The effects of task difficulty, background noise and noise reduction on recall

Andreea Micula, Elaine Hoi Ning Ng, Fares El-Azm & Jerker Rönnberg

To cite this article: Andreea Micula, Elaine Hoi Ning Ng, Fares El-Azm & Jerker Rönnberg (2020): The effects of task difficulty, background noise and noise reduction on recall, International Journal of Audiology, DOI: 10.1080/14992027.2020.1771441

To link to this article: https://doi.org/10.1080/14992027.2020.1771441
The effects of task difficulty, background noise and noise reduction on recall

Andreea Micula\textsuperscript{a,b,c}, Elaine Hoi Ning Ng\textsuperscript{a,b}, Fares El-Azm\textsuperscript{a} and Jerker Rönberg\textsuperscript{b,c}

\textsuperscript{a}Oticom A/S, Smårum, Denmark; \textsuperscript{b}Linnaeus Centre HEAD, Swedish Institute for Disability Research, Linköping University, Linköping, Sweden; \textsuperscript{c}Department of Behavioural Sciences and Learning, Linköping University, Linköping, Sweden

**ABSTRACT**

**Objective:** In the present study, we investigated whether varying the task difficulty of the Sentence-Final Word Identification and Recall (SWIR) Test has an effect on the benefit of noise reduction, as well as whether task difficulty predictability affects recall. The relationship between working memory and recall was examined.

**Design:** Task difficulty was manipulated by varying the list length with noise reduction on and off in competing speech and speech-shaped noise. Half of the participants were informed about list length in advance. Working memory capacity was measured using the Reading Span.

**Study sample:** Thirty-two experienced hearing aid users with moderate sensorineural hearing loss.

**Results:** Task difficulty did not affect the noise reduction benefit and task difficulty predictability did not affect recall. Participants may have employed a different recall strategy when task difficulty was unpredictable and noise reduction off. Reading Span scores positively correlated with the SWIR test. Noise reduction improved recall in competing speech.

**Conclusions:** The SWIR test with varying list length is suitable for detecting the benefit of noise reduction. The correlation with working memory suggests that the SWIR test could be modified to be adaptive to individual cognitive capacity. The results on noise and noise reduction replicate previous findings.

**Introduction**

Hearing aids play an essential role in the treatment of hearing loss by providing amplification in order to reduce the effects of auditory impairment (Chisolm et al. 2007; Singh and Launer 2016). Furthermore, advanced hearing aid signal processing technologies are designed to improve audibility and provide listening comfort, especially in difficult listening environments. However, the degree of satisfaction with the hearing aids varies across the population of hearing aid users, which may lead to rare and irregular use of the devices (Lopez-Poveda et al. 2017). This variation may be related to individual differences in cognitive abilities (Lunner, Rudner, and Rönberg 2009; Tognola et al. 2019). In fact, previous research suggests that it is important to take into consideration, and understand the relationship between cognition and speech recognition in noise in individuals with hearing impairment, as cognitive factors may account for individual differences in the benefit obtained from hearing aid signal processing (Akeroyd 2008; Arehart et al. 2013; Arehart et al. 2015; Davies-Venn and Souza 2014; Desjardins 2016; Foo et al. 2007; Lunner 2003; Ng et al. 2013; Ohlenforst, Macdonald, and Souza 2015; Rönberg et al. 2013; Tognola et al. 2019). However, evidence is mixed, as not all research studies support the relationship between hearing and cognition (Desjardins and Doherty 2014; Füllgrabe and Rosen 2016; Lopez-Poveda et al. 2017; Neher et al. 2014). It seems that the relationship is more solid and replicable for older adults, and especially for those with an age-related hearing loss, or for persons with more profound hearing loss. Despite the growing evidence on the importance of this relationship, there is no agreement on how to include cognition in practical matters, such as hearing loss treatment (Souza, Arehart, and Neher 2015).

Cognitive functions need to be addressed in audiological praxis, since they are important in communicative settings. The working memory model of Ease of Language Understanding (ELU) is one of the more comprehensive models of speech understanding in a communicative context (Carroll et al. 2015; Edwards 2016; Rönberg et al. 2013; Rönberg et al. 2008; Schmidtke 2016; Wingfield, Amichetti, and Lash 2015). According to the ELU model, under optimal listening conditions speech input is automatically and rapidly matched to phonological representations in semantic long-term memory, which unlocks the lexicon. This implicit processing does not burden working memory (WM, Lunner, Rudner, and Rönberg 2009; Rönberg et al. 2013), which is the ability to simultaneously phonologically/semantically process, store and make inferences for prediction and postdiction purposes (Baddeley 2012; Rönberg, Holmer, and Rudner 2019). Under sub-optimal listening conditions (e.g. in a noisy environment), however, the matching process may fail. Mismatch is assumed to trigger the explicit processing loop, which places a higher demand on the processing function in WM, because information stored in semantic and/or episodic long-term memory is necessary in order to complete or infer missing information. The higher the demand placed on the processing function, the fewer resources remain available for the storage function, due to the limited capacity of the WM system (Lunner, Rudner, and Rönberg 2009; Rönberg et al. 2013). For people with hearing loss, the...
probability of mismatch is higher than for people with normal hearing due to the distorted auditory perception, as well as artefacts or distortion caused by hearing aid signal processing.

Resource allocation can be measured using the Sentence-final Word Identification and Recall (SWIR) Test (Ng et al. 2013; Ng et al. 2015). The SWIR test is an auditory recall test designed to measure the effect of signal processing in hearing aids on memory for highly intelligible speech in background noise. The test is administered at a signal-to-noise ratio (SNR) corresponding to 95% speech intelligibility, which is considered to have higher ecological validity than the levels used in traditional speech tests, as it is closer to SNRs encountered during real life communication. Since speech intelligibility is high, it can be assumed that nearly all the last words are correctly perceived. Consequently, an improvement in recall performance using noise reduction would likely be due to the increased availability of resources that can be allocated to the storage function in WM. Previous research using the SWIR test (Ng et al. 2013, 2015) suggests that noise reduction improves recall performance for people with hearing loss. This finding indicates that noise reduction frees up some resources that would otherwise be used for processing, making them available for the storage function (Lunner, Rudner, and Rönberg 2009). Implementing a similar paradigm, Neher, Wagener, and Fischer (2018) found that directional microphones significantly improved recall performance in a noisy environment. The study by Neher, Wagener, and Fischer (2018) provides further evidence demonstrating that hearing aid signal processing can provide a cognitive benefit in noisy and ecologically valid listening conditions.

In the first study using the SWIR test, Ng et al. (2013) found that noise reduction improved recall performance only for people with high working memory capacity (WMC). In a subsequent study (Ng et al. 2015), where the list length of the SWIR test was reduced from eight to seven sentences per list, thus reducing the task difficulty, both participants with low and high WMC benefitted from noise reduction in terms of recall performance. This outcome is comparable to that obtained by Neher, Wagener, and Fischer (2018), which does not support the finding that WM modulates benefit from noise reduction. However, while both groups in the study by Ng et al. (2015) benefitted from noise reduction, for the participants with low WMC the benefit was limited to the items at the end of the list. For participants with high WMC, the benefit was spread across all list positions.

A possible explanation of the outcomes from Ng et al. (2013, 2015) is that the task difficulty of the SWIR test might have exceeded the cognitive capacity of some participants in the first study, thus preventing detection of a potential benefit from noise reduction. One way of addressing this problem may be to adapt the task difficulty to individual cognitive capacity. Task difficulty can be manipulated by varying the list length, that is, the amount of sentences per list. In an adaptive procedure, the list length could be increased and decreased based on correct or incorrect responses respectively, in order to individually find the appropriate list length required for a certain performance level.

Before designing an adaptive SWIR test, it is necessary to investigate whether the effect of noise reduction can be captured using varying list lengths. Thus, the current standard of seven sentences per list was included in the present study, alongside three, five and nine sentences per list. These list lengths were chosen so as to gradually increase the difficulty from very easy to very challenging. In order to create a contrast between each difficulty level, the difference between each list length was of two sentences. It is not known how noise reduction affects recall of words in lists of varying length. In the study by Neher, Wagener, and Fischer (2018), which included lists of two, four and six sentences, no interaction between noise reduction and list length was found, suggesting that varying list length may be appropriate for measuring the effect of noise reduction on recall. However, the outcomes may differ for longer lists, which are more cognitively taxing.

Including lists of varying length offers the possibility of evaluating the predictive component of the ELU model, whose main role is preparation for goal attainment. Participants may or may not know in advance how many sentences will be heard in the following list (e.g. how difficult the task will be). Unknown list length is considered to have more ecological validity than known list length, because during real life dialogue the length of the next utterance that will be heard is essentially unpredictable. We assume that not knowing the task difficulty will add uncertainty to the task, presumably making it more difficult, stimulating change in processing strategy (Rönberg 1980, 1981; Rönberg and Nilsson 1987). Not knowing the list length in advance may prevent the participants from using a preferred strategy depending on task difficulty (Grenfell-Essam and Ward 2012). Evidence from studies by Grenfell-Essam and Ward (2012) suggests that there is no difference between participants with and without prior knowledge of list length in terms of proportion of words recalled. As the SWIR test lists are composed of sentences rather than words, we will investigate whether this applies by including a group of participants for whom task difficulty was predictable and one for whom task was unpredictable.

Several studies have demonstrated that recall order is influenced by list length (Grenfell-Essam, Ward, and Tan 2013; Tan et al. 2016; Ward, Tan, and Grenfell-Essam 2010). Evidence from these studies shows that recall of shorter lists tends to be initiated from the first item of the list, even when participants are not instructed to recall in the correct order. For longer lists, the tendency is to start with one of the last list items (Ward, Tan, and Grenfell-Essam 2010). However, the tendency to begin recall of shorter lists with the first item only occurred when the participants were instructed to recall as many items as possible. If only one or two items had to be recalled, they were usually recalled from the end of the list (Grenfell-Essam, Ward, and Tan 2013; Tan et al. 2016). These findings are relevant for the present study, since we are using a modified version of the SWIR test with varying list length. The SWIR test always requires recall of as many words as possible, hence a similar pattern to the one described by Grenfell-Essam, Ward, and Tan (2013) and Tan et al. (2016) could be expected. However, the mentioned studies used list material composed of words, while the SWIR test is composed of sentences. To our knowledge, this type of material has not been used in research studying the probability of first recall (PFR). This method could shed light on differences in storage and retrieval of information as a result of signal processing.

In terms of background noise, Ng et al. (2013) have used four-talker (4T) babble and steady-state-noise (SSN) in order to investigate the effect of noise reduction on recall. They found that the improvement in recall was only significant in 4T babble, but not in SSN. However, as it is not known whether task difficulty affects the benefit of noise reduction in different types of background noise, both 4T babble and SSN were also included in this study.

**Aims**

The first aim of this study was to investigate the effect of task difficulty, determined by list length, on the benefit of noise
reduction. Furthermore, the correlation between WMC and recall performance at each task difficulty level was investigated.

The second aim was to investigate the effect of task difficulty predictability on recall performance. This refers to knowing compared to not knowing list length in advance.

The last aim was to investigate the effect of background noise and noise reduction on recall performance. The findings were compared to those reported by Ng et al. (2013, 2015), in order to assess whether a different noise reduction technology can provide a cognitive benefit.

Method

Participants

An a priori power calculation was conducted using G*Power 3.1, which resulted in a total sample size of 14 participants. The α-error probability, power and effect size were based on values from studies by Ng et al. Although they did not investigate task difficulty predictability, these were the only studies that were considered similar to the current one.

Thirty-two patients (13 female, 19 male) from the audiology clinic of the University Hospital of Linköping (Sweden) were included for participation in this study. The mean age was 65 years (SD = 6.4, range: 46–74 years). All participants were native Swedish speakers with moderate symmetrical sensorineural hearing loss. The bilateral average four frequency pure-tone audiometry threshold (4F PTA) for 500, 1000, 2000 and 4000 Hz was 49.8 dB HL (SD = 5.7). All participants had been using hearing aids for a minimum of 2 years.

Assessment tools

Audiometry

Pure-tone audiometry thresholds were measured according to the American National Standard Methods for Manual Pure-Tone Threshold Audiometry (ANSI S3.21-1978, R1997) (Dobie and Van Hemel 2005).

Montreal Cognitive Assessment (MoCA) Test

In order to screen for mild cognitive dysfunction, the validated Swedish version of the MoCA test was administered (Borland et al. 2017; Nasreddine et al. 2005). The assessment includes major cognitive domains, such as visuospatial/executive functions, naming, attention, language, abstraction, delayed recall and orientation. Nasreddine et al. (2005) recommend using a cut-off score of 26 out of 30 points as an indication of normal cognitive function, although the scores for the group of normal controls in their study ranged between 25.2 and 29.6 (mean = 27.4, SD = 2.2). A more recent meta-analysis found that the recommended cut-off of 26 points is too high and inflates the rate of false positives, especially for older persons (Carson, Leach, and Murphy 2018). Although Carson, Leach, and Murphy (2018) recommend a cut-off of 23 points, we used a cut-off score of 25 points, as it is closer to the one recommended by Nasreddine et al. (2005), while still allowing participation of otherwise unnecessarily excluded older participants. In the present study the mean score was 27.3 (SD = 1.4).

Reading span

A Swedish version of the Reading Span Test was used as a measure of WMC, as it involves simultaneous semantic processing and storage of information (Ng et al. 2013, 2015; Rönnerberg et al. 2016). The test is composed of three-, four- and five-sentence lists, which are presented visually on a computer screen in ascending order. There are two lists of each length (total 24 sentences). The participants were instructed to read each sentence and judge whether it is sensible or absurd immediately after the sentence. After each list, the participants were asked to recall either the first or the last words of the sentences in the respective list. The Reading Span Test was scored as the percentage of correctly recalled words out of the total (Rönnerberg et al. 1989), irrespective of recall order (Petersen et al. 2016). In the present study the mean percentage score was 57.2% (SD = 12.4).

Speech Recognition Test

The Swedish version of the Hearing In Noise Test (HINT) (Hållgren, Larby, and Arlinger 2006; Nilsson, Soli, and Sullivan 1994) was used, in order to measure the speech reception threshold corresponding to 80% speech intelligibility. The task consists of repeating sentences as accurately as possible. The SNR corresponding to 80% speech intelligibility was obtained using a modified up/down procedure, such that the SNR was decreased by 0.8 dB after correct repetition and increased by 3.2 dB after incorrect repetition. For the first five sentences, however, the step size was twice as large. The starting SNR was 0 dB (target sentence and masker both at 65 dB SPL). Subsequently the level of the masker was adapted based on the response, while the target sentences remained fixed at 65 dB SPL.

Sentence-Final Identification and Recall (SWIR) Test

The task of the original SWIR test (Ng et al. 2013) is to listen to lists of sentences and repeat the last word of each sentence immediately after the sentence is finished. At the end of every list, indicated by a beep tone, the participants are asked to recall as many of the repeated words as possible regardless of presentation order. The assumption is that recall performance reflects the resources allocated to storage in relation to the resources needed for sentence processing as such.

In the present study, the SWIR test was modified to contain three, five, seven and nine sentences per list. Since the test was administered in conditions where speech intelligibility was high and as Ng et al. (2015) found no effect of last word repetition, the participants were not asked to repeat the last words in this study. In previous studies by Ng et al., a recall score was calculated based on the number of recalled words out of the number of repeated words. The advantage of this scoring method was that misheard words could be included if they were recalled correctly. Since the last word was not repeated in the current study, scoring was based on the percentage of correctly recalled words out of the total number of last words in the list. The drawback of omitting repetition was that if participants recalled words that did not appear in the list, it could be difficult to assess whether a misheard word was being recalled. Hence, a word that did not appear in the list was marked as correctly recalled, as long as it was phonologically or semantically closely related to one of the last words of a sentence in the list. Amongst the semantically related, some examples were “badrummet” (“the bathroom”) instead of “badkaret” (“the bathtub”) and “lekplatsen” instead of “lekparken”, both meaning “the playground”. An example of a phonological similarity was “isen” (“the ice”) instead of
“spisen” (=“the stove”). Out of the total 6070 recalled words amongst all test participants, only 40 were judged as correctly recalled misheard words.

**Test conditions**

The SWIR test was administered in 16 test conditions: four list lengths, two types of background noise and two types of hearing aid signal processing (within-subject factors). Moreover, the participants were divided in two groups: unpredictable task difficulty group and predictable task difficulty group (between-subjects factor).

**SWIR test lists and task difficulty**

The SWIR test sentence material is composed of recordings of a female voice and is the same as in the previous SWIR test version, rearranged in order to create different lists of varying length. All sentences were repeated twice, but in lists with different sentence combinations and of a different length. The lists have been balanced so that the frequency of the last word is on average approximately equal across lists. The current version of the SWIR test contains 48 test lists, excluding training lists, in order to allow three repetitions per condition (16 conditions \(\times\) 3 repetitions = 48 lists). Additionally, six training lists of seven sentences were used for procedural training. The SWIR test shares the same sentence corpus as the HINT. However, no sentence was repeated in both the HINT and the SWIR test.

**Task difficulty predictability**

Each participant was randomly assigned to the unpredictable or predictable task difficulty group (16 participants in each group). The participants for whom task difficulty was unpredictable were aware about the test containing lists of varying length, but they were not informed about the length before each list. The participants for whom task difficulty was predictable were shown signs containing the numbers three, five, seven or nine before each list. These numbers represented the number of sentences in the following list.

**Hearing aid signal processing**

Signal processing was provided binaurally using a hearing aid simulator. This hardware makes it possible to input sound and record the output. It functions as a real hearing aid and it was programmed via fitting software based on each participant’s PTA thresholds. Two conditions of hearing aid signal processing were used: noise reduction off and on. Noise reduction off was the baseline setting and consisted of amplification only, provided using the Voice Aligned Compression (VAC+) rationale. This rationale is partly based on loudness data from Buus and Florentine (2002). Table 1 shows the insertion gain configuration based on the average PTA thresholds of the test participants for a speech signal at input levels of 50, 65 and 80 dB SPL.

The noise reduction system was based on Boldt et al. (2008), but it used a smoothed SNR estimate, which had been mapped to a continuous gain for each time-frequency unit, rather than the ideal binary mask used by Ng et al. The time resolution was of 1 ms and the frequency resolution of 156 Hz/64 frequency bands. The SNR is estimated in each time-frequency unit and attenuation is applied when the SNR is low, i.e. the probability of those units containing a speech signal is low. The smoothed SNR gain is achieved by estimating the average over a longer period of time.

**Background noise**

The SWIR test was administered in two types of background noise, 4T babble and SSN. The 4T babble was composed of recordings of two female and two male speakers reading different newspaper paragraphs. The SSN background noise was stationary speech-shaped noise. Both types of background noise have long-term average spectra resembling the spectrum of the HINT sentences.

**Test set-up**

The auditory stimuli of the HINT and SWIR test were generated via an external 36-channel soundcard (RME Fireface UC) and transmitted to the hearing aid simulator. Noise reduction was pre-processed in Matlab. The hearing aid simulator was connected to a pre-amplifier (Behringer HA400) from which the auditory stimuli were binaurally transmitted to the participants’ ears via insert earphones (Radioear IP30). The participants were seated inside an anechoic chamber.

**Procedure**

The data were collected during a single test session. Regular breaks were introduced in order to avoid fatigue, as this was a relatively long and demanding test session. The test session consisted of the following sections:

1. First, pure-tone audiometry was measured and hearing loss was compensated for based on the outcome. MoCA and Reading Span were then administered.
2. Next, the individual SNR that corresponds to 80% speech intelligibility was estimated. The HINT was administered in 4T babble and with noise reduction off, which serves as the baseline condition. Four 10-sentence lists from the original HINT training were combined to two lists, thus each list containing 20 sentences. The first list served as procedural training and the second list as the test list determining the SNR to be used at the beginning of the SWIR test training.
3. Four SWIR test training lists, of seven sentences each, were administered with noise reduction off in 4T babble for procedural training as well as for verification of the SNR level. The SNR resulting from the HINT should approximately correspond to 95% speech intelligibility in the SWIR test. Thus, if this level is reached, the participants should be able to correctly repeat six or seven (86–100%) of the last words in a list of seven sentences. We followed the guidelines used in a previous study on how to adjust the SNR in order to yield 95% speech understanding (Lunner et al. 2016). If six or seven words are repeated correctly, no changes are made to the SNR. If four or five words are repeated correctly, the
SNR is increased by 1 dB. If only three or less words are repeated correctly, the SNR is increased by 2 dB. The SNR level obtained after the fourth training list was used for the remainder of the SWIR test. Two additional training lists were administered after the participants were instructed not to repeat the last word.

4. The participants were randomly assigned to one of the two task difficulty predictability groups prior to the test session. Half of the participants performed the SWIR test in 4T babble first and then in SSN, the other half in the reversed order. In each type of background noise, 24 lists were administered. List length and signal processing were randomised.

Results

Independent samples t-tests were conducted in order to assess whether participants in the unpredictable task difficulty group and participants in the predictable task difficulty group significantly differed from each other in regard to age, 4F PTA, MoCA and Reading Span. No significant differences were found (age: t(30) = −0.63, p = 0.53; PTA: t(30) = 1.38, p = 0.18; MoCA score: t(30) = 1.88, p = 0.07 and Reading Span score: t(30) = −0.49, p = 0.63). Thus, differences in SWIR test outcomes between the groups were not likely driven by differences in these background characteristics.

In order to investigate the effects of the within-subject factors task difficulty, background noise and noise reduction on recall, a $2 \times 2 \times 4$ ($\times 2$) analysis of variance (ANOVA) was carried out with task difficulty predictability group as a between-subjects factor. An overview of all main effects and interaction effects is presented in Table 2.

Task difficulty and task difficulty predictability

The ANOVA showed a main effect of list length, $F(2, 90) = 206.64, p < 0.001$, $\eta^2 = 0.87$, indicating that the longer the list, the lower the recall performance as a proportion of the total number of words (3 sentences/list = 95.79%, 5 sentences/list = 78.45%, 7 sentences/list = 61.83%, 9 sentences/list = 52.56%). The recall performance for each list length was significantly different from the others ($p < 0.001$). It is expected that the proportion of recalled words decreases as task difficulty increases. Previous studies have found similar effects of list length (Grenfell-Essam and Ward 2012; Neher, Wagener, and Fischer 2018; Ward 2002). No significant interactions between list length and noise reduction were found. It should be noted that the lack of a significant interaction between list length and noise reduction, $F(3, 90) = 1.33, p = 0.27$, indicates that lists of varying length can be used to measure the effect of noise reduction on recall. Hence, task difficulty does not seem to affect the benefit of noise reduction.

In order to further investigate how WMC is related to the recall of lists of varying length, we carried out a correlation analysis between recall performance and the Reading Span. There was a significant positive correlation (adjusted for multiple comparisons) between recall performance and the Reading Span, but only for lists containing five (Pearson’s $r = 0.62$, $p < 0.001$), seven (Pearson’s $r = 0.57$, $p = 0.001$) and nine (Pearson’s $r = 0.52$, $p = 0.003$) sentences per list. The correlation between recall performance with seven sentences and the Reading Span in this study is comparable to the one reported by Ng et al. (2013). This finding demonstrates that the higher the WMC, the better the recall performance. Consequently, it can be assumed that the insignificant correlation with lists containing three sentences (Pearson’s $r = 0.28$, $p = 0.11$) is likely due to the task difficulty being relatively low and recall performance being close to 100%.

Regarding task difficulty predictability, the test of between-subjects effects did not show a significant difference between the groups in terms of recall performance in the SWIR test, $F(1, 30) = 0.03$, $p = 0.87$. This is in line with evidence from Grenfell-Essam and Ward (2012), whose outcomes likewise demonstrate that there is no difference in terms of proportion of recalled words between groups with and without prior knowledge on list length.

Background noise and noise reduction

The ANOVA showed a main effect of background noise, indicating that recall performance was lower in 4T babble (70.65%) than in SSN (73.67%), $F(1, 30) = 7.51, p = 0.010$, $\eta^2 = 0.20$. Additionally, a main effect of noise reduction was found, $F(1, 30) = 10.37, p = 0.003$, $\eta^2 = 0.26$, indicating that recall performance was significantly improved when noise reduction was on (73.34%) compared to when it was off (70.97%). These findings replicate the outcomes from the studies by Ng et al.

A significant two-way interaction was also found: noise x noise reduction, $F(1, 30) = 10.18, p = 0.003$, $\eta^2 = 0.25$. This is illustrated in Figure 1.

Post hoc pairwise comparisons using a Bonferroni correction to adjust for multiple comparisons at the 0.05 level indicate that recall performance was significantly better when noise reduction was on only in 4T babble ($p = 0.001$). Ng et al. (2013) have previously reported the same finding. This evidence suggests that noise reduction is more efficient in terms of recall benefit when the background noise is competing speech rather than stationary noise.

Furthermore, the three-way interaction noise x noise reduction x group was marginally significant, $F(1, 30) = 4.04$, $p = 0.054$, $\eta^2 = 0.12$. This interaction is shown in Figure 2.

Post hoc pairwise comparisons showed that recall performance was significantly better when noise reduction was on in 4T babble when task difficulty was unpredictable ($p < 0.001$).

### Table 2. Overview of the main effects and interaction effects of the main ANOVA.

<table>
<thead>
<tr>
<th>Effect</th>
<th>$F$</th>
<th>$p$</th>
<th>$\eta^2$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Main effects</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise</td>
<td>7.51</td>
<td>0.010*</td>
<td>0.20</td>
</tr>
<tr>
<td>List length</td>
<td>206.64</td>
<td>0.000*</td>
<td>0.87</td>
</tr>
<tr>
<td>Noise reduction</td>
<td>10.37</td>
<td>0.003*</td>
<td>0.26</td>
</tr>
<tr>
<td>Group</td>
<td>0.03</td>
<td>0.866</td>
<td>0.00</td>
</tr>
<tr>
<td>Two-way interactions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise x Group</td>
<td>0.31</td>
<td>0.585</td>
<td>0.01</td>
</tr>
<tr>
<td>Noise x List length</td>
<td>1.13</td>
<td>0.340</td>
<td>0.04</td>
</tr>
<tr>
<td>Noise x Noise reduction</td>
<td>10.18</td>
<td>0.003*</td>
<td>0.25</td>
</tr>
<tr>
<td>List length x Group</td>
<td>0.59</td>
<td>0.623</td>
<td>0.02</td>
</tr>
<tr>
<td>List length x Noise reduction</td>
<td>1.33</td>
<td>0.270</td>
<td>0.04</td>
</tr>
<tr>
<td>Noise reduction x Group</td>
<td>0.51</td>
<td>0.481</td>
<td>0.02</td>
</tr>
<tr>
<td>Three-way interactions</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise x List length x Group</td>
<td>1.62</td>
<td>0.189</td>
<td>0.05</td>
</tr>
<tr>
<td>Noise x Noise reduction x Group</td>
<td>4.04</td>
<td>0.054</td>
<td>0.12</td>
</tr>
<tr>
<td>Noise x List length x Noise reduction</td>
<td>2.09</td>
<td>0.108</td>
<td>0.07</td>
</tr>
<tr>
<td>List length x Noise reduction x Group</td>
<td>1.45</td>
<td>0.233</td>
<td>0.05</td>
</tr>
<tr>
<td>Four-way interaction</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Noise x List length x Noise reduction x Group</td>
<td>1.65</td>
<td>0.184</td>
<td>0.05</td>
</tr>
</tbody>
</table>

* $p < 0.05$. 

INTERNATIONAL JOURNAL OF AUDIOLOGY
The aims of this study were to investigate the effect of task difficulty on the benefit of noise reduction and the effect of task difficulty predictability on recall performance, as well as the effects of background noise and noise reduction on recall performance.

**Task difficulty and task difficulty predictability**

Regarding task difficulty, the outcomes of the current study demonstrate that it does not have an effect on the benefit of noise reduction. Moreover, a significant positive correlation was found between the Reading Span and recall performance in the SWIR test with five, seven and nine sentences per list. Lastly, task difficulty predictability did not have a significant overall effect on recall performance.

Interestingly, although a main effect of list length was found, it did not interact with noise reduction. This was expected based on the study by Neher, Wagener, and Fischer (2018). Consequently, this finding shows that the SWIR test with varying list length is suitable to capture the benefit of noise reduction on recall. Administering the SWIR test with shorter lists than in previous studies (Ng et al. 2013, 2015) would lead to a reduced test duration.

It is perhaps not surprising that the higher the WMC, the better recall performance, as shown by the positive correlation between the SWIR test and the Reading Span. This indicates that it would be suitable to use shorter lists for individuals with lower WMC and longer lists for individuals with higher WMC. However, the correlation analysis also showed that task difficulty should not be too low, so as to avoid performance at ceiling. This finding lends support to the idea of an adaptive SWIR test, where list length is adapted to individual cognitive capacity for a certain performance level (Ng et al. 2015). Further research is
needed in order to investigate how to create an adaptive version of the SWIR test. For future studies, including a larger sample size in each group or investigating task difficulty predictability as a within-subject factor would be an advantage in terms of power.

The findings related to task difficulty have implications for potential involvement of cognition in clinical praxis, which is currently lacking (Souza, Arehart, and Neher 2015). This study may be an early step in this direction, since the outcomes suggest that the SWIR test could be shortened or that the procedure could be modified to be adapted to individual cognitive capacity. A similar approach was implemented when developing the Word Auditory Recognition and Recall Measure (WARRM) (Smith, Pichora-Fuller, and Alexander 2016), which aims to measure auditory WM. In this test, the amount of words per list increases if three out of five trials are performed correctly. The test stops once this cannot be achieved, therefore avoiding testing at an inappropriate difficulty level, as well as reducing test duration. The WARRM illustrates the advantages of using varying list lengths, which could apply to the SWIR test as well. Further work on the SWIR test could eventually result in a less time-consuming, clinically feasible version of the test. Hence, such a test could be used to evaluate and optimise hearing aid fittings taking into account the cognitive benefit that hearing aids can provide.

**Background noise and noise reduction**

In summary, our outcomes are in line with those from the study by Ng et al. (2013), indicating that the noise reduction technology used in this study improves recall performance in 4T babble.

Previous research has demonstrated that noise reduction can improve recall performance (Neher, Wagener, and Fischer 2018; Ng et al. 2013, 2015). Interpreting this finding from the perspective of the ELU model (Ng et al. 2013), it can be assumed that noise reduction likely decreases the signal degradation caused by background noise and thus facilitates the matching process between the heard input and phonological representations stored in semantic long-term memory. Demands on WM are thereby lowered and more resources can be allocated to the storage function.

Previous studies have also demonstrated that noise reduction is more efficient in 4T babble than in SSN in terms of recall performance measured using the SWIR test (Ng et al. 2013). Albeit using a different noise reduction system, the current study replicates the findings by Ng et al. Attenuating the 4T babble, which poses higher cognitive demands than SSN due to its semantic content, leads to easier segregation of the target sentences (Ng and Rönnberg 2020). According to the ELU model, this could facilitate lexical access by preventing semantic mismatch and consequently facilitating encoding (Sööqvist and Rönnberg 2012).

In this case, noise reduction is needed to counteract the detrimental effects of competing speech on memory when they interfere with WM processes described in the ELU model. Since SSN does not pose the same challenges as 4T babble at such a high speech intelligibility level, the effect of the noise reduction system used in the present study on memory is limited.

The ANOVA conducted in this study also resulted in a three-way interaction between background noise, noise reduction and group (unpredictable vs. predictable task difficulty). A PFR analysis was conducted in order to investigate the improvement in recall performance with noise reduction on when task difficulty was unpredictable in 4T babble. The PFR analysis suggests that there is a higher tendency to begin recall with the first input position when noise reduction is off compared to when it is on. This can be seen on the left side of Table 3, where the PFR is relatively similar between the first and last input positions for all list lengths when noise reduction is off. The increased PFR of the first item in this condition may suggest a higher degree of rehearsal of the items at the beginning of the list (Grenfell-Essam, Ward, and Tan 2013; Rundus and Atkinson 1970). From the perspective of the ELU model, this can be interpreted as the lack of prediction preventing a distinct recall strategy. It is a safer strategy to focus on the items at the beginning of the list, since at least the first three are expected independently of list length, but it is not possible to predict the end of the list. When noise reduction is on, it can be seen that the PFR for the first input position of the longer lists (seven and nine sentences) is decreased compared to when noise reduction is off. This is in line with previous research on recall order as a function of list length, which shows that recall of longer lists tends to begin with one of the last input positions (Grenfell-Essam, Ward, and Tan 2013; Tan et al. 2016; Ward, Tan, and Grenfell-Essam 2010). Since this pattern is not seen with noise reduction off, it may be interpreted as an indication of a less effective encoding strategy contributing to worse recall performance (Unsworth, Miller, and Robson 2019). The interaction between background noise, noise reduction and group should, however, be interpreted with caution. This interaction could be driven by SSN, due to the larger difference between the differences in recall performance with noise reduction on and off with unpredictable (–1.4%) and predictable (2.2%) task difficulty (Figure 2). On the other hand the difference between the differences in recall performance with noise reduction on and off with unpredictable (5.0%) and predictable (3.6%) task difficulty is less in 4T babble. Taking this into consideration, the interaction may not reflect the significance of the effect of noise reduction in 4T babble. This is supported by the main effect of noise reduction as well as the two-way interaction between background noise and noise reduction.

---

**Table 3.** The probability of first recall for each input position of each list length in 4T babble with noise reduction on and off for the unpredictable and the predictable task difficulty groups.

<table>
<thead>
<tr>
<th>Input position</th>
<th>Unpredictable task difficulty</th>
<th>Predictable task difficulty</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>Noise reduction off</td>
<td>0.55</td>
<td>0.09</td>
</tr>
<tr>
<td>Noise reduction on</td>
<td>0.46</td>
<td>0.02</td>
</tr>
</tbody>
</table>
Qualitative/visual inspection of Table 3 reveals some interesting patterns. First, when task difficulty is predictable in 4T babble, there is a clear tendency to start recall from the first word when the list only contains three sentences, corroborating previous findings (Grenfell-Essam, Ward, and Tan 2013; Tan et al. 2016; Ward, Tan, and Grenfell-Essam 2010). However, this pattern does not occur when task difficulty is unpredictable. It is possible that participants who know list length in advance can decide on a recall strategy depending on task difficulty (Grenfell-Essam and Ward 2012). Second, the right side of Table 3 shows that there is a tendency to begin recall of longer lists with the last word both when noise reduction is off and on when task difficulty is predictable. This is expected based on previous research (Grenfell-Essam, Ward, and Tan 2013; Tan et al. 2016; Ward, Tan, and Grenfell-Essam 2010). As mentioned above, this pattern was only prominent when noise reduction was on when task difficulty was unpredictable. These findings may shed light on the relationship between noise reduction in hearing aids and the strategies that hearing aid users implement when recalling speech. Moreover, the length of the following utterance in real life communication is unknown to the listener (i.e. task difficulty is unpredictable), which makes the unpredictable task difficulty condition more ecologically valid.

Conclusions

The findings of the current study demonstrate that task difficulty does not have an overall effect on the benefit of noise reduction. Hence, test administration time could be reduced by using shorter lists. Furthermore, a positive correlation between the SWIR test and the Reading Span at moderate and high task difficulty was found. Thus, although using shorter lists could be useful to reduce administration time, it is important to avoid performance at ceiling. The correlation also indicates that it may be suitable to adapt the task difficulty of the SWIR test to individual cognitive capacity. Further research on the SWIR test is necessary in order to develop such an adaptive procedure or even explore potential clinical application with the aim of evaluating the cognitive benefit of hearing aids in audiological praxis. Task difficulty predictability did not seem to affect overall recall performance. However, the results suggest that listeners employ different recall strategies with noise reduction on and off when task difficulty is unpredictable. The relationship between recall strategies and noise reduction requires further investigation. Moreover, the findings of this study showed using the SWIR test that noise reduction decreases the amount of resources needed to process heard speech, thus allowing allocation of resources to storage of information. Recall performance is hereby improved when noise reduction is activated in competing speech for people with hearing loss. This is in line with previous findings by Ng et al. (2013, 2015).

Acknowledgements

The authors would like to thank Michael Syskind Pedersen and Christian Stender Simonsen from Oticon A/S, Sørum, Denmark, for their technical support with the test set-up of this study, Tomas Bjuumar from Linköping University Hospital, Linköping, Sweden, for his support with participant recruitment and Tirdad Seifi Ala from Eriksholm Research Centre, Snekkersten, Denmark, for helpful discussions on data analysis.

Declaration of interest

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

This project is a collaboration between Oticon A/S and Linnaeus Centre HEAD, Linköping University and is funded by the William Demant Foundation.

ORCID

Andreea Micula http://orcid.org/0000-0002-9155-0715
Elaine Hoi Ning Ng http://orcid.org/0000-0002-7693-1122
Jerker Rönningberg http://orcid.org/0000-0001-7311-9959

References


