Using Speech Recall in Hearing Aid Fitting and Outcome Evaluation Under Ecological Test Conditions

Thomas Lunner,[1,2] Mary Rudner,[1] Tove Rosenbom,[3] Jessica Ågren,[3] and Elaine Hoi Ning Ng[1]

In adaptive Speech Reception Threshold (SRT) tests used in the audiological clinic, speech is presented at signal to noise ratios (SNRs) that are lower than those generally encountered in real-life communication situations. At higher, ecologically valid SNRs, however, SRTs are insensitive to changes in hearing aid signal processing that may be of benefit to listeners who are hard of hearing. Previous studies conducted in Swedish using the Sentence-final Word Identification and Recall test (SWIR) have indicated that at such SNRs, the ability to recall spoken words may be a more informative measure. In the present study, a Danish version of SWIR, known as the Sentence-final Word Identification and Recall Test in a New Language (SWIRL) was introduced and evaluated in two experiments. The objective of experiment 1 was to determine if the Swedish results demonstrating benefit from noise reduction signal processing for hearing aid wearers could be replicated in 25 Danish participants with mild to moderate symmetrical sensorineural hearing loss. The objective of experiment 2 was to compare direct-drive and skin-drive transmission in 16 Danish users of bone-anchored hearing aids with conductive hearing loss or mixed sensorineural and conductive hearing loss. In experiment 1, performance on SWIRL improved when hearing aid noise reduction was used, replicating the Swedish results and generalizing them across languages. In experiment 2, performance on SWIRL was better for direct-drive compared with skin-drive transmission conditions. These findings indicate that spoken word recall can be used to identify benefits from hearing aid signal processing at ecologically valid, positive SNRs where SRTs are insensitive.

Key words: Bone-anchored hearing aids, Competing speech, Free recall, Noise reduction, Working memory.

(Ear & Hearing 2016;37;145S–154S)

INTRODUCTION

Hearing aids are intended to help reduce the activity limitations and participation restrictions encountered in daily listening environments by individuals who are hard of hearing (e.g., World Health Organization 2001; Danermark et al. 2013). To evaluate hearing aids under realistic, ecologically valid conditions, there is a need for methods that take into account listening conditions that are encountered in daily life, especially where listening effort may be high. In this introduction, we first review the listening conditions that are encountered in daily life and contrast these to the conditions under which typical clinical behavioral measures are administered to illustrate a potential problem with current clinical methods. We then present a review of measures that have the potential to be used to assess listening effort under conditions more similar to those encountered in daily life.

Listening Conditions in the Lab and in Daily Life

Behavioral measures such as speech reception thresholds (SRTs) that are administered in conditions with competing noise (e.g., Plomp & Mimpen 1979; Nilsson et al. 1994; Hagerman & Kinnefors 1995; Akeroyd et al. 2015) are currently used to compare different hearing aid signal processing schemes or settings. Adaptive SRT testing in noise (e.g., Hagerman & Kinnefors 1995; Brand & Kollmeier 2002) has been applied in hearing aid research, hearing aid development, and the clinical evaluation of hearing aid outcomes. In the SRT test, sequences of prerecorded spoken sentences are presented at a fixed level and background noise is adapted according to how well the participant is able to repeat the words in the sentences. Typically, an SRT is obtained using an adaptive procedure to reach the signal to noise ratio (SNR) at which 50% of the words are correctly repeated. The 50% point is chosen because that is where the slope of the psychometric function is steepest and thus most sensitive to changes caused by hearing aids. For people with mild to moderate hearing impairment who are tested with hearing aids, the aided SRT for 50% word recognition accuracy typically ranges from −10 to +5 dB SNR, depending on speech corpus (e.g., open- or closed-set sentences, e.g., Lunner et al. 2012; high- or low-predictability, e.g., Pichora-Fuller et al. 1995), gender of the target speech (Helfer & Freyman 2008), scoring method (word-by-word scoring or full sentence scoring, Boothroyd & Nittrouer 1988), type and number of background maskers (Rosen et al. 2013), the spatial location of maskers (Neher et al. 2009), and the individual's performance.* Thus, adaptive SRT-in-noise methods that aim for a fixed performance level (e.g., 50% correct) are by definition not aiming for successful perception, as the performance criterion must be lower than 100% to be able to converge on the predetermined performance level.

Smeds et al. (2015) reported that typical daily listening environments have SNRs that range from +5 to +15 dB SNR. Hearing aids are typically evaluated using adaptive SRTs, and such tests typically result in negative SNRs for 50% correct performance. Furthermore, the noise reduction (NR) schemes in hearing aids that are designed to reduce background noise and enhance perception of the speech target are typically most effective at positive SNRs (e.g., Brons et al. 2013; Smeds et al. 2015). Thus, if hearing aids are to be evaluated under more ecologically valid daily life conditions, methods other than SRTs in noise are needed.

Haverkamp (2015) used an ecological momentary assessment smartphone application to allow participants who were hard of hearing to rate speech intelligibility and listening effort in a number of real-life situations that they found important for communication. At the same time, the application monitored the sound levels in those situations. The results showed that the important communication situations were sometimes those in

---

[1] Department of Behavioural Sciences and Learning, Linnaeus Centre HEAD, Swedish Institute for Disability Research, Linköping University, Linköping, Sweden; [2] Eriksholm Research Centre, Oticon A/S, Snekkersten, Denmark; and [3] Oticon Medical A/S, Smørum, Denmark.

* It should however be noted that monosyllabic word-in-noise tests may show SNRs higher than +5 dB when tested unaided in co-located target-and masker conditions (Wilson 2003).

0196/0202/2016/37Supplement 1-145S/0 • Ear & Hearing • Copyright © 2016 Wolters Kluwer Health, Inc. All rights reserved • Printed in the U.S.A.

<zdoi; 10.1097/AUD.0000000000000294>

which listening effort was rated as being high. However, the scale of listening effort was more broadly exploited by the participants than the scale of speech intelligibility; speech intelligibility was typically rated as good or excellent (Haverkamp et al. 2015). That is, while the listening effort of the participants in everyday listening situations varied substantially, speech was almost always considered intelligible, indicating that, even when people can hear what is being said, different amounts of listening effort may still be required depending on other aspects of the situation.

Taken together, these findings (Smeds et al. 2015; Haverkamp et al. 2015) indicate that ecological listening conditions for hearing aid users are characterized by (a) successful performance at or near ceiling (here defined as better than 95% correct word recognition accuracy), and (b) positive SNRs of +5 dB and higher. Therefore, adaptive SRTs yield measures of performance for SNR conditions that are not within the SNR range that is typical in most real-life situations. It may seem contradictory that persons who are hard of hearing report problems in daily life that occur in listening conditions where word recognition accuracy is almost perfect, so these reports must stem from difficulties that are not reflected in measures of word recognition accuracy, for example, the high listening effort required in such situations (consensus paper, Pichora-Fuller et al. 2016, this issue, pp. 5S–27S).

Measuring Listening Effort as an Alternative to SRTs in Noise

Increasingly, hearing researchers and clinicians are becoming more interested in the concept of listening effort and the potential usefulness of measuring it for hearing aid evaluation (Gosselin & Gagné 2010; McGarrigle et al. 2014; Rudner & Lunner 2014) and to assess activity limitations and participation restrictions (e.g., Kramer et al. 2006). However, theoretical understanding, the definition and clinical meaning of listening effort and methods for measuring it remain in development (McGarrigle et al. 2014; Rudner & Lunner 2014). McGarrigle et al. categorized methods of measuring listening effort according to type: subjective listening effort ratings, behavioral measures of listening effort and physiological measures of listening effort, pointing out that they most likely reflect different aspects of effortful listening and should not be considered to be different measures of the same phenomenon.

Importantly, in SNR conditions in which listeners who are hard of hearing would have high accuracy scores on word recognition tests, recall can be measured to study listening effort in ecological conditions. When auditory input is degraded, recognizing words may consume attentional or working memory resources that might otherwise have been allocated to higher-level processing, such as comprehension of the meaning of the words. For example, in an early study, Rabbitt (1991) used a test paradigm in which listeners shadowed (repeated) each word in lists of 12 nouns. After 50 lists, the participants were asked to recall as many items as possible in any order. The results indicated that the listeners who were hard of hearing recalled fewer words from lists in which all words had been repeated correctly than did listeners with normal hearing. In an earlier study, Rabbitt (1968) compared recall of digits in early list positions, when digits in subsequent list positions were presented in noise, and in quiet. Digits in early positions were less well remembered when digits in later list positions had to be discriminated through noise, indicating cognitive resources were being diverted from storage to listening in noise.

McCoy et al. (2005) showed that under successful word recognition conditions, participants with poor hearing recalled significantly fewer of the final three items in lists of spoken words of unpredictable length than did the participants with better hearing. The results were taken as support for the notion that the extra effort a listener who is hard of hearing must expend to achieve successful word recognition comes at the cost of processing resources that might otherwise be available to encode the semantic content in memory. In addition, Baldwin and Ash (2011) showed the importance of audibility for recall of spoken words. They showed that lower presentation levels had a detrimental effect on memory recall in a listening span task.

Specifically, compared with a listener with normal hearing, the extra effort that a listener who is hard of hearing must expend to achieve successful word recognition could be inferred if they recalled fewer of the words that had been recognized. Consistent with this research, the Word Auditory Recognition and Recall Measure (Smith et al. 2015 under review; see also Smith & Pichora-Fuller 2015) has been developed recently for use by audiologists to test recall using well-known clinical word recognition testing materials in American English. Preliminary tests indicate the potential clinical feasibility of the test insofar as when the Word Auditory Recognition and Recall Measure was administered in quiet and word recognition accuracy was at or near ceiling, older listeners with hearing loss had poorer recall than peers with normal hearing.

The research showing that recall can be affected by the quality of the input signal or hearing loss suggests that recall might improve if hearing aid signal processing can reduce the extra effort the hearing aid user must expend for successful listening, as suggested by Pichora-Fuller and Singh (2006). Accordingly, a recall measure could also be used to compare the effects of different hearing aid settings or different hearing aid signal processing schemes on the allocation of resources or the effort expended during listening. One such method was successfully implemented by Sarampalis et al. (2009), who showed that recall was greater for participants when a NR scheme was used. However, that study was not performed in listeners who were hard of hearing nor in conditions that met the criteria for successful word recognition.

Ng et al. (2013, 2015) used a memory recall paradigm referred to as the sentence-final word identification and recall (SWIR) that was inspired by the studies of Pichora-Fuller et al. (1995) and Sarampalis et al. (2009) to evaluate the effect of NR on higher-level processing of speech in participants who were hard of hearing (see also Rudner 2016, this issue, pp. 69S–76S). The Ng et al. (2013) study was conducted in conditions in which listeners achieved 85% correct word recognition (i.e., slightly below ceiling), and the sequel study (Ng et al. 2015) was conducted in more favorable conditions in which listeners achieved 95% correct word recognition. The test paradigm involved auditory presentation of sets of Swedish Hearing in Noise Test (HINT) sentences (Hällgren et al. 2005) in background babble. In the first study (Ng et al. 2013), sets of eight sentences were presented, but this was reduced to seven sentences in the follow-up study (Ng et al. 2015) to avoid floor effects in participants with poorer cognitive skills. In the first study (Ng et al. 2013), the participants were asked to repeat the final word of each sentence as it was presented to verify that the performance criterion was met. In the second study, we established that word repetition had no significant effect on memory performance. Participants were asked to repeat the words in only half of the lists and to recall as many of the final words as possible in any order after each set of sentences. The results confirmed that recall of Swedish words heard in competing Swedish speech improved when NR was used by hearing aid users.

In summary, studies employing measures of memory recall have been used when word recognition is nearly perfect (successful perception criterion) and at high SNRs (ecological validity criterion). Such measures of recall offer a promising way to evaluate benefit from hearing aids under ecological conditions. In the present article, we report on two studies that evaluated hearing aids and bone-anchored hearing solutions with a method of testing memory recall based on the studies of Ng et al. (2013, 2015), but using materials translated into a new language (Danish). The Danish test paradigm is called SWIRL, an acronym of the Sentence-final Word Identification and Recall test in a new Language. It is reasonable to assume that the Swedish test of recall of spoken words would generalize to another Scandinavian language, but it is important to validate a new test and to replicate findings in other labs and under varied conditions. The first experiment compared memory recall under conditions in which binary masking NR was turned on and off, thereby helping us to understand more about how recall of spoken words can be used to quantify the effects of hearing instrument signal processing under conditions more akin to those under which the instruments are intended to operate.

EXPERIMENT 1: SWIRL TESTING OF BINARY MASKING NOISE REDUCTION

The goal was to repeat a Danish counterpart to the Swedish SWIR test described by Ng et al. (2015) in approximately the same number of listeners to compare this test method across languages. The experimental contrast was the same as that investigated by Ng et al. (2013, 2015), with the hypothesis being that recall in the Danish SWIRL test would be greater with aided binary masking signal processing compared with aided signal processing alone as had been found in the original Swedish test.

Materials and Methods

- Participants

Twenty-five native Danish speakers (11 women and 14 men) with symmetrical moderate to moderately severe acquired sensorineural hearing loss were recruited from the Eriksholm list of participants. Their average age was 70 years (SD = 7.7, range: 54 to 79 years), and their average pure-tone threshold (PTA) at 0.5, 1, 2, and 4 kHz, averaged across ears, was 49.7 dB HL (SD = 9.9 dB HL). Figure 1 shows the hearing loss configuration for each of the participants. All were experienced hearing aid users with at least 1 year of hearing aid use. All participants were fitted with hearing aids bilaterally. No history of otological problems or psychological disorders was reported.

- Reading Span Test

The Reading Span test (Daneman & Carpenter 1980; Rönnberg et al. 1989) was administered as a measure of working memory. Reading span is a working-memory test designed to tax memory storage and processing simultaneously (Daneman & Carpenter 1980). The subjects' task was to comprehend sentences and to recall either the first or the final words of a presented sequence of sentences. The words were presented in a word-by-word fashion, at a rate of one word per 0.80 sec. Half of the sentences were absurd (e.g., "The train sang a song"), and half were normal (e.g., "The girl brushed her teeth"). The subjects' task was to respond "yes" verbally (for a normal sentence) and "no" verbally (for an absurd sentence) during a 1.75-sec interval after each sentence. After a sequence of sentences (three, four, five, or six sentences in ascending order), one of the words "first" or "final" was displayed on the screen, indicating that the subjects should start to recall either the first or the final words of all the three, four, five, or six sentences in the sequence. The order ("first" or "final") was randomized. The scoring procedure was different to the original Daneman and Carpenter procedure who measured the maximum reading span. In this study set sizes of three, four, five, and six sentences were presented, and repeated three times summing up to a maximum 54 correctly recalled words. The reading span result was calculated as the percentage of the maximum number of recalled items.

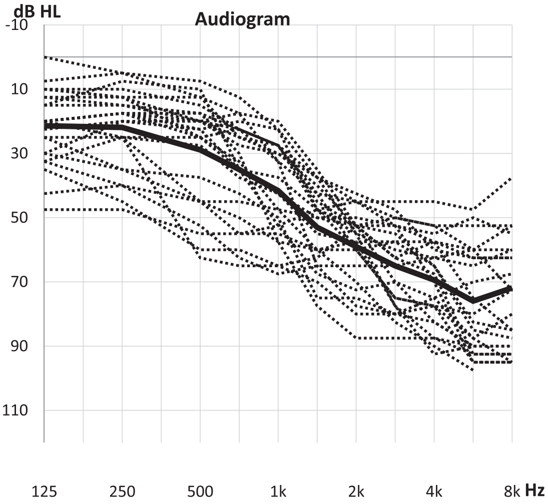


Fig. 1. Individual air conduction pure-tone hearing thresholds for the 25 participants with sensorineural hearing losses (*dotted lines*). The *solid line* indicates average hearing threshold across participants.

- SWIRL

Danish HINT sentences (Nielsen & Dau 2010) ending in two- or three-syllable words were used as stimuli. As in the study of Ng et al. (2015), each list consisted of seven sentences, and stimuli were presented at favorable SNRs estimated for each individual to produce 95% word recognition (SRT95) accuracy in competing four-talker babble in Danish (see Fig. 2). There were two tasks in the SWIRL test: an identification task (repeat the final word after listening to each sentence) and a free recall task, which was administered after all seven sentences had been presented. Specifically, participants were instructed to recall, in any order, as many as possible of the seven final words that had been previously repeated.

- Speech babble

The Danish speech babble consisted of recordings of two male and two female native speakers reading different paragraphs of a newspaper text. The duration of the recording of each speaker was approximately 3 min. The four-talker babble was postfiltered to resemble the long-term average spectrum of the HINT sentences. The speech babble was introduced 3 sec before the onset of the sentence and was terminated 1 sec after sentence offset. For each sentence, different portions of the babble were used.

- Noise reduction

A binary masking NR algorithm was used in this study. This signal processing algorithm reduces the masking effect of interfering speech noise by removing noise-dominant spectro-temporal regions in the speech-in-noise mixture (Wang et al. 2009). A 64-channel gammatone filterbank followed by time-windowing was applied to speech-in-noise mixtures to form time-frequency units. For each time-frequency unit in the binary mask, when the local SNR of any time-frequency unit is less than 0 dB (i.e., the energy of the noise exceeds the energy of the target speech) that unit is reduced by 10 dB. Otherwise, the unit is retained in the binary mask. This procedure is performed to optimize the SNR gain with the binary masks (Li & Wang 2009). More simply, the algorithm attenuates time-frequency bins where the SNR is low and passes that signal in time-frequency bins where the SNR is high. There were two NR conditions in this study: (a) binary masking NR (see Boldt et al. 2008 for details), which is the nonideal estimation of NR referred to here as NR-on; and (b) unprocessed, which is referred to as NR-off.

- Procedure

Each participant completed two 2-hr sessions. Audiometric measurements and the Reading Span test were conducted in the first session. In the second session, an individualized SRT95 was obtained for each participant using the NR-off condition before the administration of the SWIRL test, and this SRT95 was applied to all test conditions in the SWIRL test. The SRT95 for the NR-off condition was used as baseline to ensure that word accuracy was at ceiling in all conditions to prevent any further improvement in word accuracy by the binary masking algorithm. However, it is difficult to reliably estimate an SRT95 *words* correct using an adaptive procedure because the adaptation criteria require many correct words and few incorrect words, and this procedure requires testing of many words from a limited HINT corpus. It is therefore usually better to use a lower adaptation target, here, 84% *sentences* correct, according to the recommendations for performing the HINT test (Nilsson et al. 1994; Nielsen & Dau 2010). An SRT was obtained at 84% sentences correct, using the Hearing-In-Noise Test with a modified adaptive procedure (4-up-1-down; Levitt 1971). From this 84% sentence correct target, a procedure was developed to reach 95% words correct. Table 1 summarizes the steps in the procedure. The SNR for 95% words correct was found during SWIRL training, where the noise was turned down by 1 dB if 4 or 5 words were correctly repeated and by 2 dB if 0 to 3 words were correctly repeated. The noise level was left unchanged if six or seven words were correctly repeated. In the SWIRL test, the individually defined masker level was fixed. Four practice sentence lists were administered in the SWIRL training. The NR-off condition was considered to be the most difficult in terms of performance. To ensure successful perception (ca. 95% word recognition accuracy), the NR-off condition was used for HINT training, HINT testing, and SWIRL training. In the SWIRL test, condition of NR-off and NR-on was randomized across the five repetitions of the seven-sentence lists. The participants were blind to test condition.

All stimuli were preprocessed using MATLAB (2012). Auditory signals were presented using a 24-bit external PC soundcard at a sampling rate of 22.05 kHz and transmitted to the microphone of an Oticon Epoq XW behind-the-ear hearing aid (programmed to an individualized linear gain) in a small anechoic chamber. An IEC711 ear simulator was coupled to the receiver of the hearing aid, and the auditory signals from the hearing aid were transmitted to a pair of ER3A insert earphones through an equalizer and a measuring amplifier. The tests were conducted in a single-walled sound booth. Details on the setup and preparation of the test material are described elsewhere (Vatti 2013).

- Scoring method

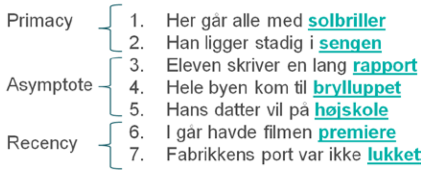
Performance on the word repetition task was scored as the percentage of responses that were correctly repeated. Performance on the free recall task was scored as the percentage of correctly recalled words (including words that included correctly recalled misperceptions). Scoring was performed online and all responses were audio-recorded to allow for retrospective checking of the scoring. The list position of the recalled words was also analyzed; the primary, asymptote, and recency list positions corresponded to the 1st to 2nd, 3rd to 5th, and 6th to 7th items, respectively, in each list. It is generally assumed that items in the recency position are held in short-term memory (working memory), while items in the primacy position have been transferred to long-term memory (see Rönnberg et al. 2013; Fig. 2).

TABLE 1. Steps in conducting the SWIRL test in experiment 1 (NR-off vs. NR-on)

Test	Target Sentence Level	Masker Level	Masker: Fixed/Adaptive	Adapt Toward % Discrimination	Number of Lists	Number of Sentences	Condition
HINT training	70 dB SPL	Starts at 64 dB SPL	Adaptive	84% sentences correct	1 list	1 × 20 sentences	NR-off
HINT test	70 dB SPL	Starts at 64 dB SPL	Adaptive	84% sentences correct	2 lists	2 × 20 sentences	NR-off
SWIRL training	70 dB SPL	Start at best (lowest) SNR from HINT test	Adaptive using scoring rules	95% words correct	4 lists	4 × 7 sentences	NR-off
SWIRL test	70 dB SPL	SNR reached in SWIRL training	Fixed	95% words correct	10 lists	2 × 5 × 7 sentences	Randomization NR-off/NR-on

HINT training was followed by the HINT test to obtain an SRT in noise at 84% sentences correct. Thereafter followed SWIRL training to adapt toward 95% words correct. The scoring rules included that the noise was turned down by 1 dB if 4 or 5 words were correctly repeated and by 2 dB if 0 to 3 words were correctly repeated. The noise level was left unchanged if six or seven words were correctly repeated. In the SWIRL test, the individually defined masker level was fixed.

NR, noise reduction; SNR, signal to noise ratio; SRT, speech reception threshold; SWIRL, Sentence-final Word Identification and Recall Test in a New Language.



Primacy
1. Her går alle med **solbriller**
2. Han ligger stadig i **sengen**

Asymptote
3. Eleven skriver en lang **rapport**
4. Hele byen kom til **brylluppet**
5. Hans datter vil på **højskole**

Recency
6. I går havde filmen **premiere**
7. Fabrikkens port var ikke **lukket**

Fig. 2. Example of a SWIRL sentence list in which the list positions (primacy, asymptote and recency) are illustrated. (The first Danish sentence "Her går alle I solbriller" has the English translation "All people wear sunglasses here.") SWIRL indicates Sentence-final Word Identification and Recall Test in a New Language.

- Statistical analysis and a priori hypotheses

Based on the previous finding that a significant main effect on recall of NR being on or off could be observed with a sample size of 26 (Ng et al. 2015, who showed a large effect size ($\eta_p^2 = 0.57$) of NR-on versus NR-off), an effect size of 0.55 was expected for the 25 participants in the present study. General linear model repeated measures analysis of variance (ANOVA) with planned contrasts was used to analyze these data except where noted. For all analyses, any p value of 0.05 or lower was considered statistically significant. The a priori hypotheses, based on the previous findings of Ng et al. (2013, 2015), were as follows: we expected a main effect of NR-on versus NR-off with higher recall of spoken words for the NR-on condition, and we expected an interaction effect with serial position such that recall of spoken words for items in the recency position would be higher in the NR-on condition than in the NR-off condition.

Results

- Reading Span test

The average reading span score was 42.6% (23 words correct out of a maximum of 54) with a standard deviation 9.2% and a range of 29.6% to 68.5%. The results were in line with previously reported Reading Span results for similar populations (e.g., Lunner 2003).

- Individualized SNR for 95% words correct

The procedure for finding the individual SNR for 95% words correct (Table 1) resulted in a mean individualized SNR of 9.6 dB (SD = 2.3), ranging from 6 to 15 dB. Thus, the SWIRL test was conducted in the positive SNR range.

- SWIRL recall

As seen in Table 2, recall performance was lower for the asymptote condition and was better for the recency and primacy conditions. Outlier analysis indicated no outliers. Variables from the present study were checked for normality using the Kolmogorov–Smirnov test and Lilliefors test for normality and the Shapiro-Wilk W test for normality. All variables could be considered to be normally distributed except for the data for the recency items in the NR-on condition; the condition of normality was not satisfied for these data (K–S, $d = 0.23$, $p < 0.15$; Lilliefors, $p = 0.01$). However, it was determined that rationalized arcsine unit-transformation (Studebaker 1985) of all variables, which would compensate for the eventual ceiling effect, did not affect the bimodal distribution and that rationalized arcsine unit transformation would also make numeral interpretations less straightforward. Thus, all variables were maintained in the percentage (fraction) format. The ANOVA analysis revealed the expected main effects of NR-on/off [$F(1,24) = 21.7$, $p < 0.0001$] and interaction effects between NR-on/off and serial position [$F(2,48) = 49.3$, $p < 0.0001$].

- Effects of processing and planned comparisons

Table 3 and Figure 3 show the results of the planned comparisons performed as paired t tests. From Table 3 in the first two rows, it can be seen that the average recall of spoken words was higher for NR-on compared with NR-off, and this description was confirmed by the paired t tests [$t(24) = 4.4$; $p < 0.001$].

Table 3 in rows three to six, as well as Figure 3, demonstrate that recall of spoken words for items in the recency position was better in the NR-on condition compared with the NR-off condition [$t(24) = 2.2$; $p < 0.05$]. The same was true for items in the primacy position [$t(24) = 2.9$; $p < 0.01$], but no significant difference was seen between NR-on and NR-off for items in the asymptote position.

- Predicting recall of spoken words

A multiple regression analysis was performed to investigate whether the four-frequency PTA (average hearing thresholds at 0.25, 0.5, 1, and 2 kHz of both ears) and the Reading Span score could be used to predict the average recall of spoken

TABLE 2. Means and standard deviations for the SWIRL free recall task obtained in the present study and in Ng et al. (2015) for the two processing conditions, binary masking noise reduction (NR-on) and unprocessed (denoted as NR-off in the present study and NoP in Ng et al.)

Study	Processing	Total	Primacy	Asymptote	Recency
Present	NR-on				
	M	0.58	0.60	0.41	0.81
	SD	0.13	0.23	0.17	0.16
	NR-off				
	M	0.50	0.51	0.36	0.72
	SD	0.13	0.23	0.17	0.14
Ng et al. (2015)	NR-on				
	M	0.62	0.61	0.45	0.79
	SD	0.18	0.21	0.17	0.15
	NoP (NR-off)				
	M	0.54	0.52	0.43	0.68
	SD	0.17	0.22	0.13	0.16

The values are presented as percent/100.
NR, noise reduction; SWIRL, Sentence-final Word Identification and Recall Test in a New Language.

TABLE 3. Planned comparisons between free recall performance in the two processing conditions: binary masking noise reduction (NR-on) and unprocessed (NR-off)

Variable	Mean	Std. Dv.	N	Diff.	Std. Dv. Diff.	*T*	Df	*p* (one-tailed)
NR-on average	0.58	0.13						
NR-off average	0.50	0.13	25	0.08	0.09	4.43	24	**0.0001**
NR-on primacy	0.60	0.23						
NR-off primacy	0.51	0.23	25	0.09	0.21	2.15	24	**0.02**
NR-on asymptote	0.41	0.17						
NR-off asymptote	0.36	0.17	25	0.05	0.15	1.52	24	0.07
NR-on recency	0.81	0.16						
NR-off recency	0.72	0.14	25	0.09	0.16	2.91	24	**0.004**

Significant p values (<0.05) are shown in bold. The units of the mean and standard deviation are percent/100.
NR, noise reduction.

words score for the NR-on condition. Only the Reading Span score had a significant partial correlation coefficient ($\beta = 0.49$; $p < 0.01$), indicating that working memory, but not the average PTA, was significantly associated with the recall of spoken words score in the NR-on condition. The regression model explained approximately 24% of the variance [$F(2,22) = 4.7$; Adjusted $R^2 = 0.236$; $p < 0.05$].

Discussion

The results of experiment 1 showed that the free recall of native language speech heard against a background of native language babble for Danish speakers with hearing impairment improves with the implementation of NR-on in hearing aids, at an SNR where perception is successful and occurs at the positive SNRs (individual SNRs from 6 to 15 dB) at which NR systems operate most efficiently. In general, the average recall score was very close to that reported by Ng et al. (2015) for comparable conditions; Table 2 shows that the differences in recall of spoken words between the two studies was only a few percentage points (fractions) in all NR conditions and serial positions. Thus, we replicated the results of Ng et al. in a new language. This result suggests that the translation of the test to different languages will likely be relatively unproblematic in countries where there are validated HINT sentences or other openly available set speech corpuses, although verification will be necessary.

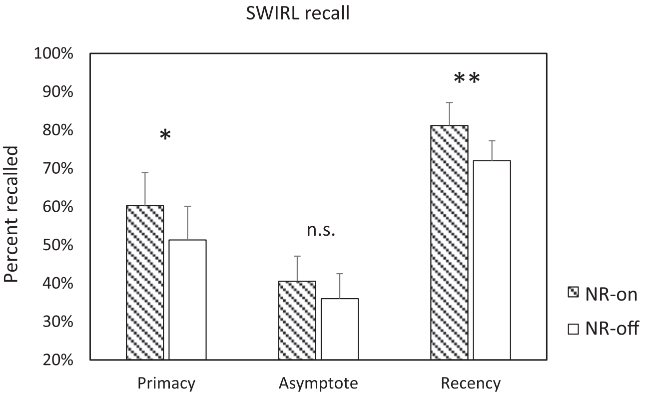


Fig. 3. Mean performance on the SWIRL recall task for items at different serial positions, primacy, asymptote, and recency. Error bars represent 95% confidence intervals (*$p < 0.05$; **$p < 0.01$). NR-on indicates binary masking noise reduction on; NR-off, binary noise reduction off; ns, not significant; SWIRL, Sentence-final Word Identification and Recall Test in a New Language.

Recall of spoken words in noise with NR-on was predicted by recall accuracy on the reading span working memory test, but not by the PTA hearing thresholds. Successful recall of speech is dependent on the ability to encode it. According to the working memory model for ease of language understanding (Rönnberg et al. 2013), when speech input is degraded by background noise or hearing impairment explicit cognitive (i.e., working memory) resources are needed to achieve encoding. Thus, the association between reading span performance and SWIRL performance with NR-on probably reflects the requirement for explicit processing and the allocation of working memory resources during encoding.

Hearing thresholds did not predict recall of spoken words in the NR condition. This result probably indicates that audibility did not affect word recognition accuracy and that the hearing aid successfully compensated for the loss of audibility. This result is also in line with the findings of Humes (2007), who argued that the importance of cognitive factors was revealed once the speech stimuli were made sufficiently audible.

Improvements in recall of spoken words in conditions of binary masking NR were seen for items in both the primary and recency serial positions, but not for those in the asymptote position, indicating that both short-term memory (working memory storage) and transfer to long-term memory were improved by NR. Performance for items in the asymptote position did not differ between NR conditions but did have the expected u-shape for the serial position curve (Murdock 1974).

The results of experiment 1 replicate the results of Ng et al. (2013, 2015), demonstrating that NR signal processing results in improved recall of fully audible speech, and extends them to a new language: Danish. It also shows that speech recall with NR signal processing is associated with working memory such that individuals who recalled more on the reading working memory test had better auditory recall of words presented with NR signal processing, again in line with the results of Ng et al. It is important to emphasize that these studies were performed under ecological test conditions (at SNRs from 6 to 15dB), where many of the current sentence-based SRT-in-noise outcome measures are insensitive.

EXPERIMENT 2: RECALL OF SPOKEN WORDS FOR WORDS TRANSMITTED BY DIFFERENT BONE-ANCHORED PATHWAYS

Experiment 2 was conducted to compare two different types of bone conduction transducers to gain insights into the effects of differences in sound quality on memory. Sound quality is most often tested with subjective ratings (Gabrielsson et al. 1988; Bramsløw 2004). Given that listening to speech when the quality of the signal is poor may require the allocation of attentional and/or working memory resources (for a discussion see Mattys et al. 2012), it is of interest to determine if memory recall paradigms could show such effects in users of hearing devices for which sound quality may vary depending on the implementation of a hearing device. For example, different transmission pathways in bone-anchored hearing technologies may result in different sound quality.

Bone-anchored hearing technologies can use percutaneous or passive transcutaneous solutions, which are also called direct-drive and skin-drive solutions, respectively. Percutaneous bone-anchored hearing systems use an osseo-integrated implant with a skin-penetrating abutment, which serves as the interface for connecting a bone-anchored sound processor to the skull bone. The sound processor includes a transducer that generates vibrations. The (sound) vibrations are transmitted directly to the cochlea via the skull bone. In passive transcutaneous bone-anchored solutions, the sound processor is connected to a footplate that is held in place either by a softband or an implanted magnet; the vibrations are transmitted through intact skin to the skull bone. When passing through the skin, the sound is attenuated by an average of 10 to 20 dB in the mid-to-high-frequency region (Håkansson et al. 1984). Therefore, in skin-drive solutions, higher gain is required to compensate for the attenuation, which results in lower headroom insofar as saturation of the transducer will be reached at a lower input level. Bosman et al. (2014) showed that in bone-anchored sound processors, increased headroom due to higher maximum force output resulted in better overall sound quality. Accordingly, sound quality may be poorer in a skin-drive compared with a direct-drive solution.

In the present experiment, the SWIRL test was used to test the memory effects of two conditions (a) bone-anchored sound processor connected on an abutment, and (b) bone-anchored sound processor connected on a softband. In both cases, the sound processor Ponto Plus Power (Oticon Medical) was used.

Materials and Methods

- Participants

Sixteen native Danish speakers (10 women and 6 men) with conductive (N = 7) or mixed conductive-sensorineural (N = 9) hearing loss were recruited from the Oticon Medical list of participants. Their average age was 57.9 years (SD = 13.6, range: 26 to 78 years), and their average bone conduction in-situ pure-tone threshold average for 0.5, 1, 2, and 4 kHz, as measured with the sound processor connected to the abutment, were 27.0 (SD = 14.9) dB HL BC (bone conduction), ranging from 9 to 55 dB HL (Fig. 4). All were users of the Oticon Medical Ponto bone conduction hearing system before the test. Figure 5 shows the average BC in-situ pure-tone threshold, as measured with the sound processor connected on the abutment and on a softband. As seen from the figure, the BC thresholds that were measured with the sound processor connected on a softband were up to 18 dB higher (worse) than the thresholds that were measured with the sound processor connected to the abutment because of the skin attenuation at mid- and high frequencies. To compensate for this attenuation in the softband condition, the transducer must generate higher gain for equal loudness compared with when the skull is vibrated through the abutment. Thus, in the softband condition, the transducer's maximum output level will be reached at a lower input level than that needed for the abutment condition, creating more distortions in the softband condition due to saturation and will thus affecting sound quality.

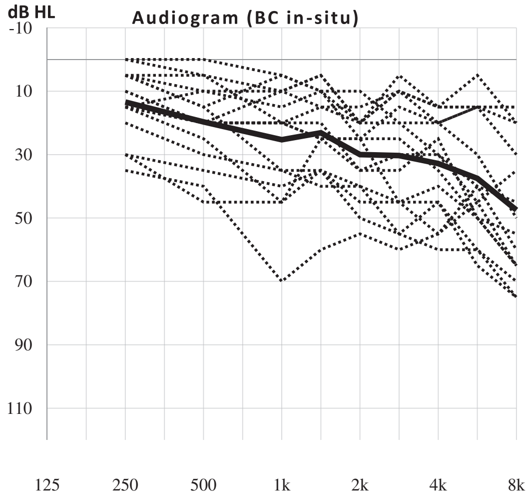


Fig. 4. Bone conduction hearing thresholds (*dotted lines*) of the participants, as measured with the sound processor connected to the abutment (N = 16). The *solid line* indicates average bone conduction hearing thresholds across participants. BC indicates bone conduction.

- Setup

The test was conducted in a sound field with a spatial setup to enable testing both hearing solutions. The loudspeaker setup consisted of four loudspeakers, one in front and three behind the listener, at a height corresponding to the level of the ears of the participants. The target sentences were presented at 0 degree and the masker input, from −110, 180, and +110 degrees azimuth on a ring with a radius of 1.5 m. Target speech was presented at 70 dB SPL (C-weighted). The speech stimuli were the same HINT stimuli described in experiment 1. The competing speech masker consisted of the international speech test signal (ISTS) masker (Holube et al. 2010).

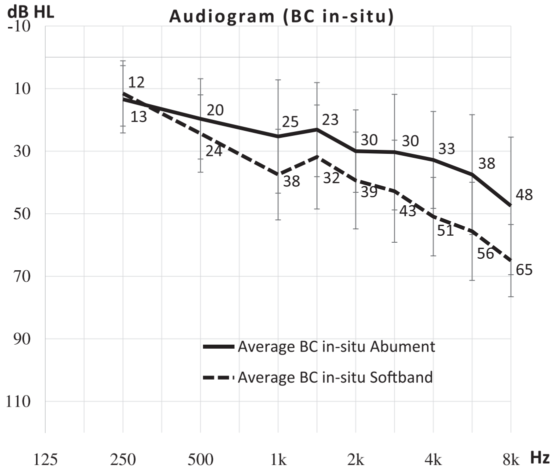


Fig. 5. Average bone conduction hearing thresholds, as measured with the sound processor fitted on an abutment (*solid line*) and softband (*dotted line*). N = 16. BC indicates bone conduction.

- Procedure

The SWIRL method described above was modified. Instead of finding the 84% HINT SRT, a procedure was employed in which the 80% HINT SRT was the starting point. Otherwise the procedure was the same as in experiment 1 (Table 1).

To blind the participants to the procedure, the sound processors on the abutment and softband were worn at the same time, but only one of them was activated at a time. The order of activation was randomized. Before testing, the outputs of the two sound processors were verified using an Interacoustics SKS-10 skull simulator to ensure they were performing equally technically. The prescribed gains in the sound processors were individually determined based on the bone conduction threshold measured via the devices with Genie Medical 2013.1. Figure 6 shows the typical positioning of the sound processor when connected to a softband and abutment.

- Statistical procedure

General linear model repeated measures ANOVA with planned contrasts was used to analyze the data. For all analyses, any p value of 0.05 or lower was considered statistically significant. The a priori hypothesis was as follows: we expected a main effect of transmission pathway, with better recall of spoken words for the direct-drive condition compared with the skin-drive condition due to the better sound quality from the direct drive. No a priori hypotheses were made about interactions between type of drive and serial position recall because no previous comparison data were available.

Results

The SWIRL training procedure resulted in individual SNRs that ranged from +4 to +22 dB, with an average of +10.0 dB SNR (SD = ±4.8 dB). Thus, the SWIRL test was conducted in the positive SNR range.

Word recognition accuracy during the SWIRL test was 96% (SD = ±5%) for the softband condition and 96% (SD = ±5%) for the abutment condition, with no significant differences between conditions [$t(15) = 0.08$; $p = 0.94$]. Thus, successful perception was obtained for both test conditions.

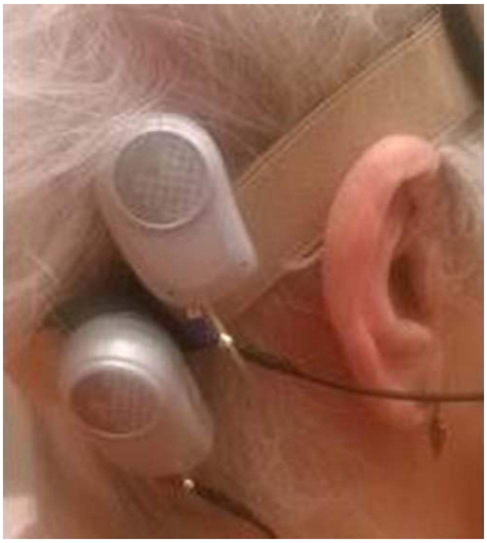


Fig. 6. Oticon medical ponto plus power bone-anchored sound processors fitted on a softband (lower left aid) and abutment (upper right aid) for single-blind direct comparison.

Average recall was higher for the abutment condition (52%, SD = 14%) compared with the softband condition (46%, SD = 11%) and the serial position effects were observed (Fig. 7). These descriptions were confirmed by a two-way repeated measures ANOVA with drive type (direct, skin) and serial position as factors (primacy, asymptote, recency) as within-subject factors was performed. Significant main effects of drive type [$F(1,15) = 5.0$, $p < 0.05$] and serial position [$F(2,30) = 41.4$, $p < 0.001$] were found, but there was no significant interaction between drive type and serial position.

Discussion

The present study showed that stimulation via the abutment (direct drive) resulted in better free recall than that found via the softband (skin drive). The key difference in transmission pathway between the percutaneous and passive transcutaneous solutions is that the abutment provides more efficient energy transduction to vibrate the skull bone, especially at higher frequencies, where the softband fitting saturates at a lower input level. The saturation creates distortions that affect sound quality. The findings of experiment 2 show that this sound quality degradation with softband transmission resulted in poorer recall, suggesting that working memory resources were diverted to listening from storage. Thus, experiment 2 demonstrates that performance on the SWIRL memory task is influenced by signal fidelity and/or by perceived sound quality. Experiment 2 shows the potential usefulness of SWIRL as a new way to *indirectly* evaluate sound quality under ecological test conditions. In contrast to experiment 1, in experiment 2, there was no interaction between device and serial position, perhaps because of insufficient power. Future studies with a larger number of participants may reveal if sound quality differences fundamentally differ from NR effects in terms of the serial position effect on recall.

GENERAL DISCUSSION AND CONCLUSION

The review in the introduction described studies in which people who were hard of hearing were tested under conditions of positive SNR and with successful performance, conditions

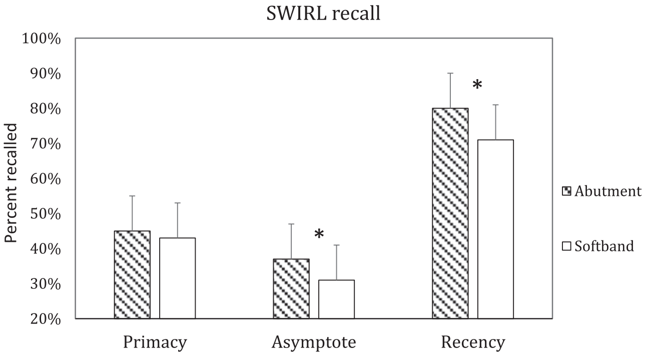


Fig. 7. Mean performance on the SWIRL recall task for items at different serial positions, primacy, asymptote, and recency. Error bars represent 95% confidence intervals (*$p < 0.05$). Abutment indicates direct-drive bone conduction transmission. Softband indicates skin-drive bone conduction transmission. SWIRL indicates Sentence-final Word Identification and Recall Test in a New Language.

which resemble the conditions found in daily life according to Smeds et al. (2015) and Haverkamp et al. (2015). Prior studies have used various methods to measure listening effort when word recognition accuracy was high, including subjective ratings of listening effort, physiological measures such as pupillometry and cognitive measures such as different versions of memory recall. In the present study, it was shown that hearing aid-mediated binary masking NR improved auditory memory recall under conditions of positive SNR and successful perception. Also, testing of different sound transmission pathways under ecological conditions revealed significantly better recall with bone-anchored sound processors that were fitted on an abutment (direct drive) compared with a soft-band (skin-drive).

An assumption that might have been previously made is that if 100% correct word recognition can be achieved, then no further benefit can be gained from hearing aid signal processing. This assumption may explain why adaptive the SRT at 50% correct has been such a popular measure in evaluations of hearing aid processing (in addition to the fact that the SRT method is most sensitive at approximately 50%).

The findings in this study contribute to our understanding of listening effort in the sense that we have shown that cognitive spare capacity can be influenced by signal processing in conditions where all words are correctly repeated. Taking experiments 1 and 2 together, the results seem to suggest that high signal fidelity is important to avoid redirection of cognitive resources away from storage and to listening. In experiment 1, the audibility of the target signal was successfully ensured by the individualized hearing aid gain, and the binary masking NR probably "cleaned" background sounds from the target speech so that the target speech became more salient; therefore, less storage resources were diverted to listening in the NR condition compared with the condition without NR. The skin-drive bone conduction transmission provides more distortion than the direct drive, and this resulted in more cognitive resources being diverted from storage and to listening.

The findings of Smeds et al. (2015) and Haverkamp et al. (2015) suggest that if clinical testing is to resemble daily life conditions, then it should be conducted under much more favorable SNRs than typical SRT measures. The findings of the present study indicate that this can be achieved using recall of spoken words measures in line with the SWIRL paradigm and that such measures may be an alternative that provide useful outcomes under simulated conditions that resemble daily life.

Considering the broader context of the Framework for Understanding Effortful Listening consensus (Pichora-Fuller 2016, this issue, pp. 92S–100S), the results from experiments 1 and 2 seem to suggest that by improving the fidelity of the incoming signal results can reduce the consumptions of resources needed to achieve success on listening tasks. Thus, improved fidelity can be related to Figure 1B in Pichora-Fuller (2016, this issue, pp. 92S–100S) by a reduction in *input-related demands*, here in the form of *transmission factors*, and as a consequence *attention-related responses* improve, here in the form of improved *working memory spare capacity*. Considering Figure 2 in Pichora-Fuller et al. (2016, this issue, pp. 5S–27S), the improved fidelity of the incoming signal will result in lower demands (e.g., going from T_4 to T_3 in Fig. 2) and thereby reducing effort. In summary, the Framework for Understanding Effortful Listening suggest that it is important for hearing instrument manufacturers to provide hearing solutions that provide signal fidelity that is as high as possible in a given listening condition to improve attention-related responses and to reduce effort.

ACKNOWLEDGMENTS

The authors thank Maria Grube Sorgenfrei, Dragan Gusatovic, and Michael Syskind Pedersen from OticonA/S, Smørum, Denmark, and Renskje Hietkamp and Marianna Vatti, Eriksholm Research Centre, Oticon A/S, Denmark, for their contributions to complete Experiment 1.

The authors have no conflicts of interest to disclose.

Address for correspondence: Thomas Lunner, Eriksholm Research Centre, Oticon A/S, Rørtangvej 20, DK-3070 Snekkersten, Denmark. E-mail: thlu@eriksholm.com

Received December 2, 2015; accepted January 29, 2016.

REFERENCES

Akeroyd, M. A., Arlinger, S., Bentler, R. A., et al. (2015). International Collegium of Rehabilitative Audiology (ICRA) recommendations for the construction of multilingual speech tests. *Int J Audiol*, *54*(Suppl 2), 17–22.

Baldwin, C. L., & Ash, I. K. (2011). Impact of sensory acuity on auditory working memory span in young and older adults. *Psychol Aging*, *26*, 85–91.

Boldt, J. B., Kjems, U., Pedersen, M. S., et al. (2008). Estimation of the ideal binary mask using directional systems. Proceedings of the 11th International Workshop on Acoustic Echo and Noise Control, URL (consulted April 2015). Retrieved December 15, 2015, from http://www.iwaenc.org/proceedings/2008/contents/papers/9062.pdf.

Boothroyd, A., & Nittrouer, S. (1988). Mathematical treatment of context effects in phoneme and word recognition. *J Acoust Soc Am*, *84*, 101–114.

Bosman, A. J., Hol, M., Snik, A. F. M., et al. (2014) Evaluating oticon medical ponto plus and ponto pro sound processor. *Audiol Online*, Article 12616.

Bramsløw, L. (2004). An objective estimate of the perceived quality of reproduced sound in normal and impaired hearing. *Acta Acust United Acust*, *90*, 1007–1018.

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am*, *111*, 2801–2810.

Brons, I., Houben, R., Dreschler, W. A. (2013). Perceptual effects of noise reduction with respect to personal preference, speech intelligibility, and listening effort. *Ear Hear*, *34*, 29–41.

Daneman, M., & Carpenter, P. A. (1980). Individual differences in working memory and reading. *J Verb Learn Verb Behav*, *19*, 450–466.

Danermark, B., Granberg, S., Kramer, S. E., et al. (2013). The creation of a comprehensive and a brief core set for hearing loss using the international classification of functioning, disability and health. *Am J Audiol*, *22*, 323–328.

Gabrielsson, A., Schenkman, B. N., Hagerman, B. (1988). The effects of different frequency responses on sound quality judgments and speech intelligibility. *J Speech Hear Res*, *31*, 166–177.

Gosselin, P. A., & Gagné, J. P. (2010). Use of a dual-task paradigm to measure listening effort. *Can J Speech-Lang Pathol Audiol*, *34*, 43–51.

Hagerman, B., & Kinnefors, C. (1995). Efficient adaptive methods for measurements of speech reception thresholds in quiet and in noise. *Scand Audiol*, *24*, 71–77.

Haverkamp, L. (2015). Erfassung von Alltagssituationen mithilfe von Echtzeitaufnahmen und subjektiven Bewertungen [In German]. Unpublished Bachleor Thesis, 2015, Jade Hochschule, Oldenburg, Germany.

Haverkamp, L., Gablenz, P. V., Kissner, S., et al. (2015). Erfassung von Alltagssituationen mithilfe von Echtzeitaufnahmen und subjektiven Bewertungen [In German]. Jahrestagung der Deutschen Gesellschaft für Audiologie DGA, Bochum, Germany.

Håkansson, B., Tjellström, A., Rosenhall, U. (1984). Hearing thresholds with direct bone conduction versus conventional bone conduction. *Scand Audiol*, *13*, 3–13.

Hällgren, M., Larsby, B., Lyxell, B., et al. (2005). Speech understanding in quiet and noise, with and without hearing aids. *Int J Audiol*, *44*, 574–583.

Helfer, K. S., & Freyman, R. L. (2008). Aging and speech-on-speech masking. *Ear Hear*, *29*, 87–98.

Holube, I., Fredelake, S., Vlaming, M., et al. (2010). Development and analysis of an International Speech Test Signal (ISTS). *Int J Audiol*, *49*, 891–903.

Humes, L. E. (2007). The contributions of audibility and cognitive factors to the benefit provided by amplified speech to older adults. *J Am Acad Audiol*, *18*, 590–603.

Kramer, S. E., Kapteyn, T. S., Houtgast, T. (2006). Occupational performance: Comparing normally-hearing and hearing-impaired employees using the Amsterdam Checklist for Hearing and Work. *Int J Audiol*, *45*, 503–512.

Levitt, H. (1971). Transformed up-down methods in psychoacoustics. *J Acoust Soc Am*, *49*, 467–477.

Li, Y., & Wang, D. (2009). On the optimality of ideal binary time-frequency masks. *Speech Commun*, *51*, 230–239.

Lunner, T. (2003). Cognitive function in relation to hearing aid use. *Int J Audiol*, *42*, S49–S58.

Lunner, T., Hietkamp, R. K., Andersen, M. R., et al. (2012). Effect of speech material on the benefit of temporal fine structure information in speech for young normal-hearing and older hearing-impaired participants. *Ear Hear*, *33*, 377–388.

Mattys, S. L., Davis, M. H., Bradlow, A. R., et al. (2012). Speech recognition in adverse conditions: A review. *Lang Cognitive Proc*, *27*, 953–978.

McCoy, S. L., Tun, P. A., Cox, L. C., et al. (2005). Hearing loss and perceptual effort: downstream effects on older adults' memory for speech. *Quart J Exp Psychol A*, *58*, 22–33.

McGarrigle, R., Munro, K. J., Dawes, P., et al. (2014). Listening effort and fatigue: What exactly are we measuring? A British Society of Audiology Cognition in Hearing Special Interest Group "white paper." *Int J Audiol*, *53*, 433–440.

Murdock, B. B. (1974). *Human Memory: Theory and Data*. Potomac, MD: Lawrence Erlbaum Associates.

Neher, T., Behrens, T., Carlile, S., et al. (2009). Benefit from spatial separation of multiple talkers in bilateral hearing-aid users: Effects of hearing loss, age, and cognition. *Int J Audiol*, *48*, 758–774.

Ng, E. H. N., Rudner, M., Lunner, T., et al. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *Int J Audiol*, *52*, 433–441.

Ng, E. H. N., Rudner, M., Lunner, T., et al. (2015). Noise reduction improves memory for target language speech in competing native but not foreign language speech. *Ear Hear*, *36*, 82–91.

Nielsen, J. B., & Dau, T. (2010). The Danish hearing in noise test. *Int J Audiol*, *50*, 202–208.

Nilsson, M., Soli, S. D., Sullivan, J. A. (1994). Development of the hearing in noise test for the measurement of speech reception thresholds in quiet and in noise. *J. Acoust Soc Am*, *95*, 1085.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of age on auditory and cognitive processing: Implications for hearing aid fitting and audiologic rehabilitation. *Trends Amplif*, *10*, 29–59.

Pichora-Fuller, M. K., Schneider, B. A., Daneman, M. (1995). How young and old adults listen to and remember speech in noise. *J Acoust Soc Am*, *97*, 593–608.

Pichora-Fuller, M. K., Kramer, S. E., Eckert, M., et al. (2016). Hearing impairment and cognitive energy: A framework for understanding effortful listening (FUEL). *Ear Hear*, *27*, 5S–27S.

Plomp, R., & Mimpen, A. M. (1979). Speech-reception threshold for sentences as a function of age and noise level. *J Acoust Soc Am*, *66*, 1333–1342.

Rabbitt, P. M. A. (1968) Channel-capacity, intelligibility and immediate memory, *Quart J Exp Psychol*, *20*, 241–248.

Rabbitt, P. (1991). Mild hearing loss can cause apparent memory failures which increase with age and reduce with IQ. *Acta Oto-laryng*, *476*(Suppl), 167–176.

Rönnberg, J., Arlinger, S., Lyxell, B., et al. (1989). Visual evoked potentials: Relation to adult speechreading and cognitive function. *J Speech Hear Res*, *32*, 725–735.

Rönnberg, J., Lunner, T., Zekveld, A., et al. (2013). The ease of language understanding (ELU) model: Theoretical, empirical and clinical advances. *Front Syst Neurosci*, *7*, 13.

Rosen, S., Souza, P., Ekelund, C., et al. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *J Acoust Soc Am*, *133*, 2431–2443.

Rudner, M., & Lunner, T. (2014). Cognitive spare capacity and speech communication: A narrative overview. *BioMed Res Int*, 2014, Article ID 869726, 10 pages.

Rudner, M. (2016). Cognitive spare capacity as an index of listening effort. *Ear Hear*, *27,* 69S–76S.

Sarampalis, A., Kalluri, S., Edwards, B., et al. (2009). Objective measures of listening effort: Effects of background noise and noise reduction. *J Speech Lang Hear Res*, *52*, 1230–1240.

Smeds, K., Wolters, F., Rung, M. (2015). Estimation of signal-to-noise ratios in realistic sound scenarios. *J Am Acad Audiol*, *26*, 183–196.

Smith, S. L., & Pichora-Fuller, M. K. (2015). Associations between speech understanding and auditory and visual tests of verbal working memory: Effects of linguistic complexity, task, age, and hearing loss. *Front Psychol*, *6*, 1394.

Smith, S., Pichora-Fuller, M. K., Wilson, R., et al. (2015). Development of the word auditory recognition and recall measure (WARRM): A working memory test for use in rehabilitative audiology. *Ear Hear*. Revision submitted.

Studebaker, G. A. (1985). A "rationalized" arcsine transform. *J Speech Hear Res*, *28*, 455–462.

Vatti, M. (2013). SWIRL test setup description. Internal Eriksholm Research Centre report no. 2014-14-I.

Wang, D. L., Kjems, U., Pedersen, M. S., et al. (2009). Speech intelligibilityin background noise with ideal binary time-frequency masking. *J Acoust Soc Am*, *125*, 2336–2347.

Wilson, R. H. (2003). Development of a speech-in-multitalker-babble paradigm to assess word-recognition performance. *J Am Acad Audiol*, *14*, 453–470.

World Health Organization. (2001). International Classification of Functioning, Disability and Health. Geneva, Switzerland: World Health Organization.