 0.29$ (phoneme perception test⁴); $z = -1.89$, $p = 0.06$ (speech intelligibility in quiet⁵); $z = -1.28$, $p =$
 338 0.11 (speech intelligibility in noise); TM: $z = -2.06$, $p < 0.05$ (PPT⁴); $z = -1.63$, $p = 0.09$ (speech
 339 intelligibility in quiet⁵); $z = -0.41$, $p = 0.36$ (speech intelligibility in noise)], outcomes were averaged
 340 across sessions to investigate perceptual differences between the FF and TM fitting strategies. The
 341 results are shown in Figure 4. Statistical analysis revealed significantly lower aided thresholds at 5000
 342 Hz with TM compared to FF [Mdn_{FF} = 28 dBA SPL vs Mdn_{TM} = 25 dBA SPL; $z = -2.00$, $p < 0.05$], as well as
 343 a lower averaged aided threshold across 3000, 5000 and 6000 Hz with TM [Mdn_{FF} = 31 dBA SPL vs
 344 Mdn_{TM} = 29 dBA SPL; $z = -1.84$, $p < 0.05$]. Also, participants reported higher scores on the SSQ12 for
 345 the speech domain with a TM fitting [Mdn_{FF} = 4.9 vs Mdn_{TM} = 6; $z = -1.89$, $p < 0.05$]. For the other
 346 perceptual outcome measures, no significant differences were found [$z = -1.60$, $p = 0.13$ (phoneme
 347 perception test - 3000 Hz), $z = -0.56$, $p = 0.30$ (phoneme perception test - 6000 Hz), $z = -1.15$, $p = 0.16$
 348 (speech intelligibility in quiet), $z = -0.97$, $p = 0.19$ (speech intelligibility in noise)].

⁴Based on the averaged aided threshold for 3000, 5000 and 6000 Hz.

⁵Based on $N = 6$. For the first 4 participants, no significant differences were found across sessions for the averaged percentage correct score at 45, 50 and 65 dBA SPL [FirstFit: $z = 0.00$, $p = 0.56$; TargetMatch: $z = -0.37$, $p = 0.44$].

Discussion

To increase REM uptake by hearing healthcare professionals, fitting strategies in which REMs are integrated in the fitting software are becoming increasingly popular (Koehler & Kulkarni, 2014; Beck & Crowe, 2017). Reports on the reliability of integrated and automated REM-based fittings are scarce, and literature on improved perceptual outcomes, such as aided audibility and speech intelligibility, as well as on perceived real-life benefits with hearing aids being fitted based on REM as compared to a FF, is lacking. This study aimed to conduct a preliminary investigation into the reliability and matching accuracy of TM, a new automated fitting strategy integrated in Phonak fitting software in HI participants with a sensorineural hearing loss compared to a traditional REM-based approach with MF. Also, a fitting based on TM was compared to a FF approach with respect to perceptual outcomes and perceived benefits.

A counterbalanced single-blinded repeated-measures cross over design was used, allowing a direct comparison of the fitting strategies and perceptual outcomes within the same participant. Validated test materials were used, and perceptual tests were conducted with adaptive features of the hearing aid in default mode. Cognitive and linguistic confounds were minimized by careful participant selection. Only experienced hearing aid users were enrolled in the study, to rule out amplification acceptance issues. As such, we assessed the added value of conducting REMs with respect to perception, ruling out possible confounding influences.

Despite the careful study design, a major limitation of this study is its low sample size. Study findings are promising, but should be validated in a larger and more varied sample.

Reliability

Test-retest reliability was investigated for the REUG and the REAR. For the unaided response, test-retest reliability estimates for a single measurement below 3 dB are documented in the literature for

frequencies between 250 and 4000 Hz. Generally, test-retest reliability decreases with increasing frequency (Valente & Goebel, 1991), due to interactions between standing waves and probe tube insertion depth (Storey & Dillon, 2001; Mueller, 2001). Our data for the REUG are consistent with this literature. Test-retest errors below or around 1.5 dB up to 6300 Hz using a guided probe tube placement (TM) were obtained. Test-retest reliability with a constant-insertion depth method (MF) was slightly lower, with errors just exceeding 3 dB at 2000 and 5000 Hz. Overall, test-retest reliability for TM was better than for MF for the majority of frequencies measured. This is most probably the result of using the guided probe tube feature, indicating when adequate insertion depth has been reached.

More studies have looked into the test-retest reliability of the real-ear insertion gain across frequencies using REMs for different hearing aid styles and measurement systems. In general, differences of 2 dB can be considered as measurement variability (Jespersen & Møller, 2013). In accordance with the literature, test-retest reliability for the REAR observed in our study were better than or around 2 dB for the majority of frequencies, and almost never exceeded 2.5 dB (except for MF). Test-retest reliability for TM was best for frequencies between 1000 and 5000 Hz, and worse for very low or very high frequencies. Better test-retest reliability was obtained with TM compared to MF for all frequencies except 630 and 800 Hz, with test-retest differences for TM being generally smaller for frequencies between 1250 and 5000 Hz. Also test-retest reliability was worse for very low and very high frequencies, with test-retest differences exceeding 2.5 dB with MF for the lowest and very high frequencies. The poorer test-retest reliability with a MF might be related to the experience level of the fitter. The audiologist who fitted the HAs (first author) had no prior experience with fitting hearing aids based on REM.

Test-retest reliability for REMs is relatively understudied for frequencies above 4000-6000 Hz, with only recent studies examining the reliability for higher frequencies (Vaisberg et al, 2016; Jespersen & Møller, 2013). This study showed that test-retest reliability is somewhat worse for high frequencies. However, compared to a MF strategy, better values are obtained with integrated and automatic REMs.

This underlines the importance of adequate probe tube placement and the benefit of automated target matching, bypassing the experience level of the hearing healthcare professional.

Matching accuracy

The matching accuracy of TM was compared to mismatches to target with a FF as well as with a MF. Targets were generally matched within ± 5 dB for frequencies up to 5000 Hz. Because of individual differences in ear canal geometry, discrepancies between predicted (FF) and measured REAR (REM) emerge, especially for high frequencies (Hawkins & Cook, 2003; Aazh & Moore, 2007; Aazh et al, 2012), with possible underamplification of 10 dB and more. Mismatches to target with FF fittings reported in literature are the highest for frequencies between 3000 and 4000 Hz. This is partly related to the loss of ear canal resonance when placing an earmold in the ear (Aazh & Moore, 2007). Our data also show the largest target mismatch at 4000 Hz, using a FF. For the majority of participants, target mismatches exceed 5 dB for frequencies between 2500 and 4000 Hz, which are key to speech intelligibility (Pavlovic, 1987). Not only did FF fittings lead to significantly larger mismatches compared to TM, manually adjusted fitting based on REM also resulted in larger mismatches for frequencies between 800 and 5000 Hz. Mismatch values exceeded 5 dB for 4000 Hz and higher frequencies. Indeed, significantly lower REARs were measured for MF compared to TM, especially for higher frequencies (4000 and 5000 Hz). Again, the experience level of the hearing healthcare professional could play a role, especially since mismatches to target have been associated with the slope of the audiogram between 2000 and 4000 Hz, i.e. steeply sloped audiograms failing to reach fitting targets. Mixed results are found for the relation between target mismatches and hearing loss degree (Swan & Gatehouse, 1995; Aazh & Moore, 2007). As such, some ears might be more difficult to fit. Whereas mismatches to target are generally lower for TM compared to a FF and a MF, the differences only reach significance for frequencies up to 5000 Hz. For higher frequencies, there was no significant

difference between strategies in match to target. But also, at these higher frequencies, variability was larger. Part of this variability might be explained by the hearing loss degree of the patients, and as such the maximum amplification the devices could produce. Certainly, the gain controls were at their maximum setting in the high frequencies for some participants, which may have contributed to the high target mismatch across fitting strategies in this frequency range.

Perceptual outcomes

Given that more accurate target matching for higher frequencies is obtained, and given that prescription rules aim at maximizing intelligibility, improved perceptual outcomes with TM over a FF fitting were expected. Our results indicated an improved overall aided audibility with TM and significantly better thresholds at 5000 Hz. No differences were found at 3000 and 6000 Hz. The latter is consistent with target mismatches being equally large for both fitting strategies in the high frequency range. At 3000 and 5000 Hz, there were some floor effects. Only a small improvement in aided audibility was found, with differences between 1 and 2 dB, which are clinically insignificant. No better speech intelligibility in quiet and in noise was obtained with TM. Possibly, this can partly be explained by the small improvement observed in aided audibility. However, 3/10 participants obtained better SRTs with TM for SIN, with differences > 1 dB. The worst performers with FF also performed the worst with TM. It should be noted that the majority (8/10) of our HI participants had a severe high frequency hearing loss (> 60 dB HL). Desensitization, i.e. being less able to extract information from audible speech with increasing hearing loss, has been shown to drop the maximum of extractable information to about 60% of that of a person with hearing thresholds at 0 dB HL when hearing thresholds reach 60 dB HL (Ching & Dillon, 2013). With respect to real-life outcomes, a benefit for hearing speech in a variety of competing contexts was experienced with TM over a FF fitting, using the SSQ12. These results are in agreement with Abrams et al (2012) who found significantly higher mean scores for the APHAB communication subscales with a verified fitting compared to a FF. Whereas speech-in-noise

testing in the lab was done with speech-shaped noise, noise shapes encountered in daily life might be different relative to the speech spectrum. For instance, many real-life noise spectra have predominantly more low frequencies than speech or have a more pronounced roll-off in the high frequencies than speech (approximately 6 dB per octave), leading to higher SNRs at high frequencies. Thus, high frequency cues become more important and better high frequency audibility might be beneficial.

This study demonstrates improved test-retest reliability for an integrated REM-based fitting with automatic target matching, compared to FFs and REM-based fittings with manual adjustment. Targets are accurately matched for frequencies up to 5000 Hz. This results in improved high frequency audibility, as well as real-life benefits for speech intelligibility compared with FF. Better speech intelligibility with a fitting based on REM compared to a FF could be demonstrated based on the SSQ12 data, but could not be confirmed with a speech-in-noise test in the lab. Given that our group consisted of HI participants with severe high frequency hearing losses, possible effects of desensitization could explain the absence of such an improvement. Also, our sample size was rather low. No fine tuning was done up to this point, as this study aimed to investigate the added value of conducting REMs. Ensuring a match to prescribed amplification should merely be the start of the hearing aid fitting process. Possibly, not all patients prefer prescribed targets being accurately matched. Especially, new hearing aid users might have acceptance issues when the amount of amplification provided exceeds their preferred levels. Nonetheless, REMs ensure the right shape of the gain-frequency response (Aazh et al, 2012), possibly reducing the number of visits that is required to reach a best fit fitting.

Automated REMs that can be conducted from within the fitting software are a growing trend. Hearing aid manufacturers are hopeful that these methods meet the clinician's needs by overcoming some of the barriers of conducting REMs, such as time constraints. Whether the availability of these fitting strategies will lead to an increased use of REMs in hearing aid fitting practices remains to be seen.

478 Conclusion

479

480 This study demonstrates high test-retest reliability and matching accuracy of TM, an automated REM-
481 based fitting integrated in the fitting software. Compared to a traditional REM-based fitting with
482 manual fitting to target, REUGs and REARs can be obtained with increased reliability and targets are
483 more accurately matched. The use of a guided probe tube placement feature and bypassing the
484 experience level of the audiologist are plausible explanations. Confirming literature, relying on average
485 REUGs conducting a FF, results in large mismatches to target. However, at a group level, improved
486 speech intelligibility with TM compared to a FF was not observed. Yet, improved speech intelligibility
487 in daily life was experienced. A fitting based on TM improves high frequency audibility and is a good
488 starting point for further fine tuning by the hearing healthcare professional. Findings need to be
489 verified in a larger and more varied sample.

490 References

- 491 Aarts, N.L. & Caffee, C.S. 2005. Manufacturer predicted and measured REAR values in adult hearing
492 aid fitting: accuracy and clinical usefulness. *Int J Audiol*, 44(5), 293–301.
- 493 Aazh, H., Moore, B. & Prasher, D., 2012. The accuracy of matching target insertion gains with open-fit
494 hearing aids. *Am J Audiol*, 21, 175–181.
- 495 Aazh, H. & Moore, B.C.J., 2007. The value of routine real ear measurement of the gain of digital
496 hearing aids. *J Am Acad Audiol*, 664(8), 653–664.
- 497 Abrams, H.B., Chisolm, T.H., McManus, M. & McArdle, R., 2012. Initial-fit approach versus verified
498 prescription: comparing self-perceived hearing aid benefit. *J Am Acad Audiol*, 23(10), 768–778.
- 499 Alcantara, J.I., Moore, B.C.J. & Marriage, J. 2004. Comparison of three procedures for initial fitting of
500 compression hearing aids. *Int J Audiol*, 43(1), p.3–14.
- 501 Amlani, B.A.M., Pumford, J. & Gessling, E. 2016. Improving patient perception of clinical services
502 through real-ear measurements. *Hear Rev*, 23(12).
- 503 Baumfield, A. & Dillon, H. 2001. Factors affecting the use and perceived benefit of ITE and BTE
504 hearing aids. *Br J Audiol*, 35(4), 247–258.
- 505 Beck, D.L. & Crowe, N. 2017. Easy, fast, and accurate hearing aid fittings via an automated REM
506 system using IMC 2. *Hear Rev*, 24(4), 30–31.
- 507 Byrne, D. 1986. Effects of frequency response characteristics on speech discrimination and perceived
508 intelligibility and pleasantness of speech for hearing-impaired listeners. *J Acoust Soc Am*, 80(2),
509 494–504.
- 510 Ching, T.Y.C. & Dillon, H. 2013. A brief overview of factors affecting speech intelligibility of people
511 with hearing loss: implications for amplification. *Am J Audiol*, 22(2), 306–309.
- 512 Digiovanni, J.J. & Pratt, R.M. 2010. Verification of in situ thresholds and integrated real-ear
513 measurements. *J Am Acad Audiol*, 21(10), 663–670.
- 514 Fikret-Pasa, S. & Revit, L.J. 1992. Individualized correction factors in the preselection of hearing aids. *J*
515 *Speech Lang Hear Res*, 35(2), 384–400.
- 516 Francart, T., Hofmann, M., Vanthornhout, J., Van Deun, L., van Wieringen, A, et al. 2017. APEX/SPIN:
517 a free test platform to measure speech intelligibility. *Int J Audiol*, 56(2), 137–143.
- 518 Francart, T., van Wieringen, A. & Wouters, J. 2008. APEX 3: a multi-purpose test platform for auditory
519 psychophysical experiments. *J Neurosci Methods*, 172, 283–293.
- 520 Hawkins, B.D.B. & Cook, J.A. 2003. Hearing aid software predictive gain values: How accurate are
521 they? *Hear J*, 56(7), 26–34.
- 522 Holube, I., Fredelake, S., Vlaming, M. & Kollmeier, B. 2010. Development and analysis of an
523 International Speech Test Signal (ISTS). *Int J Audiol*, 49(12), 891–903.

- 524 Jespersen, C.T. & Møller, K.N. 2013. Reliability of real ear insertion gain in behind-the-ear hearing
525 aids with different coupling systems to the ear canal. *Int J Audiol*, 52(3), 169–76.
- 526 Johnson, E.E. 2013. Modern prescription theory and application: realistic expectations for speech
527 recognition with hearing aids. *Trends Hear*, 17(3), 143–170.
- 528 Keidser, G., Dillon, H., Flax, M., Ching, T. & Brewer, S. 2011. The NAL-NL2 prescription procedure.
529 *Audiol Res*, 1(e24), 88–90.
- 530 Kochkin, S., Beck, D.L., Christensen, L.A., Compton-Conley, C., Fligor, B.J., et al. 2010. MarkeTrak VIII:
531 the impact of the hearing healthcare professional on hearing aid user success. *Hear Rev*, 17(4),
532 12–34.
- 533 Koehler, E.D. & Kulkarni, S. 2014. Fast and easy fitting and verification with integrated real-ear
534 measurement. *Hear Rev*, 21(10), 36–40.
- 535 Margolis, R.H. & Saly, G.L. 2008. Asymmetric hearing loss: definition, validation, and prevalence. *Otol*
536 *Neurotol*, 29(4), 422–431.
- 537 Mueller, B.H.G. & Picou, E.M. 2010. Survey examines popularity of real-ear probe-microphone
538 measures. *Hear J*, 63(5), p.27–32.
- 539 Mueller, H.G. 2001. Probe microphone measurements: 20 years of progress. *Trends Ampl*, 5(2), 35–
540 68.
- 541 Nasreddine, Z.S., Phillips, N.A., Bedirian, V., Charbonneau, S., Whitehead, V., et al. 2005. The
542 Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. *J*
543 *Am Geriatr Soc*, 53(4), 695–699.
- 544 Naylor, G., Oberg, M., Wanstrom, G. & Lunner, T. 2015. Exploring the effects of the narrative
545 embodied in the hearing aid fitting process on treatment outcomes. *Ear Hear*, 36(5), 517–526.
- 546 Noble, W., Jensen, N.S., Naylor, G., Bhullar, N. & Akeroyd, M.A. 2013. A short form of the speech,
547 spatial and qualities of hearing scale suitable for clinical use: the SSQ12. *Int J Audiol*, 52(6), 409–
548 412.
- 549 Schmitt, N., Winkler, A., Boretzki, M. & Holube, I. 2016. A phoneme perception test method for high-
550 frequency hearing aid fitting. *J Am Acad Audiol*, 27(5), 367–379.
- 551 Scollie, S., Seewald, R., Cornelisse, L., Moodie, S., Bagatto, M., et al. 2005. The desired sensation level
552 multistage input / output algorithm. *Trends Ampl*, 9(4), 159–197.
- 553 Storey, L. & Dillon, H. 2001. Estimating the location of probe microphones relative to the tympanic
554 membrane. *J Am Acad Audiol*, 12(3), 150–154.
- 555 Swan, I.R. & Gatehouse, S. 1995. The value of routine in-the-ear measurement of hearing aid gain. *Br*
556 *J Audiol*, 29(5), 271–277.
- 557 Vaisberg, J.M., Macpherson, E.A. & Scollie, S.D. 2016. Extended bandwidth real-ear measurement
558 accuracy and repeatability to 10 kHz. *Int J Audiol*, 55(10), 580–586.

- 559 Valente, M. & Goebel, J. 1991. Reliability and intersubject variability of the real ear unaided
560 response. *Ear Hear*, 12(3), 216–20.
- 561 Van Wieringen, A. & Wouters, J. 2008. LIST and LINT: sentences and numbers for quantifying speech
562 understanding in severely impaired listeners for Flanders and the Netherlands. *Int J Audiol*,
563 47(6), 348–55.
- 564 Yanz, J.L., Pisa, J.F.D., Olson, L. & Prairie, E. 2007. Integrated REM: Real-Ear Measurement from a
565 hearing aid. *Hear Rev*, p.1–11.
- 566 Wouters, J., Damman, W. & Bosman, A.J. 1994. Vlaamse opname van woordenlijsten voor
567 spraakaudiometrie. *Logopedie*, 7, 28–34.

568 Tables

569

570 Table 1. Test-retest reliability for the Real-Ear Unaided Gain (REUG) and the Real-Ear Aided Response
 571 (REAR) per frequency (f) and fitting strategy (TM = TargetMatch, MF = Manual Fitting to Target).

f (Hz)	REUG (dB)		REAR (dB)	
	TM	MF	TM	MF
250	0.1	0.3	2.3	3.2
315	0.1	0.3	1.4	2.5
400	0.4	0.4	1.2	2.2
500	0.4	0.4	1.6	2.0
630	0.3	0.4	2.1	1.7
800	0.9	0.5	1.2	1.0
1000	0.3	0.6	1.0	1.3
1250	1.6	0.6	0.5	1.6
1600	1.1	2.1	0.2	1.4
2000	1.3	3.1	0.4	1.6
2500	0.4	1.2	0.6	2.2
3150	1.7	0.9	0.7	2.0
4000	0.9	1.1	0.8	2.2
5000	1.5	3.3	0.9	2.9
6300	1.4	1.9	2.3	2.6
8000	2.7	2.7	2.5	4.2

572 Test-retest reliability was calculated as the root mean square of the standard deviation of repeated
 573 measurements. Data were averaged across ears (REUG) and levels (REAR), and should therefore be
 574 multiplied by $\sqrt{2}$ (REUG) and $\sqrt{6}$ (REAR) to deduce the test-retest reliability for an individual
 575 measurement.

576

577 Table 2. Median differences and statistics for mismatches to target per frequency (f) for manual fitting
 578 to target (MF) and FirstFit (FF) versus TargetMatch (TM). Negative values indicate smaller mismatches
 579 for TM.

f (Hz)	Target Mismatch (dB)	
	MF vs TM	FF vs TM
250	5.1; z = -0.56, p = 0.31	0.1; z = -0.15, p = 0.46
315	2.0; z = -0.05, p = 0.50	-1.3; z = -2.40, p < 0.01
400	0.2; z = -1.48, p = 0.08	-1.8; z = -1.78, p < 0.05
500	-0.8; z = -1.68, p = 0.05	-2.4; z = -2.19, p < 0.05
630	-1.0; z = -1.53, p = 0.07	-1.6; z = -1.89, p < 0.05
800	-2.7; z = -2.70, p < 0.01	-1.3; z = -2.40, p < 0.01
1000	-4.4; z = -2.80, p < 0.01	-2.9; z = -2.09, p < 0.05
1250	-5.7; z = -2.80, p < 0.01	-3.5; z = -2.60, p < 0.01
1600	-2.2; z = -2.40, p < 0.01	-1.0; z = -1.27, p = 0.12
2000	-2.4; z = -1.78, p < 0.05	-3.5; z = -2.50, p < 0.01
2500	-1.5; z = -1.78, p < 0.05	-5.3; z = -2.40, p < 0.01
3150	-4.9; z = -2.80, p < 0.01	-4.8; z = -2.60, p < 0.01
4000	-8.1; z = -2.81, p < 0.01	-7.2; z = -2.81, p < 0.01
5000	-6.3; z = -2.70, p < 0.01	-2.7; z = -2.40, p < 0.01
6300	-3.0; z = -1.17, p = 0.14	1.3; z = -0.56, p = 0.31
8000	-3.9; z = -1.38, p = 0.10	2.7; z = -1.58, p = 0.07

580

581 Figures

582

583 Figure 1. Audiograms for left (upper panel) and right (lower panel) ears. Grey lines indicate air
 584 conduction thresholds for the individual participants. Black lines represent the average hearing
 585 thresholds as a function of frequency. Dotted lines represent the fitting range of the HAs used.

586

587 Figure 2. Schematic overview of the study protocol and outcome measures obtained at the different
 588 study visits for hearing impaired (HI) participants. MF = Manual Fitting to target, FF = FirstFit, TM =
 589 TargetMatch, REUG = Real-Ear Unaided Gain, REAR = Real-Ear Aided Response, PPT = Phoneme
 590 Perception Test, SIQ = Speech Intelligibility in Quiet, SIN = Speech Intelligibility in Noise, SSQ = Speech
 591 Spatial and Qualities questionnaire.

592

593 Figure 3. Mismatches to Target For Manual Fitting to Target compared to TargetMatch (DSL-v5) (upper

594 panel), and for FirstFit compared to TargetMatch (NAL-NL2) (lower panel). The dotted line indicates a

595 perfect match to target. Positive values are indicative of underamplification. Negative values indicate

596 overamplification. Individual data points (light grey) and medians are shown. Error bars represent the

597 IQR. Significant differences are indicated by asterisks.

598
 599 Figure 4. Aided detection thresholds (upper left), speech reception thresholds in quiet (SRTq, upper
 600 right), speech reception thresholds in noise (SRTn, lower left) and SSQ12 scores (lower right) for FirstFit
 601 compared to TargetMatch. Outcomes were averaged across sessions (before and after home trial).
 602 Individual data are plotted in grey. Medians and IQRs are also shown. Significant differences are
 603 indicated.