An Integrative Evaluation of the Efficacy of a Directional Microphone and Noise-Reduction Algorithm under Realistic Signal-to-Noise Ratios

Francis Kuk[1] Christopher Slugocki[1] Petri Korhonen[1]

[1] Widex Office of Research in Clinical Amplification (ORCA-USA), Lisle, IL

J Am Acad Audiol 2020;31:262–270.

Address for correspondence Francis Kuk, Widex Office of Research in Clinical Amplification (ORCA-USA), Lisle, IL 60532 (e-mail: fkuk@widex.com).

Abstract

Background Many studies on the efficacy of directional microphones (DIRMs) and noise-reduction (NR) algorithms were not conducted under realistic signal-to-noise ratio (SNR) conditions. A Repeat-Recall Test (RRT) was developed previously to partially address this issue.

Purpose This study evaluated whether the RRT could provide a more comprehensive understanding of the efficacy of a DIRM and NR algorithm under realistic SNRs. Possible interaction with listener working memory capacity (WMC) was assessed.

Research Design This study uses a double-blind, within-subject repeated measures design.

Study Sample Nineteen listeners with a moderate degree of hearing loss participated.

Data Collection and Analysis The RRT was administered with participants wearing the study hearing aids (HAs) under two microphones (omnidirectional versus directional) by two NR (on versus off) conditions. Speech was presented from 0° at 75 dB SPL and a continuous noise from 180° at SNRs of 0, 5, 10, and 15 dB. The order of SNR and HA conditions was counterbalanced across listeners. Each test condition was completed twice in two 2-hour sessions separated by one month.

Results The recall scores of listeners were used to group listeners into good and poor WMC groups. Analysis using linear mixed-effects models revealed significant effects of context, SNR, and microphone for all four measures (repeat, recall, listening effort, and tolerable time). NR was only significant on the listening effort scale in the DIRM mode at an SNR of 5 dB. Listeners with good WMC performed better on all measures of the RRT and benefitted more from context. Although DIRM benefitted listeners with good and poor WMC, the benefits differed by context and SNR.

Conclusions The RRT confirmed the efficacy of DIRM and NR on several outcome measures under realistic SNRs. It also highlighted interactions between WMC and sentence context on feature efficacy.

Keywords
- directional microphone
- listening effort
- noise reduction
- realistic signal-to-noise ratios
- Repeat-Recall Test

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Background

Speech intelligibility tests are often used to evaluate the efficacy of hearing aids (HAs) and/or their features, such as directional microphones (DIRMs) and noise reduction (NR). However, speech tests may not fully capture all the benefits. In addition, it may be more meaningful to quantify benefits under realistic signal-to-noise ratio (SNR) conditions (e.g., Smeds et al, 2015[41]). This study examined the feasibility of using the recently developed Repeat-Recall Test (RRT, Slugocki et al, 2018[39]) to capture performance with DIRMs and NR.

Copyright © 2020 by the American Academy of Audiology. All rights reserved. Thieme Medical Publishers, Inc., 333 Seventh Avenue, New York, NY 10001, USA. Tel: +1(212) 760-0888.

DOI https://doi.org/10.3766/jaaa.19009.
ISSN 1050-0545.

DIRMs in HAs have been available since the 1970s (Ricketts, 2001[35]). From single microphone designs with two ports to the multiple microphone designs of today, the laboratory efficacy of DIRMs has been demonstrated to range from a 1- to 2-dB improvement in SNR in the open-fit mode (Kuk et al, 2005a[17]) to ~6 dB in a closed-fit mode (Ricketts and Hornsby, 2006[36]). Other studies have shown that DIRMs also reduce listening effort (Holmes et al, 2018[14]) and improve the acceptable noise level (Freyaldenhoven et al, 2005[11]).

Data on the efficacy of NR algorithms are less clear. Chong and Jenstad (2018[7]) reviewed studies from 2000 to 2016 on the efficacy of commercially available single microphone NR algorithms on adults and children. Most of the studies failed to report any improvement in speech intelligibility. Rather, these studies report improved sound quality, reduced annoyance, increased listening comfort, reduction in effort, reduced pupil dilation, improved acceptable noise level, improved ability to learn novel stimuli, improved secondary visual tracking, improved word recall, and increased preference. These results suggest that present-day commercial NR algorithms may reduce the cognitive load on the listeners but may not improve SNR sufficiently to improve speech understanding. Ofnote, some non-real-time NR systems, such as the binary mask algorithm, reportedly show improvements in speech intelligibility in noise (e.g., Wang et al, 2009[45]).

One of the issues with laboratory studies is that many are conducted under test conditions that optimize the outcome of the HA evaluation. However, such test conditions may or may not represent the range of SNRs that listeners encounter in real life. Many studies test at listeners' individualized speech reception thresholds (SRTs) for 50% (e.g., Brons et al, 2014;[5] Desjardins, 2016;[9] Miller et al, 2017[25]) and/or 95% correct identification (e.g., Ng et al, 2013;[28] 2015;[29] Wendt et al, 2017[46]). Hence, actual SNRs cover a broad range at the group level. For example, for SRT50, Miller et al (2017)[25] reported mean SNRs around 0 dB (range from 0 to −1.57 dB) and Brons et al (2014)[5] reported a mean of 1.5 dB (range from 0 to 2.4 dB). Ng et al (2013;[28] 2015[29]) reported a mean SRT_{95} of 4.1 dB (standard deviation [SD] = 1.85 dB) and 7.5 dB (SD = 1.9 dB) in their 2013 and 2015 studies, respectively. Neher et al (2018)[27] used an SNR of 6 dB, whereas Desjardins and Doherty (2014)[10] used an SNR of 8 dB to optimize the likelihood of an observed benefit. Other studies optimize test conditions based on functional considerations of the feature under evaluation. For example, Wang et al (2009)[45] studied the efficacy of an ideal binary mask NR algorithm and reported a SRT ranging from −8 dB with a speech-shaped noise to −20 dB with a cafeteria noise. Magnusson et al (2013)[24] demonstrated that a DIRM (occluded and open fit) improved SNRs from around 0 dB to −10 or −12 dB. It is commonly known that directional benefits decrease as SNRs increase (e.g., Kuk et al, 1999;[16] Ricketts and Hornsby, 2006[36]).

Clearly, differences in technology or algorithm implementation impose unique requirements on the test design. However, individualization and/or optimization of test conditions leads one to question whether the efficacy observed in laboratory studies may be generalized to more realistic SNR conditions. Smeds et al (2015)[41] and Wu et al (2018)[48] reported that people with a mild-to-moderate degree of hearing loss tend to experience day-to-day communication at SNRs of ~5 and 10 dB. If this is true, then the results of some of the studies reported previously may occur infrequently in real life. Thus, a speech measure that includes a realistic range of SNRs may help streamline the evaluation of HA features for all patients. On the other hand, testing at realistic SNRs will likely result in performance ceilings and will decrease the sensitivity of the test to possible differences between HA signal processing conditions (e.g., Smeds and Wolters, 2017[40]). Speech materials that yield a shallower slope on the performance-intensity (P-I) functions and/or that prevent plateaus at realistic SNRs may be useful in overcoming these issues. A solution is the use of low-context (LC) speech materials. This could make correct identification more difficult and reduce the slope of the P-I function, making the test more sensitive to changes at the higher, more realistic SNRs (>5-10 dB). Indeed, in a previous study (Kuk et al, 2019[20]), we were able to demonstrate a difference in SRT at an 85% performance criterion between variable speed compression and fast/slow compression using LC sentences.

Recently, the concept of working memory capacity (WMC) has been introduced to explain a listener's speech-in-noise difficulties (Lunner, 2003;[21] Akeroyd, 2008;[1] Ronnberg et al, 2008;[37] Rudner et al, 2011;[38] Besser et al, 2013;[4] Pichora-Fuller et al, 2016[33]). WMC is defined as the collection of cognitive resources that individuals use to encode, store, and process information (Baddeley and Hitch, 1974[2]). Listeners with large WMCs may be able to allocate resources to processing degraded speech and still have spare capacity for storage. Conversely, listeners with limited WMCs may engage all of their cognitive resources to process speech in noise, leading to feelings of effortful listening and leaving less room for storage. Hence, individual variability in WMC might explain some of the variance observed in studies on the efficacy of specific HA features. Indeed, Gatehouse et al (2006)[12], Lunner and Sundewall-Thoren (2007)[23], and Souza et al (2015)[42] observed that listeners with poor WMCs performed better with slow-acting compression, whereas those with a larger WMC performed better with fast-acting compression. Ng et al (2013;[28] 2015[29]) and Lunner et al (2016)[22] observed that listeners with better WMCs recalled more speech in noise with a NR algorithm. Although there is no reason to believe that a DIRM selectively favors good or poor WMC listeners, a difference in benefit between the two groups may be revealed under conditions where the performance of one group has plateaued and that of the other has not. In those scenarios, people with poor WMC may continue to experience DIRM benefit because there is still room for improvement, whereas those with good WMC may not. A test that includes an estimate of WMC and varying levels of difficulty may provide evaluative (i.e., actual performance) and explanatory (i.e., why such performance) value.

The RRT (Slugocki et al, 2018[39]) assesses speech intelligibility and WMC at realistic SNRs (0, 5, 10, 15 dB, and quiet; Smeds et al, 2015;[41] Wu et al, 2018[48]) using high-context (HC) and LC sentences. During the test, listeners repeat a list of six sentences one at a time. The correct target words are scored. After all six sentences are repeated, listeners recall as many of the sentences (or fragments of the sentences) as they can. Afterward, listeners

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

rate how much effort they spent listening to the sentences (i.e., listening effort) and estimate the time they are willing to spend (i.e., tolerable time) communicating under the specific SNR condition.

To date, Slugocki et al (2018)[39] have determined the list equivalence of the speech materials. The P-I functions, test-retest reliability, and validity of the test on 20 normal hearing listeners and 16 hearing-impaired listeners were also determined. Repeat performance (SRT_{50}) correlates with the listeners' Hearing in Noise Test scores ($r = 0.40$, $p < 0.05$) and recall performance correlates with the listeners' scores ($p = 0.53$, < 0.01) on the Reading Span Test (Van den Noort et al, 2008[44]). Furthermore, intra-class correlation coefficients (ICC) indicate that a single administration of the RRT produced reliable measures of repeat (ICC = 0.83) and recall (ICC = 0.75) performance. Together, these results suggest that the integrated RRT produces a valid measure of speech-in-noise performance and that the recall scores may be used to assess a listener's WMC.

The purpose of this study was to reconfirm the efficacy of DIRMs and NR using the RRT. First, we wanted to evaluate whether the range of realistic SNRs included in the RRT is sufficient to demonstrate the efficacy of DIRMs and NR. If so, the results of the evaluation may offer some insights into the real-world effectiveness of these noise management algorithms. Second, we wanted to use the recall score on the RRT to examine potential interactions between WMC and efficacy of HA features. Previous studies had not examined the interaction between WMC and DIRM benefit. Third, we wanted to confirm that the RRT could capture differences in how DIRMs and NR affect behavioral measures of speech-in-noise processing. Previous research has shown that a DIRM results in changes in speech intelligibility and listening effort, whereas NR only results in changes in listening effort. Because the RRT uses the same set of stimuli and test conditions for all four outcome measures, it may offer a more comprehensive single-test approach to evaluating a listener's speech-in-noise difficulties.

Methods

Participants

Twenty hearing-impaired adults (mean age = 73.6 years, range = 56-86 years) were recruited from the local community. All participants were native speakers of American English. Two participants discontinued after the first session because of an illness in the family. A new participant was recruited after the second dropout. Thus, the data analyzed and reported here were from this final sample of 19 participants (8 females).

The four-frequency pure-tone average of the participants was 48.6 and 49.8 dB HL (SD = 3.6) for the left and right ears, respectively (►**Figure 1**). Two of the 19 participants never wore HAs, although both had participated in HA studies previously. Of the HA wearers, six wore Widex HAs of various models, four wore Phonak HAs, and the rest wore HAs from other manufacturers. All but one participant scored above 23 (out of 30) on the Montreal Cognitive Assessment (MoCA, Nasreddine et al, 2005[26]), which is considered as normal performance (Carson et al, 2018[6]). The data of the participant with a MoCA score of 22 were included in the analysis. The study was approved by an external independent institutional review board. Written informed consent was obtained from all participants before the study. Participants were financially compensated for their time.

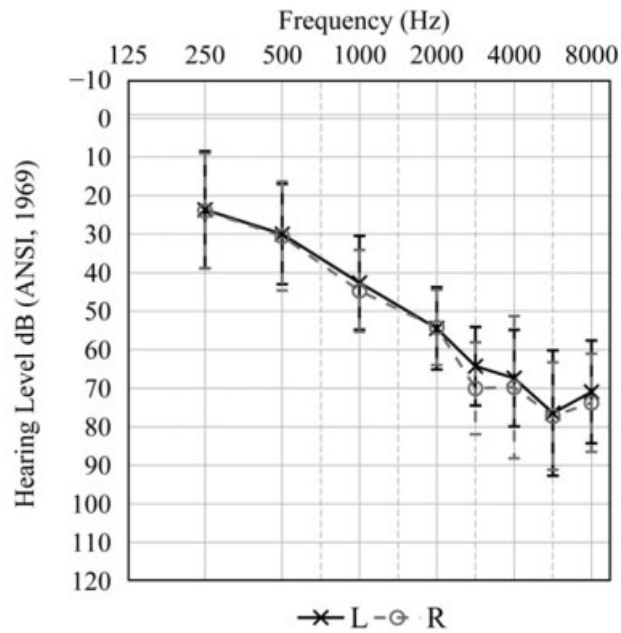

Fig. 1 Average pure tone thresholds for the left (black exes) and right (gray circles) ears of 19 hearing-impaired listeners. Error bars represent one SD.

Hearing Aid Conditions

Participants completed all testing in the aided mode with bilaterally fitted receiver-in-canal HAs. The study HA used a 15-channel automatic adaptive DIRM with speech preservation (Kuk et al, 2005b[18]). A fixed hypercardioid mode was used during the testing to eliminate any unpredictable change in polarity. The Speech Enhancer NR algorithm is a modulation-based algorithm that reshapes the frequency response in noise to optimize the speech intelligibility index. When activated, it provided a maximum gain reduction of 12 dB and a maximum gain increase of 4 dB in the mid-frequencies (Kuk and Paludan-Muller, 2006[19]). The study HAs were coupled to a receiver that had a peak output sound pressure level at 90 dB SPL input (OSPL90) 114 dB SPL as measured in a 2-cc coupler. All fittings used fully occluding "double-dome" instant-fit ear tips to minimize the influence of direct sounds mixing with the processed sounds. The target gain of the HAs was set based on the National Acoustics Laboratory-Nonlinear fitting formula version 2 (NAL-NL2) rationale (Keidser et al, 2011[15]). Output from the HAs was verified using the SoundTracker feature (Oeding and Valente, 2013[31]) on the HA fitting software to ensure audibility of the test materials in the reference condition (omnidirectional, no NR). Only one of the participants had previous experience with the study HA. Listener performance in noise was assessed at four combinations of microphone and NR conditions: omnidirectional microphone (OMNI) with NR

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

enabled (OMNI.NR.ON), OMNI with NR disabled (OMNI.NR.OFF), directional microphone with NR enabled (DIRM.NR.ON), and directional microphone with NR disabled (DIRM.NR.OFF).

Test Materials

The RRT (Slugocki et al, 2018[39]) drew speech materials from five sets of thematically related sentences. The themes included food and cooking, books and movies, music, shopping, and sports. Under each theme, seven lists of six sentences (in a list) were available so that a unique list was used for each SNR. Each sentence contained three to four target words (mostly nouns, adjectives, and verbs) so that 20 target words were scored for every list. All sentences were targeted at a fourth-grade reading level as measured by the Flesch-Kincaid reading level scale.

Semantic context has been documented as a cue for speech understanding in noise (e.g., Pichora-Fuller et al, 1995[34]; Obleser and Kotz, 2010[30]; Davis et al, 2011;[8] Zekveld et al, 2011;[49] Winn, 2016;[47] Holmes et al, 2018[14]). The RRT estimated context use by comparing listener performance for HC and LC sentences. HC sentences were meaningful sentences that were related to the same theme (or topic) such that listeners can draw upon within-sentence and between-sentence cues for word identity. LC sentences were generated by randomizing target words among the HC sentences in a list. This process resulted in six sentences that were syntactically valid but semantically meaningless (both within-sentence and between-sentence). This process also ensured that the word difficulty and long-term spectra of HC and LC materials were similar.

Procedure

The study followed a double-blind within-subjects design. Participants completed two 2-hour sessions at the Widex ORCA-USA office. All testing took place in a double-walled sound-treated booth (Industrial Acoustics, Bronx, NY; internal dimensions: $3 \times 3 \times 2$ m, W × L × H). At the beginning and end of each visit, participants' thresholds at 500 Hz and 4000 Hz were measured to ensure no change in threshold occurred.

On qualifying for the study, participants' behavioral speech-in-noise abilities were assessed using the RRT. A unique sentence set (i.e., sports, shopping, music, and books and movies) was used to assess each HA condition. Listeners were instructed on the RRT using a standardized script. A practice RRT trial was administered using a dedicated LC passage presented at 75 dB SPL at an SNR of +10 dB. Testing was then carried out in blocks, where each block assessed both LC and HC passages across all SNRs for a given HA condition. To minimize any carry-over effects from semantically meaningful sentences, testing always began with LC passages. The order of HA program blocks and SNRs within a block was counterbalanced across listeners. The HAs were programmed by another staff member who did not participate in data collection. At no time was the participant or the tester aware of which HA features were enabled on the study aids. Each list of six sentences took about two to three minutes to complete. The whole RRT (LC/HC sentence lists at four SNRs) was completed within 20–25 minutes for a single HA condition.

Speech stimuli were delivered in the free-field at a fixed level of 75 dB SPL via a KRK ST6 loudspeaker (KRK systems, Nashville, TN) (±2 dB from 62 Hz to 20 kHz) driven by the output of a Rotel 1048 power amplifier (Rotel, North Reading, MA). The amplifier received input from a Shure Auxpander line mixer that routed channel output from an Echo Audio Gina 24 (Echo, Santa Barbara, CA) sound card. The speech loudspeaker was positioned at a distance of 1 m directly in front (i.e., 0°) of the participant. The center of the loudspeaker driver was 107 cm above the floor. A spectrally matched, continuous speech-shaped noise was presented from a second KRK ST6 loudspeaker, driven by a different channel on the same equipment, located at a distance of 1 m directly behind the listener (i.e., 180°). Background noise was presented at fixed levels to produce SNRs of 0, 5, 10, and 15 dB. A sound level meter (Quest Technologies Model 1800; TSI incorporated, Shoreview, MN) was used for daily calibration of the stimulus levels. Visual prompts used in the RRT (to alert listeners to respond) were presented on a touchscreen computer monitor (17" Planar PT 1700 MU; Planar, Beijing, China) placed on a small table directly in front of the participant at a 45° downward angle in the median plane. The position of the monitor did not obstruct a direct line between the loudspeaker and the listener's ears.

All test participants returned in about a month for a retest on the RRT. Testing followed the same procedure outlined for the first visit but with a new counterbalancing order. Before analysis, performance metrics (i.e., repeat and recall) and subjective ratings (i.e., listening effort and tolerable time) from the RRT were averaged across tests and retests for each combination of SNR and HA condition.

Results

To group listeners based on WMC, we first examined the distribution of repeat scores for HC speech materials to determine which test condition showed perfect or near-perfect repeat performance. Repeat performance at SNR = +15 dB in the DIRM.NR.OFF condition was at or above 97.5% for all participants. Based on previous research (Slugocki et al, 2018[39]), this level of performance satisfied the requirements of adequate audibility while requiring some effort from the listeners (average listening effort ratings = 4). At this test condition, recall scores ranged from 28% to 65%, with the median at 43%. Hence, participants with recall performance ≥43% were placed into the "good" WMC group and those with recall performance <43% were placed into the "poor" WMC group. There were ten participants in the good WMC group and nine in the poor WMC group. Good and poor WMC groups were similar in mean age (73 years ± 9.4 SD versus 74 years ± 7.6 SD), four-frequency pure-tone averages (47 dB HL ± 6.5 SD versus 51 dB HL ± 10.5 SD), and MoCA scores (27 ± 2.2 SD versus 26 ± 2.1 SD). It should be noted that here we used good versus poor WMC in a relative sense. Other measures of WMC may result in different groupings and different outcomes.

The lme4 package (Bates et al, 2015[3]) for R was used to compute linear mixed-effects models that assessed the fixed effects of microphone (OMNI versus DIRM), NR(NR On versus

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

NR Off), passage context (HC versus LC), SNR (0, 5, and 10), and group (good WMC versus poor WMC) on each of the RRT outcome measures. Whereas the P-I functions displayed subsequently included scores at SNR +15 dB, this SNR was excluded from all statistical analyses because recall performance at this SNR was used to group subjects and because of potential ceiling effects. Unique slopes were modeled as random effects across SNRs for each participant. Before statistical analysis, listeners' repeat and recall scores were transformed into rationalized arcsine units according to the method defined in Studebaker (1985)[43]. Visual inspection of residual plots did not reveal any obvious deviations from normality. ***p*** values were obtained by Wald tests (Type IISS) of linear hypotheses using the Chi-square statistic. Only significant factors and interactions were reported in the corresponding text describing the P-I functions of the repeat, recall, listening effort, and tolerable time in good and poor WMC groups for HC and LC speech materials.

Repeat Performance

Repeat performance was measured as the number of correct target words repeated after each sentence. ►**Figure 2** summarizes repeat performance over the range of SNRs for both groups of participants. Repeat scores increased as the SNR increased [$\chi^2_{(2)} = 484.62$, $p < 0.000$] and plateaued at z10 dB for the HC materials. Repeat scores for the HC sentences were higher than those for the LC sentences [$\chi^2_{(1)} = 249.68$, $p < 0.000$] and were higher for the DIRM mode than for the OMNI mode [$\chi^2_{(1)} = 1{,}406.28$, $p < 0.000$]. It was higher by over 50 percentage points at the SNR of 0 dB condition. This is equivalent to about 6.5 dB improvement in SNR when estimated at the 75% correct level. Participants in the good WMC group outperformed those in the poor WMC group [$\chi^2_{(1)} = 3.9$, $p = 0.048$]. A microphone × SNR interaction confirmed that the slope of the P-I function with the DIRM was shallower than that of the OMNI condition [$\chi^2_{(2)} = 118.76$, $p < 0.000$]. This occurred because repeat performance was less sensitive to SNR changes at or above 5 dB in the DIRM compared with the OMNI mode. A context × SNR interaction [$\chi^2_{(2)} = 23.12$, $p < 0.000$] occurred because the effect of context was small at the poorest SNRs of 0 and 5 dB. Last, there was a three-way interaction of microphone × context × SNR [$\chi^2_{(2)} = 10.57$, $p = 0.005$] reflecting a different behavior between OMNI and DIRMs with context at the poorest SNRs. With DIRM processing, the repeat P-I functions for HC and LC passages were similar, albeit offset (HC > LC). With OMNI processing, repeat performance did not differ between HC and LC passages at SNR = 0 dB, presumably because of a floor effect. Above that SNR, repeat performance increased by a greater amount with SNR for HC than for LC sentences. It was also noted that the DIRM benefit (DIRM minus OMNI repeat scores) at SNR = 15 dB differed between good WMC and poor WMC listeners, especially between HC and LC materials. Because performance at SNR = 15 dB was used to group listeners, data at that SNR were not included into the statistical model for test of significance. Any observed positive effects of NR (2–6% improvement) were not significant ($p > 0.05$).

Recall Performance

Recall performance was measured as the number of correctly recalled target words that were also correctly repeated. ►**Figure 3** shows the P-I functions. Recall performance of the good WMC group was higher than that of the poor WMC group [$\chi^2_{(1)} = 17.45$, $p < 0.000$]. With DIRM processing, recall

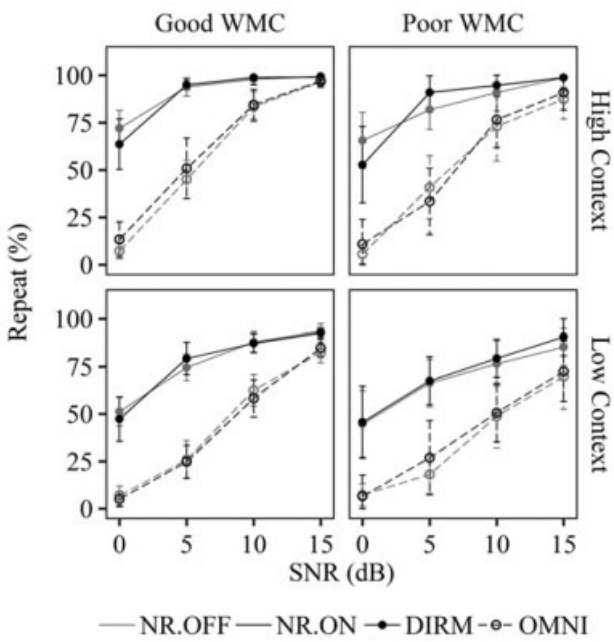

Fig. 2 P-I functions of repeat performance for HC (top panels) and LC (bottom panels) passages in good (left) and poor (right) WMC listeners. Data are shown for directional (solid lines) and omnidirectional (dashed lines) microphones with NR enabled (black) and disabled (gray). Error bars represent 95% confidence intervals of the mean.

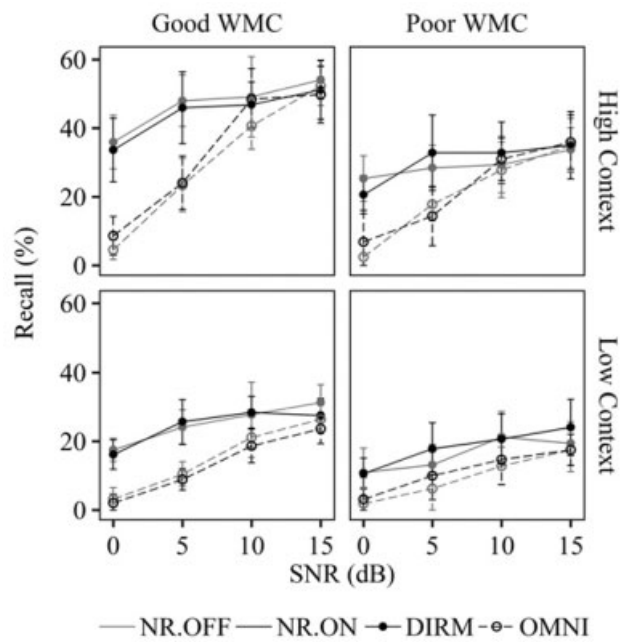

Fig. 3 P-I functions of recall performance for HC (top panels) and LC (bottom panels) passages in good (left) and poor (right) WMC listeners. Data are shown for directional (solid lines) and omnidirectional (dashed lines) microphones with NR enabled (black) and disabled (gray). Error bars represent 95% confidence intervals of the mean.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

plateaued at SNRs of 5 and 10 dB for HC and LC passages, respectively. It was about 10–15 dB for the OMNI condition with the HC materials. In addition, there were significant effects of microphone [$\chi^2_{(1)} = 468.27$, $p < 0.000$], context [$\chi^2_{(1)} = 411.72$, $p < 0.000$], and SNR [$\chi^2_{(2)} = 204.45$, $p < 0.000$]. A group × microphone interaction [$\chi^2_{(1)} = 11.25$, $p < 0.001$] confirmed the greater difference in recall between the DIRM and OMNI conditions in the good WMC group than in the poor WMC group. In addition, a context × group effect reflected a greater difference in recall between HC and LC materials in the good WMC group than in the poor WMC group [$\chi^2_{(1)} = 14.03$, $p < 0.000$]. Microphone, context, and SNR interacted with each other in two-way (microphone × context [$\chi^2_{(1)} = 5.0$, $p = 0.025$]; microphone × SNR [$\chi^2_{(2)} = 122.93$, $p < 0.000$]; and context × SNR [$\chi^2_{(2)} = 18.19$, $p < 0.000$]) and three-way (microphone × context × SNR [$\chi^2_{(2)}) = 22.6$, $p < 0.000$]) interactions. These interactions occurred when context improved performance in the DIRM but not the OMNI mode at SNR = 0 dB condition. Again, NR did not result in any significant differences in recall performance.

Ratings of Listening Effort

Ratings of listening effort and tolerable time were provided after all six sentences were repeated and recalled. ►**Figure 4** summarizes the changes in reported listening effort with SNRs. Listening effort decreased with increasing SNRs [$\chi^2_{(2)} = 191.35$, $p < 0.000$] and was generally lower for the DIRM than for the OMNI microphone condition [$\chi^2_{(1)} = 480.75$, $p < 0.000$]. In addition, HC materials were generally rated as less effortful than LC materials [$\chi^2_{(2)} = 218.29$, $p < 0.000$]. Even at SNR = +15 dB, participants still rated the test conditions to be somewhat effortful (i.e., >4). A group × context interaction [$\chi^2_{(1)} = 11.88$, $p < 0.001$] reflected that the good WMC group rated HC materials as less effortful and LC materials as more effortful than did the poor WMC group. A significant microphone × NR × SNR interaction [$\chi^2_{(2)} = 8.2$, $p < 0.017$] occurred because NR reduced listening effort at SNR = 5 dB when used with the DIRM but not for other SNRs or when used with the OMNI. The benefit of NR also appeared to be stronger in listeners with poor WMC, although this trend was not significant. A microphone × context interaction [$\chi^2_{(1)} = 3.86$, $p = 0.049$] occurred when the DIRM was associated with lower ratings oflistening effort relative to OMNI, but this difference was larger for HC than for LC materials. A context × SNR interaction [$\chi^2_{(2)} = 27.54$, $p < 0.000$] reflected a greater decrease in listening effort for HC than for LC materials with increasing SNR. A three-way microphone × context × SNR interaction [$\chi^2_{(2)} = 8.32$, $p = 0.016$] reflected a constant effect of context in the DIRM mode but not in the OMNI mode at SNR = 0 dB.

Estimates of Tolerable Time

Estimates of tolerable time were transformed into log units before display and statistical analysis. Such a transformation made the visual display of tolerable time (►**Figure 5**) more closely resemble those of the other measures. Tolerable time significantly increased with SNR [$\chi^2_{(2)} = 83.7$, $p < 0.000$] and was longer in the DIRM than in the OMNI mode [$\chi^2_{(1)} = 328.45$, $p < 0.000$] by about 15 minutes. A significant two-way SNR × microphone interaction [$\chi^2_{(2)} = 75.94$, $p < 0.000$] occurred wherein the benefit of the DIRM decreased with increasing SNR. Tolerable time was also longer for HC than for LC materials

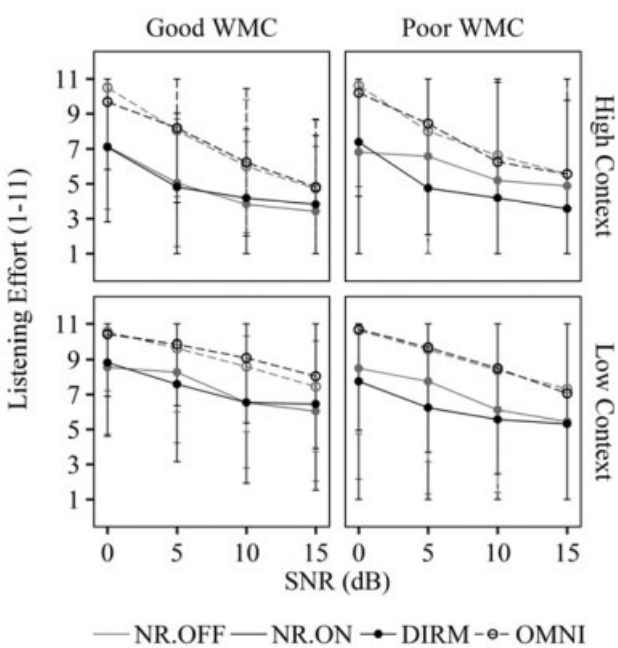

Fig. 4 P-I functions of listening effort for HC (top panels) and LC (bottom panels) passages in good (left) and poor (right) WMC listeners. Data are shown for directional (solid lines) and omnidirectional (dashed lines) microphones with NR enabled (black) and disabled (gray). Error bars represent 95% confidence intervals of the mean.

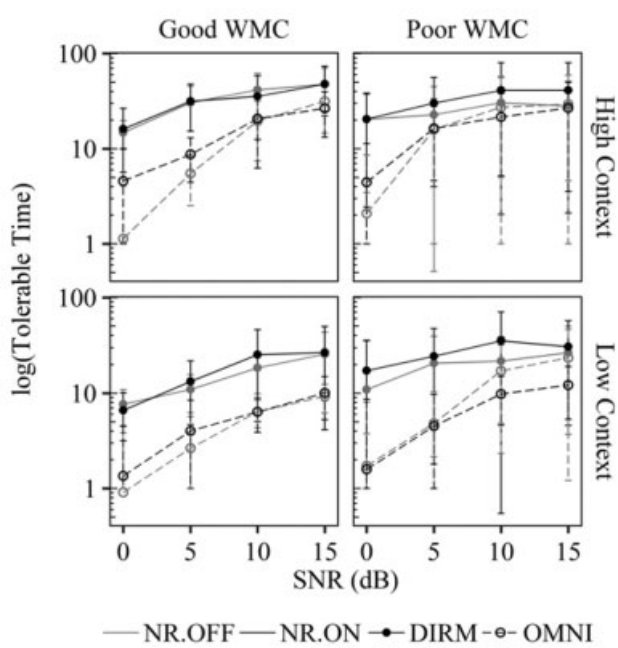

Fig. 5 P-I functions of log-transformed tolerable time for HC (top panels) and LC (bottom panels) passages in good (left) and poor (right) WMC listeners. Data are shown for directional (solid lines) and omnidirectional (dashed lines) microphones withNRenabled (black) and disabled (gray). Error bars represent 95% confidence intervals of the mean.

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

[$\chi^2_{(1)} = 80.0$, $p < 0.000$]. The main effects of context and microphone were further qualified by significant two-way context × group [$\chi^2_{(1)} = 6.0$, $p = 0.014$] and microphone × group interactions [$\chi^2_{(1)} = 10.11$, $p < 0.001$]. Listeners with good WMC exhibited a greater difference in tolerable time with DIRM over OMNI modes, and with HC over LC materials than did listeners with poor WMC. The effect of NR was not statistically significant.

Discussion

The current study reaffirmed the efficacy of DIRM and NR algorithms on the RRT. The range of realistic SNRs used in the RRT is sufficient to capture some of the multidimensional effects of these HA features in listeners evaluated in this study. Furthermore, recall scores were useful in grouping participants. This grouping revealed that listeners with good WMC performed better on all measures of the RRT and benefitted more from context than those with poor WMC. They also differed in how they benefitted from a DIRM.

The benefits of DIRMs permeated all RRT outcome measures and decreased with increasing SNRs. For the repeat task, both groups of listeners benefitted from DIRMs to a similar degree for HC sentences at low to moderate SNRs. However, differences between groups were noted for the LC sentences at SNR = 15 dB. ►**Figure 2** shows that the performance of listeners with good WMC plateaued at SNR = 15 dB, offering no room for improvement on repeat scores even if the SNR is improved. On the other hand, listeners with poor WMC only scored 65%, suggesting that there is room for improvement if available. Thus, the improvement in SNR from DIRMs resulted in as much as 20% DIRM benefit for the poor WMC (and not in the good WMC) listeners with the LC materials.

►**Figure 3** shows that the improvement in recall of LC materials associated with DIRM (over OMNI) processing was greater for the good WMC group than for the poor WMC group. One interpretation is that the poor WMC group was unable to use the SNR improvement provided by the DIRM to help with recall of the LC materials. In other words, the improvement in SNR brought by the DIRM was not sufficient to improve recall to the same magnitude as that of the good WMC group or that for the HC materials. That is, ensuring similar repeat scores (from better SNR) is a necessary but not a sufficient condition for proper recall. The WMC of the individuals and contextual cues are also important determiners. These two observations advance our understanding of the benefits of DIRMs in that contextual cues, the WMC of the listeners, the SNR of the environment, and the outcome measures used (such as repeat or recall tasks) also affect the expression of the DIRM benefits.

The current study confirmed previous reports that NR algorithms improve subjective listening effort (e.g., Holmes et al, 2018[14]). When in the DIRM mode, NR lowered ratings of listening effort, most significantly at SNR = 5 dB. Hoetink et al (2009)[13] suggest that the efficacy of NR is dependent on the input SNR. At SNR = 0 dB, the improvement by NR may not be perceptible because of a potential floor effect. As the SNR increases to 10 or 15 dB, the amount of gain reduction from NR decreases, thus reducing the contrasts between NR states. We ***speculate*** that the subjective improvement associated with activation of NR, such as captured by ratings of listening effort, likely results from an internal comparison between changes in cognitive load from the NR and cognitive resource allocation for the task. If this ratio is large (i.e., large reduction in cognitive load compared with small cognitive resource allocation), listeners may notice a subjective improvement; otherwise, no change in perception is likely. In the OMNI mode, listening is effortful for both groups of listeners; thus, more cognitive resource is required. Because the improvement in cognitive load from NR may be small, it would be a small percentage of the total cognitive resources that the listeners need to spend on the task. Thus, no appreciable improvement with NR is reported. In the DIRM mode, listening is not as effortful as in the OMNI mode, thus necessitating a relatively smaller amount of cognitive resource. As such, the same small decrease in cognitive load from NR, when compared with the smaller size of the allocated cognitive resource, would result in a larger ratio and lead to the perception of less effort. Because people in the poor WMC group have less cognitive resources to allocate than those in the good WMC, the same decrease in cognitive load from NR would make the effect even more pronounced in the poor WMC group than in the good WMC group.

These observations suggest that it could be beneficial to test over a range of SNRs instead of testing at one fixed SNR. On the other hand, if a fixed SNR were to be used to examine the benefit of DIRMs and NR algorithm, this study suggests that an SNR of 5 dB may be the most optimal because this is the only condition where statistically significant NR and DIRM effects were seen. This SNR is similar to the mean SNR used in several studies that evaluated the efficacy of a NR algorithm (e.g., Ng et al, 2013;[28] Brons et al, 2014;[5] Neher et al, 2018;[27] Ohlenforst et al, 2018[32]).

Determining the WMC using the RRT under a speech-in-noise condition where speech intelligibility is near perfect may help explain the communication difficulties of listeners. In this study, listeners with good WMC, as compared with those with poorer WMC, have significantly higher repeat and recall performance, report less listening effort and longer tolerable time, are able to use more context cues (resulting in a higher HC score), and show different patterns of benefits from DIRMs, at least under some conditions. Conversely, listeners in the poor WMC group are less able to take advantage of the available cues from signal processing (e.g., DIRMs) or from context to help improve their listening experience even though they may have a greater need to do so.

Abbreviations

DIRM	directional microphone
HA	hearing aid
HC	high context
ICC	intra-class correlation coefficients
LC	low context
MoCA	Montreal Cognitive Assessment
NR	noise reduction
OMNI	omnidirectional microphone

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

P-I	performance-intensity
RRT	Repeat-Recall Test
SD	standard deviation
SNR	signal-to-noise ratio
SRT	speech reception threshold
WMC	working memory capacity

Conclusions

These findings suggest that the RRT may provide a framework for demonstrating the multidimensional benefits of HA features under more realistic SNRs. Using the RRT, a DIRM provides significant benefit on all four outcome measures. On the other hand, the use of NR algorithm is not likely to improve speech intelligibility but may reduce listening effort in some noisy conditions. The extent of the benefit varies with the WMC of the listener, the SNR of the listening condition, and the amount of context provided by the speech materials.

Notes
All authors are employees of Widex A/S.

References

1 Akeroyd M. Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. Int J Audiol 2008;47(2, Suppl):53–71
2 Baddeley A, Hitch G. Working memory. In: Bower GH, ed. The Psychology of Learning and Motivation: Advances in Research and Theory. vol. 8. London, UK: Academic Press; 1974:47–89
3 Bates D, Maechler M, Bolker B, Walker S. Fitting linear mixed-effects models using lme4. J Stat Software 2015 67:1–48
4 Besser J, Koelewijn T, Zekveld A, Kramer S, Festen J. How linguistic closure and verbal working memory relate to speech understanding in noise - a review. Trends Amplif 2013 17(02):75–93
5 Brons I, Houben R, Dreschler W. Effects of noise reduction on speech intelligibility, perceived listening effort, and personal preference in hearing-impaired listeners. Trends Hear 2014 18:1–10
6 Carson N, Leach L, Murphy K. A re-examination of Montreal Cognitive Assessment (MoCA) cutoff scores. Int J Geriatr Psychiatry 2018 33(02):379–388
7 Chong F, Jenstad L. A critical review of hearing aid single microphone noise reduction studies in adults and children. Disabil Rehab Assist Technol 2018 13(06):600–608
8 Davis M, Ford M, Kherif F, Johnsrude I. Does semantic context benefit speech understanding through "top-down" processes? Evidence from time-resolved sparse fMRI. J Cognit Neurosci 2011;23(12):3914–3932
9 Desjardins J. The effects of hearing aid directional microphone and noise reduction processing on listening effort in olderadults with hearing loss. J Am Acad Audiol 2016 27(01):29–41
10 Desjardins J, Doherty K. The effect of hearing aid noise reduction on listening effort in hearing-impaired adults. Ear Hear 2014;35 (06):600–610
11 Freyaldenhoven M, Nabelek A, Burchfield S, Thelin J. Acceptable noise level as a measure of directional hearing aid benefit. J Am Acad Audiol 2005 16:228–236
12 Gatehouse S, Naylor G, Elberling C. Linear and nonlinear hearing aid fittings-patterns of candidature. Int J Audiol 2006 45:153–171
13 Hoetink A, Korossy L, Dreschler W. Classification of steady state gain reduction produced by amplitude modulation based noise reduction in digital hearing aids. Int J Audiol 2009;48(07): 444–455
14 Holmes E, Folkeard P, Johnsrude I, Scollie S. Semantic context improves speech intelligibility and reduces listening effort for listeners with hearing impairment. Int J Audiol 2018;57(07): 483–492
15 Keidser G, Dillon H, Flax M, Ching T, Carmen S. The NAL-NL2 prescription procedure. Audiol Res 2011;1(01):e24
16 Kuk F, Kollofski C, Brown S, Melim A, Rosenthal A. Use of a digital hearing aid with directional microphones in school-aged children. J Am Acad Audiol 1999 10:535–548
17 Kuk F, Keenan D, Sonne M, Ludvigsen C. Efficacy of an open-fitting hearing aid. Hear Rev 2005a 12(02):26–32
18 Kuk F, Peeters H, Keenan D, Baekgaard L. Timing is (almost) everything - fully adaptive directional microphone. Hear Rev 2005b12(08):24–29
19 Kuk F, Paludan-Muller C. Noise management algorithm may improve speech intelligibility in noise. Hear J 2006 59(04): 62–65
20 Kuk F, Slugocki C, Korhonen P, Seper E, Hau O. Evaluation of the efficacy of a dual variable-speed compressor over a single fixed speed compressor. J Am Acad Audiol 2019 30(07):590–606
21 Lunner T. Cognitive function in relation to hearing aid use. Int J Audiol 2003 42(1, suppl):S49–S58
22 Lunner T, Rudner M, Rosenbom T, Agren J, Ng E. Using speech recall in hearing aid fitting and outcome evaluation under ecological test conditions. Ear Hear 2016;37(1, suppl): 145S–154S
23 Lunner T Sundewall-Thoren E. Interactions between cognition, compression, and listening conditions: effects on speech-innoise performance in a two-channel hearing aid. J Am Acad Audiol 2007 18(07):604–617
24 Magnusson L, Claesson A, Persson M, Tengstrand T. Speech recognition in noise using bilateral open-fit hearing aids: the limited benefit of directional microphones and noise reduction. Int J Audiol 2013 52(01):29–36
25 Miller C, Bentler R, Wu Y, Lewis J, Tremblay K. Output signal-to-noise ratio and speech perception in noise: effects of algorithm. Int J Audiol 2017 256(08):568–579
26 Nasreddine Z, Phillips N, Bedirian V, Charbonneau S, Whitehead V, Collin I, Cummings JL, Chertkow H. The Montreal Cognitive Assessment, MoCA: a brief screening tool for mild cognitive impairment. J Am Geriatr Soc 2005 53(04):695–699
27 Neher T, Wagener K, Fischer R-L. Hearing aid noise suppression and working memory function. Int JAudiol 2018;57(05):335–344
28 Ng E, Rudner M, Lunner T, Pedersen M, Ronnberg J. Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. Int J Audiol 2013 52(07):433–441
29 Ng E, Rudner M, Lunner T, Ronnberg J. Noise reduction improves memory for target language speech in competing native but not foreign language speech. Ear Hear 2015 36:82–91
30 Obleser J, Kotz S. Expectancy constraints in degraded speech modulate the language comprehension network. Cereb Cortex 2010 20(03):633–640
31 Oeding K, Valente M. Differences in sensation level between the Widex SoundTracker and two real-ear analyzers. J Am Acad Audiol 2013 24(08):660–670
32 Ohlenforst B, Wendt D, Kramer S, Naylor G, Zekveld A, Lunner T. Impact of SNR, masker type and noise reduction processing on sentence recognition performance and listening effort as indicated by the pupil dilation response. Hear Res 2018 365:90–99
33 Pichora-Fuller K, Kramer S, Eckert M, Edwards B, Hornsby B, Humes L, Lemke U, Lunner T, Matthen M, Mackersie C, Naylor G, Phillips N, Richter M, Rudner M, Sommers M, Tremblay K,

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.

Wingfield A. Hearing impairment and cognitive energy: the framework for understanding effortful listening (FUEL). Ear Hear 2016 37(1, suppl):S5–S27
34 Pichora-Fuller K, Schneider B, Daneman M. How young and old adults listen to and remember speech in noise. J Acoust Soc Am 1995 97(01):593–608
35 Ricketts T. Directional hearing aids. Trends Amplif 2001; 5:139–176
36 Ricketts T, Hornsby B. Directional hearing aid benefit in listeners with severe hearing loss. Int J Audiol 2006 45:190–197
37 Ronnberg J, Rudner M, Foo C, Lunner T. Cognition counts: a working memory system for Ease of Language Understanding (ELU). Int J Audiol 2008 47(2, suppl):S99–S105
38 Rudner M, Ronnberg J, Lunner T. Working memory supports listening in noise for persons with hearing impairment. J Am Acad Audiol 2011 22(03):156–167
39 Slugocki C, Kuk F, Korhonen P. Development and clinical applications of the ORCA repeat and recall test (RRT). Hear Rev 2018;25 (12):22–26
40 Smeds K, Wolters F. Towards a firm grip on auditory reality. Hear Rev 2017 24(12):20–25
41 Smeds K, Wolters F, Rung M. Estimation of signal-to-noise ratios in realistic sound scenarios. J Am Acad Audiol 2015 26(02): 183–196
42 Souza P, Arehart K, Neher T. Working memory and hearing aid processing: literature findings, future directions, and clinical applications. Front Psychol 2015 6:1–12
43 Studebaker G. A rationalized arcsine transform. J Speech Hear Res 1985 28:455–462
44 Van den Noort M, Bosch P, Haverkort M, Hugdahl K. A standard computerized version of the Reading Span Test in different languages. Eur J Psychol Assess 2008 24(01):35–42
45 Wang D, Kjems U, Pedersen M, Boldt J, Lunner T. Speech intelligibility in background noise with ideal binary time-frequency masking. J Acoust Soc Am 2009 125(04):236–2347
46 Wendt D, Hietkamp RK, Lunner T. Impact of noise and noise reduction on processing effort: a pupillometry study. Ear Hear 2017 38(06):690–700
47 Winn M. Rapid release from listening effort resulting from semantic context, and effects of spectral degradation and cochlear implants. Trends Hear 2016 20:1–17
48 Wu Y Stangl E, Chipara O Hasan S, Welhaven A, Oleson J. Characteristics of real world signal-to-noise ratios and speech listening situations of older adults with mild to moderate hearing loss. Ear Hear 2018 39(02):293–304
49 Zekveld A, Rudner M, Johnsrude I, Festen J, van Beek J, Ronnberg J. The influence of semantically related and unrelated text cues on the intelligibility of sentences in noise. Ear Hear 2011;32(06):e16-e25

This document was downloaded for personal use only. Unauthorized distribution is strictly prohibited.