Type of speech material affects Acceptable Noise Level test outcome

Xaver Koch[1, 2*], Gertjan Dingemanse[3], André Goedegebure[3], Esther Janse[1, 4]

[1]Center for Language Studies, Radboud University, Netherlands, [2]International Max-Planck Research School for Language Sciences, Netherlands, [3]Department of ENT, Erasmus Medical Center, Netherlands, [4]Max-Planck Institute for Psycholinguistics, Netherlands

Submitted to Journal:
Frontiers in Psychology

Specialty Section:
Language Sciences

ISSN:
1664-1078

Article type:
Original Research Article

Received on:
30 Nov 2015

Accepted on:
31 Jan 2016

Provisional PDF published on:
31 Jan 2016

Frontiers website link:
www.frontiersin.org

Citation:
Koch X, Dingemanse G, Goedegebure A and Janse E(2016) Type of speech material affects Acceptable Noise Level test outcome. *Front. Psychol.* 7:186. doi:10.3389/fpsyg.2016.00186

Type of speech material affects Acceptable Noise Level test outcome

Xaver Koch[1,4]*, Gertjan Dingemanse[2], André Goedegebure[2], Esther Janse[1,3]

[1]Center for Language Studies, Radboud University, Nijmegen, The Netherlands
[2]Department of ENT, Erasmus Medical Center, Rotterdam, The Netherlands
[3]Max-Planck Institute for Psycholinguistics, Nijmegen, The Netherlands
[4]International Max-Planck Research School for Language Sciences, Nijmegen, The Netherlands

*** Correspondence:** Xaver Koch, Center for Language Studies, PO-BOX 310, Nijmegen, 6500 AH, The Netherlands
x.koch@let.ru.nl

Keywords: Acceptable noise level, speech material type, meaningfulness, working memory, self-control capabilities.

Abstract

The Acceptable Noise Level (ANL) test, in which individuals indicate what level of noise they are willing to put up with while following speech, has been used to guide hearing aid fitting decisions and has been found to relate to prospective hearing aid use. Unlike objective measures of speech perception ability, ANL outcome is not related to individual hearing loss or age, but rather reflects an individual's inherent acceptance of competing noise while listening to speech. As such, the measure may predict aspects of hearing aid success. Crucially, however, recent studies have questioned its repeatability (test-retest reliability). The first question for this study was whether the inconsistent results regarding the repeatability of the ANL test may be due to differences in speech material types used in previous studies. Second, it is unclear whether meaningfulness and semantic coherence of the speech modify ANL outcome. To investigate these questions, we compared ANLs obtained with three types of materials: the International Speech Test Signal (ISTS), which is non-meaningful and semantically non-coherent by definition, passages consisting of concatenated meaningful standard audiology sentences, and longer fragments taken from conversational speech. We included conversational speech as this type of speech material is most representative of everyday listening. Additionally, we investigated whether ANL outcomes, obtained with these three different speech materials, were associated with self-reported limitations due to hearing problems and listening effort in everyday life, as assessed by a questionnaire. ANL data were collected for 57 relatively good-hearing adult participants with an age range representative for hearing aid users. Results showed that meaningfulness, but not semantic coherence of the speech material affected ANL. Less noise was accepted for the non-meaningful ISTS signal than for the meaningful speech materials. ANL repeatability was comparable across the speech materials. Furthermore, ANL was found to be associated with the outcome of a hearing-related questionnaire. This suggests that ANL may predict activity limitations for listening to speech-in-noise in everyday situations. In conclusion, more natural speech materials can be used in a clinical setting as their repeatability is not reduced compared to more standard materials.

1. Introduction

One of the most frequent complaints of adult hearing aid users is that comprehending speech is challenging in noisy environments (Nábělek et al., 2006; Cord et al., 2004; Killion et al., 2004) Indeed insufficient benefit of hearing aids in noisy situations seems to be an important reason for people fitted with a hearing aid not to use it. Hearing rehabilitation could be better attuned to the needs of hearing-impaired individuals if audiologists were able to identify those hearing-impaired individuals who will have problems with accepting higher noise levels in everyday communication situations. Individualized counseling may help hearing-impaired individuals to set realistic expectations of hearing-aid benefit in noise. Furthermore, the use of assistive listening devices could then be applied early on for individuals who can be expected to be unsatisfied with hearing devices in noisy environments in order to ultimately minimize disappointment with the device, activity limitations and participation restrictions related to hearing disabilities (cf. Nábělek et al., 2006; Kim et al., 2015).

This raises the question of how to identify future hearing aid users who may be discouraged from using hearing aids because of difficulty listening in noise. One obvious approach would be to measure the individual's objective ability to understand speech in noise (e.g., the standard speech-reception threshold measure). However, such objective performance measures are not predictive of hearing aid benefit or success (Bender et al., 1993; Humes et al., 1996; Nábělek et al., 2006). In contrast, one *subjective* measure called "acceptable noise level" or "tolerated SNR" (henceforth, ANL) seems to be predictive of hearing aid and cochlear implant success (Bender et al., 1993; Humes et al., 1996; Nábělek et al., 1991; Nábělek et al., 2006; Plyler et al., 2008; but cf. Olsen and Brännström, 2014). The ANL procedure involves the following two steps: Listeners are first asked to indicate the loudness level they find most comfortable (henceforth, Most Comfortable Loudness Level, or MCL, cf. Hochberg, 1975) for listening to a continuous speech signal. In a second step, listeners adjust the background noise level (henceforth, Background Noise Level, or BNL) to the maximum level they are *willing* to put up with while following the running speech presented at their individual MCL level. Subtracting the BNL value from the MCL value yields the Acceptable Noise Level (ANL) measure which typically ranges between -15 dB and 40 dB with a mean of around 5 to 12 dB (cf. Eddins, 2013; Nábělek et al., 1991; Nábělek et al., 2006; von Hapsburg and Bahng, 2006; Walravens et al., 2014). The lower the ANL value, the more noise the participant accepts while listening to speech. The ANL measure quantifies the individual's "willingness to listen to speech in background noise" (cf. Nábělek et al., 2006, p. 626). As such, it may be a better indicator of successful hearing aid uptake than the individual's objective ability to understand speech in noise as it is more telling about the individual's wishes, motivation, and intentions.

Speech perception is generally considered to involve an interaction between the processing of acoustic information (bottom-up processing) and linguistic and cognitive processing (top-down processing). An important question is how ANL outcome relates to this interaction, as participants are explicitly instructed to 'follow the speech' during the ANL task. Even though listeners may engage in setting up linguistic hypotheses about upcoming content when the signal is clear, top-down contextual support may be particularly helpful in reconstructing the message when the signal is presented in noise. It is unclear whether type of speech material affects ANL. The original ANL publications (e.g., Nábělek et al., 1991; Nábělek et al., 2006) used a standard stretch of read speech, making up a coherent story (the Arizona Travelogue passage). In contrast, Olsen and Brännström (2014) used the International Speech Test Signal (ISTS; Holube et al., 2010), which is non-meaningful by definition as the signal consists of roughly syllable-sized units from six different languages and speakers, concatenated into a continuous speech stream. Olsen and Brännström (2014) argue that the ISTS can be used to compare ANL values across languages. However, the use of the ISTS precludes top-down processing. In that sense, the question whether type of speech material affects acceptable noise level outcome is a question about the nature of the acceptable noise level task in the broader context of models of speech processing. Regarding the question of whether meaningfulness affects ANL outcome, ANLs obtained with unintelligible speech (i.e., reversed or unfamiliar speech) have been found to be higher (i.e., indicative of lower noise tolerance) than those obtained with intelligible speech (Gordon-Hickey and Moore, 2008). In contrast, Brännström et al. (2012a) showed that ANLs were lower for the ISTS in comparison with meaningful speech stimuli. We investigate whether ANL depends on meaningfulness and coherence by using three different stimulus types that differ in meaningfulness (ISTS vs concatenated sentences and fragments of conversational speech) and coherence (concatenated sentences vs coherent conversational speech). If meaningfulness of the test material does not affect ANL outcome, listeners' acceptance of noise while following speech may mainly rely on bottom-up processing. Consequently, following speech in noise as captured by the ANL task would deviate from speech perception and comprehension. In line with Gordon-Hickey and Moore (2008), we expect to find increased ANL values for the non-meaningful ISTS material compared to the meaningful materials. Our hypothesis regarding the direction of a semantic coherence effect is that participants will accept more noise (i.e., show lower ANLs) for the conversational stimulus type in comparison with the passage of concatenated sentences as redundant information is available on the discourse level, which facilitates speech comprehension. Alternatively, however, the faster speech rate and less careful articulation observed in conversational speech may make listening harder than in the sentence materials and may yield lower noise acceptance.

In order for ANL to be a clinically useful tool in hearing rehabilitation, it is important to establish its repeatability (i.e., consistency over repeated measures or test-retest reliability with the exact same materials). Brännström and Olsen (2014) questioned the repeatability of the existing ANL procedures using the ISTS material. In the present study we investigate whether speech material type affects ANL outcomes and repeatability. Relatedly, repetition of the exact same materials may lead to substantial priming effects, especially for the meaningful materials. Consequently, participants would accept more noise upon repeated exposure, yielding a lower repeatability. We investigate whether the use of meaningful materials yields differential repeatability compared to non-semantic ISTS material.

Nábělek et al. (2006) suggest that *future* hearing aid use can be predicted on the basis of ANL outcome for a majority of hearing aid candidates. Olsen and Brännström (2014), however, challenge the predictive value of ANL outcome for hearing-aid use, and report that results regarding the association between ANL and self-reported hearing-aid outcome measures have been mixed. These inconsistent findings may be caused by the multitude of variables that are possibly related to hearing-aid use, hearing-aid satisfaction and hearing-aid success, as reviewed by Knudsen et al. (2010) and McCormack and Fortnum (2013). Note, however, that self-reported hearing problems have been shown to be consistently associated with hearing-aid outcome measures obtained throughout the process of getting a hearing aid (help seeking, hearing-aid uptake, use, and satisfaction). We investigate whether ANL is associated with (specific components of) the Speech, Spatial, and Qualities of Hearing self-report questionnaire (SSQ; Gatehouse and Noble, 2004) and whether this relation depends on ANL test material type. Our expectation is to find differential correlations between the questionnaire outcome and ANL for three speech stimulus types with stronger associations for the more ecologically valid materials.

The central concept of the ANL measure is 'Listening comfort'. Thus, individual acceptable noise levels are not necessarily linked to the listener's objective ability to comprehend speech in noise, as shown in a number of studies (cf. Nábělek et al., 2004; Plyler et al., 2008; Mueller et al., 2006; von Hapsburg and Bahng, 2006, but cf. Gordon-Hickey and Morlas, 2015). Whether and how the concept of *comfort* in noisy listening situations relates to *listening effort* is unclear. The clinical meaning of the concept of listening effort has recently been discussed in several papers

(McGarrigle et al., 2014; Schulte et al., 2015; Francis and Füllgrabe, 2015; Rennies et al., 2014). One way to quantify *listening effort* is to ask participants to fill in effort-related subscales of self-report questionnaires (cf. McGarrigle et al., 2014). We therefore investigate whether listening effort, as measured with specific questions of the SSQ (Akeroyd et al., 2014) is associated with ANL. We hypothesize that ANL is associated with a listening effort-related subscale of the SSQ with more subjective listening effort related to lower noise acceptance (i.e., higher ANLs).

Listeners need cognitive capacity to map a noisy signal onto stored representations (McGarrigle et al., 2014), as laid out in the Ease of Language Understanding model (Rönnberg et al., 2008, 2013). Multiple studies have shown that hearing aid users' objective speech understanding in adverse conditions (such as background noise) is related to their working memory capacity, verbal working memory in particular (Akeroyd, 2008; Rudner et al., 2011; Ng et al., 2013, 2014). Given the relatively large amount of unexplained variance for individual acceptable noise levels, ANLs may also be associated with working memory. Brännström and colleagues (2012b) found a significant correlation between working memory capacity and ANL for a sample of normal-hearing participants, with lower noise acceptance (i.e., higher ANLs) relating to poorer working-memory capacity. We investigate whether ANL outcomes obtained with the different types of speech materials relate to listeners' working memory capacity, where we expect to replicate the results of Brännström et al. (2012b).

As ANL specifically asks listeners about their willingness to accept noise, ANL may be related to personality traits. Indeed, self-control abilities (i.e., the capability to control thoughts, feelings, impulses and performance; Baumeister et al., 1994), have been found to predict ANL outcomes (Nichols and Gordon-Hickey, 2012). We revisit the question to what extent ANL outcome relates to personality characteristics in this study. We expect to replicate effects of self-control on ANL with better self-control related to lower acceptable noise levels (cf. Nichols and Gordon-Hickey, 2012). Furthermore, even though earlier studies have not found a link between ANL and age (Nábělek et al., 1991; Moore et al., 2011), nor between ANL and pure-tone hearing thresholds (Nábělek et al., 1991; Freyaldenhoven et al., 2007; Plyler et al., 2007), or between ANL and speech perception accuracy in noise (Nábělek et al., 2004), we investigate whether our data replicate this pattern of results.

This study investigates whether speech material type affects ANL outcomes and repeatability for a reference sample of normal-hearing middle-aged and older participants. As addressing these questions on speech material and repeatability involves relatively long testing sessions with repeated ANL measurements, we tested a non-clinical population first so as not to burden a patient population. Future testing is then required to see whether material type effects generalize to a patient population and whether ANLs based on conversational materials better predict hearing aid success than ANL values obtained with more standard audiology materials (such as, e.g., ISTS).

The present study was set up to address the following four research questions:

1. Does ANL outcome depend on the meaningfulness (1A) and semantic coherence (1B) of the speech materials?

2. Does ANL repeatability differ across speech material types?

3. Are ANLs differentially associated with self-report measures of listening effort and of hearing-related activity limitations for the different speech materials?

4. Do participant characteristics such as working-memory (4A), and self-control abilities, age, hearing thresholds, and speech perception in noise predict ANL (4B)?

2. Materials and Methods

Participants

Seventy-one adults were recruited, all native speakers of Dutch, above 30 years of age (39 female, 33 male). From the initial sample, we excluded ten participants whose hearing loss in one or both ears exceeded the Dutch health insurance criterion for partial reimbursement of hearing aids (i.e., pure-tone average over 1000, 2000, and 4000 Hz ≥ 35 dB HL in either ear). We also excluded two participants who suffered from tinnitus and one participant who showed significant binaural low-frequency hearing loss. One participant was excluded because she did not manage to perform the ANL task in the training phase. The 57 remaining participants (34 female, 23 male) ranged in age from 30 to 77 years with an overall mean of 60.7 years (SD=11.0). All participants indicated that they had no hearing impairment and did not use hearing aids. None of the participants had a history of a neurological disease. We followed the protocols of the Radboud University Ethics Assessment Committee for the Humanities. All participants provided written informed consent and were informed that they could withdraw from the study at any time.

Speech Stimuli

Three types of speech materials were used for ANL testing that differed in meaningfulness and semantic coherence: The unintelligible speech-like International Speech Test Signal (ISTS, Holube et al., 2010), a concatenated passage of meaningful Dutch sentences taken from speech material developed by Versfeld et al. (2000; henceforth, SENT), and conversational speech (henceforth, CONV) extracted from the Dutch conversational IFADV corpus (van Son et al., 2008). The 60 seconds long ISTS signal is made up of units that are roughly syllable sized, originating from six female speakers each reading a short standard passage in their native language (being Mandarin, Spanish, English, German, French and Arabic). The ISTS signal had been developed on the basis of an automatic procedure to cut, concatenate and reassemble the roughly syllable sized segments from the original six recordings to create a smooth 60 seconds long speech-like signal including pauses at regular intervals (all pause durations being smaller than 600 milliseconds). The resulting speech rate is approximately 4 syllables per second (Holube et al., 2010). Furthermore, the ISTS signal has been shaped to spectrally match the female international long-term-average speech spectrum (ILTASS, Byrne et al., 1994).

To create the second type of material (SENT), we concatenated fifty sentences from the female speaker of the materials of Versfeld and colleagues (2000) with intervals of 500 milliseconds silence between sentences (total duration of the passage was 120 seconds). These sentences are all between five and eight words long and are semantically coherent. A translated example sentence is: “I hope to be able to catch the train”. The speech rate of the sentences ranges between 3.5 to 5.7 syllables per second (Mean=4.6 syll./sec, SD=0.6). In order to match the spectral properties of the SENT materials to the ISTS materials, the concatenated SENT material was filtered to the ILTASS (combination of male and female signal) using a finite impulse response (FIR) filter between 100 and 16000 Hz.

The third type of speech material was created by extracting two male and two female recordings from the conversational IFADV corpus (van Son et al., 2008). The Dutch open-source IFADV corpus consists of annotated high-quality recording of dialogues on daily topics such as problems in public transport, leisure time activities or vacations. As we wanted to spectrally shape these materials, we selected four longer stretches of speech (CONV1 (female speaker), CONV2 (male speaker), CONV3 (male speaker), CONV4 (female speaker) where only one speaker was speaking, without being interrupted by the dialogue partner. These stretches were based on the available corpus annotations. In a few instances we cut out verbal backchannelling (e.g. “yes”, “hmm”) of the interlocutor, which did not overlap with the target speech. All pauses longer than

500 milliseconds were shortened to 500 milliseconds. The four resulting speech files ranged in duration between 63 and 75 seconds. Speech rate calculated over the breath groups (sequence of words between inhalations) ranged between 2.6 and 7.5 syllables per second (Mean=5.7 syll./sec., SD=1.2; CONV1: 6.10 syll./sec., CONV2: 5.10 syll./sec., CONV3: 5.79 syll./sec., CONV4: 5.89 syll./sec.). In order to match the spectral contents of the conversational materials to the other types of materials, the four conversational fragments were also filtered to the ILTASS (combination of male and female signal) using a FIR filter between 100 and 16000 Hz.

Noise material

The noise stimulus used throughout the ANL test procedure was a non-stationary eight speaker babble noise (BAB8, Scharenborg et al., 2014) filtered to the ILTASS (combination of male and female spectrum) using a FIR filter between 100 and 16000 Hz. In line with the idea of aiming to approximate realistic listening conditions, we used a multi-talker babble noise since it is a typical background sound encountered in daily life.

Experimental procedure

Test set-up

All acceptable noise level (ANL) test materials were presented in a sound-attenuated booth using an Alesis multimix 4USBFX device and Behringer MS16 loudspeakers in front of the listener (0° azimuth) at a distance of 1 meter. Stimuli were presented in a custom application (cf. Dingemanse and Goedegebure, 2015) running in Matlab (v7.10.0) on a MacBook Pro (type 9,1). Participants adjusted the sound level of the speech stimuli or the noise file using the up and down keys of a customized keyboard. The starting intensity for the most comfortable loudness level (MCL) was 45 dB (SPL). The intensity of the speech file for the background noise level (BNL) task was set to the mean of the three measurements in the preceding MCL task. The step size for the intensity adjustment for both tasks was fixed at 2 dB per button press.

All speech and noise materials were scaled to have the same overall level in dB (RMS). Sound level calibration was done using a 2250 Brüel and Kjær real time sound analyzer and a 1000 Hz warble test tone with the same RMS-value as the ANL materials.

ANL instructions

Participants were instructed to first adjust the level of the speech until it was too loud (i.e., up to the first deviation point), then to reduce the intensity until the speech became very soft (being the second deviation point) and lastly find the most comfortable loudness level (MCL). Then the participant's task was to select the maximum background noise level (BNL) they were willing to accept while following the speech at their MCL. They were instructed to use the same pattern of adjustments as described for MCL: turn up the volume of the noise until it was too loud to comfortably listen to the speech (i.e., the first deviation point), then to reduce the noise intensity until the speech became very clear (i.e., the second deviation point) and lastly to find the maximal background noise level they were willing to put up with while following the speech signal (BNL).

Familiarization phase

In order to familiarize participants with the ANL procedure prior to actual testing, each participant was presented with a phonetically balanced Dutch training fragment. A two-minute-long recording of a female Dutch speaker reading a standard text passage (*Dappere fietsers* - 'Brave cyclists') served as training material. The noise stimulus (BAB8) used throughout the actual ANL test (BNL part) also served as background noise during the training session.

Participants first received written instructions on the experimental task (which was a Dutch translation of the instruction provided in Nábělek et al., 2006, p. 639). The experimenter then demonstrated the task, using scripted instructions, which again followed the translation of Nábělek et al. (2006). A visual display was available during the familiarization phase that enabled the participant, as well as the experimenter, to see the course of the presentation level during the MCL and the BNL tasks. Each participant had to demonstrate the expected intensity pattern (up-down-final adjustments, cf. deviation points above) three times in a row for both MCL and BNL components before they could proceed with the test phase.

Test phase

Unlike during the familiarization phase, visual output was available only to the experimenter during the ANL test sessions. Participants had to perform the MCL and BNL tasks for each of the six ANL test stimuli, and each of the two tasks was repeated three times in a row to decrease measurement error (cf. Brännström et al., 2014b; Walravens et al., 2014). The acceptable noise level for each fragment and for each participant was calculated by subtracting the mean BNL from the averaged MCL. Note that stimulus presentation was looped such that if participants had not provided their response before the end of the stimulus, the stimulus was automatically repeated. All participants managed to set the MCL and BNL levels within the stimulus duration in the test phase (minimal duration: 60 s. for the ISTS).

Test repetition

In order to test the repeatability of the ANL measures across the different materials, we asked the participants to do the ANL task twice for each stimulus type (ISTS, SENT, CONV) with exactly the same material. Note that we took into account that the repetition of the exact same materials across sessions could lead to substantial priming effects, especially for the meaningful materials, by including a control variable in our models to capture changes in ANL over test sessions. Participants first performed the ANL test with the different materials at the beginning of the test session, and again (approximately 1 hour later) towards the end of the session. Participant characteristics data were collected in between these two ANL test sessions. During the first ANL session (session I), six different fragments were presented: ISTS, SENT, CONV1, CONV2, CONV3 and CONV4. To restrict testing time, we only presented one fragment for each of the three material types in the test repetition (session II): ISTS, SENT and CONV4. We selected the CONV4 stimulus from the four conversational test fragments because it featured a female speaker (as was the case for the ISTS and the SENT material) and because its speech rate was typical for conversational speech (i.e., 5.89 syllables per second).

Randomization

We used a block-wise randomization procedure to minimize presentation order effects for the material types. Each participant was pseudorandomly assigned to one out of six possible block orders for the speech material types (ISTS, SENT, CONV). The order of the presented speech material types for the second test session (session II) matched the order of session I. The order in which the four conversational materials appeared in the first ANL test session was also randomized. Each participant was randomly assigned one out of 24 possible presentation orders for the conversational speech stimuli.

Tests of participant characteristics

Hearing (Pure-Tone Average)

Hearing status was screened with air conduction pure-tone audiometry using the modified Hughson-Westlake technique for octave-frequencies between 250 and 8000 kHz, including two half-octave frequencies of 3000 kHz and 6000 kHz (see Figure 1). Audiometric averaged thresholds were calculated for the better ear as auditory presentation of the ANL test was binaural. Seven participants showed an asymmetric hearing loss, defined as an interaural difference of more than 10 dB averaged over 500, 1000, 2000, and 4000 Hz (Noble and Gatehouse, 2004). In addition to the pure-tone average over 1000, 2000, and 4000 Hz, we calculated high-frequency PTA^{HF} as the mean threshold over 3000, 4000, 6000, and 8000 Hz. Table 1 displays descriptives for the two PTA measures. Higher values indicate poorer hearing.

[Figure 1 about here]

Speech perception in noise

Speech perception in noise was tested using a standard Dutch speech audiometry test, the CVC word material from Bosman and Smoorenburg (1992, 1995), which is common in clinical practice in the Netherlands. The test allows presenting the materials at SNRs which are reasonably representative of noise levels during everyday communication (Smeds et al., 2015). This test material consists of meaningful monosyllables (e.g., *kaas*, 'cheese') produced by a female speaker arranged in lists of twelve words. The material was presented in a sound-attenuated booth using Behringer MS16 loudspeakers placed in front of the listener (0° azimuth) at a distance of one meter. The CVC words were presented at an intensity level of 65 dB (SPL) mixed with a masking noise of the same intensity (long-term-average spectrum of the recorded speaker). The test score was based on the number of correctly reproduced phonemes (max. three per test item), discarding the first item of each list (which is considered a practice item). Based on Bosman and Smoorenburg's standardizations results, we expected a mean phoneme accuracy score of about 80 to 85 percent for normal hearing adult participants at an SNR of 0 dB (more favorable signal-to-noise ratios may thus lead to ceiling effects in performance). All participants were presented with five consecutive lists (list 31–35), which resulted in a maximum accuracy score of 165 phonemes correct (5 lists × 11 items × 3 phonemes). The speech perception in noise score reported here was quantified as the percentage of correct phonemes produced. Table 1 provides the descriptives for the perception in noise score. Higher values indicate better speech perception in noise.

Reading Span

We used a Dutch version of the well-established reading span test to index working memory (cf. Daneman and Merikle, 1996; Besser et al., 2013; Besser, 2015). The Dutch test consists of 54 grammatically correct sentences, consisting of a noun phrase plus verb phrase. The 54 sentences are divided in twelve sets of three, four, five or six consecutive sentences. Half of the 54 sentences make sense (e.g., The student sang a song); the other half is absurd (e.g., The daughter climbed the past). The sentences were presented orthographically in chunks: first the subject noun phrase was presented (determiner-noun, e.g., *The student*), followed by the verb (e.g., *sang*), followed by the object noun phrase (determiner-noun, e.g., *a song*; cf. Besser, 2015, p. 173). We used E-prime (2.0, Psychology Software Tools) to present the chunks of the respective test sentences (Subject, Verb and Object) consecutively on a computer screen (display time of each chunk: 800 ms, blank inter chunk interval: 75 ms). Font size was 36 pt (Verdana). The primary unspeeded task was to repeat back either the first or the last nouns of the respective test set ranging in length from three to six consecutive sentences. Thus, participants were visually prompted to (orally) recall either the subject noun phrases (first nouns) or the object noun phrases (last nouns) of the 12 test sets. The order in which participants recalled the first or last words was not taken into consideration for the scoring (cf. Besser et al., 2013). Additionally, participants were asked to perform a speeded plausibility judgement after each sentence as a secondary task. This task ensured that participants read and comprehended the sentences. Response time was restricted by imposing a time out of 1.75 s after a visual prompt appeared that initiated the plausibility judgement task. Participants gave their plausibility judgment by either pressing a red (i.e., absurd) or a green button (i.e., makes sense) on a customized standard keyboard. Participants received written task instructions and completed a training test set before the actual test started. Reading span score was quantified as the percentage of correctly recalled nouns across the 12 sets. Table 1 displays the descriptives for the Reading Span test. Higher values indicate better working memory capabilities.

Self control

Participants filled in a Dutch translation of the Brief Self-Control Scale, a 13 item questionnaire using a five-point Likert scale (cf. Kuijer et al., 2008; Tangney et al., 2004). Individual test score were quantified as the percentage of points out of the maximum of 65 points. Table 1 displays the descriptives for the self-control predictor variable. Higher values indicate better self-control abilities.

SSQ questionnaire

Prior to the ANL testing session, participants filled in an online (Dutch) version of the Speech, Spatial and Quality of Hearing Scale (SSQ, Gatehouse and Noble, 2004). The SSQ self-report scale, which consists of 49 items, is subdivided into three parts: Part 1: 'Speech hearing' (14 questions), Part 2: 'Spatial hearing' (17 questions), and Part 3: 'Qualities of hearing' (18 questions). Following Akeroyd et al. (2014), we extracted a factor related to listening effort covering question numbers 15 and 18 of the SSQ subscale 'Qualities of hearing' ('Do you have to put in a lot of effort to hear what is being said in conversation with others?';'Can you easily ignore other sounds when trying to listen to something?'). Hence, we calculated the SSQ 'effort and concentration' subscale by averaging scores over these two questions. We also calculated the average over the first and the third SSQ scale as these two were deemed most relevant. Table 1 presents the descriptive values for averaged SSQ 'Speech hearing' and 'Qualities of hearing' scores, as well as for the factor related to listening effort (SSQ 'effort and concentration'). Higher values on the SSQ scale indicate fewer limitations in self-reported activity due to hearing problems. Table 2 provides a correlation matrix of all the participant-related characteristics.

[TABLE 1 about here]

[TABLE 2 about here]

3. Analyses

RQ1

Two separate statistical regression models were run to investigate the effects of meaningfulness and coherence (RQ1) of the test material on ANL, using linear mixed-effect models with participants as random variable. The program R was used with the lme4 package (Bates et al., 2013) and restricted maximum likelihood estimation. *P*-values were calculated using the Anova function of the car package which calculates type II Wald χ^2 values. The categorical within-subject variable *meaningfulness* included two levels: not meaningful (ISTS material) versus meaningful (CONV and SENT material). The within-subject variable *coherence* featured two categories: coherent on sentence level (SENT material) versus coherent on discourse level (CONV material). Block order (order a–f) was included as additional control variable in all models. For the model on meaningfulness (model 1A), we allowed for the possibility that the effect of meaningfulness differed across participants by including a random participant slope for meaningfulness. Similarly, we allowed for the possibility that the effect of semantic coherence differed across participants by including a random participant slope for meaningfulness in the 'coherence' analysis (model 1B). Note that we also included the interaction between session number and meaningfulness (in model 1A) or between session number and coherence (in model 1B), to allow for the possibility that ANLs may systematically change with session number due to semantic priming. Consequently, we also allowed for the possibility that the effect of session number differed across participants by including a random participant slope for both models (model 1A, model 1B).

RQ2

We first ran a linear mixed-effect model (with random intercepts for participants) with ANL differences between test sessions as dependent variable. The question was whether ANL values obtained for the three types of speech materials differed in their repeatability across test sessions. One outlier was excluded from repeatability analysis of the ISTS material as the ANL difference between sessions I and II of this participant exceeded a threshold of the sample mean plus three standard deviations.

Apart from the mixed-effect analysis described above, we followed the procedures described by Brännström et al. (2014b) to assess the repeatability of the three speech materials. Hence, we inspected the Bland-Altman plots (Bland and Altman, 1986; Vaz et al., 2013) as well as the coefficient of repeatability (henceforth, CR) for each of the three test materials for which two test sessions had been run. The CR measure is a repeatability (test-retest reliability) measure. It indicates the size of the measurement error in its original measured unit (i.e., dB). In our case, it represents the size of the difference between one measurement (session) and another measurement using the exact same material (with 95% confidence level). The Bland-Altman plots show for each of the three speech materials (ISTS, SENT, CONV4) each participant's mean ANL over the two sessions on the x-axis against the difference between the two sessions on the y-axis. The CR was calculated for each material by multiplying the standard deviation of the differences between ANLs (averaged over repetitions) for the two sessions with 1.96. Additionally, we calculated the coefficients of repeatability for all test materials (i.e., incl. CONV1, CONV2 and CONV3) over their three repetitions within test sessions (repetition 1 versus repetition 2; repetition 2 versus repetition 3). This enabled us to analyze whether repeatability changed within and across test sessions.

RQ3

To assess the question whether self-reported hearing related activity limitations and listening effort differentially predict ANL outcomes for the three different speech materials (RQ3) we set up four linear mixed-effect models that included a categorical speech material variable (ISTS, SENT, CONV) in interaction with one of three variables derived from the SSQ scale (SSQ Part 1, SSQ Part 3, SSQ 'effort and concentration'). Session number was added as categorical covariate to capture repetition effects due to semantic priming. Again, we allowed for the possibility that the effects of session number and speech material differed across participants and therefore added random slopes for the variable speech material and session number to the model.

RQ4

To investigate the effects of participant characteristics (age, hearing thresholds, speech perception in noise accuracy, working memory and self-control abilities) on ANL for the three speech materials (RQ4) we performed 15 correlation analyses (Pearson's *r*) and Bonferroni corrected for multiple comparisons. ANL values were pooled across the two test sessions.

3. Results

Table 3 shows the ANL test results per speech material per test session for the three unrepeated conversational materials (CONV1-3) and the three repeated materials (CONV4, SENT, ISTS).

Mean ANLs are higher for the ISTS material than for the meaningful materials. Figure 2 gives an overview of the ANL test results per test session including the conversational materials that were only presented in test session I (i.e., CONV1, CONV2, and CONV3).

[TABLE 3 about here]

[FIGURE 2 about here]

Research Question 1A: Does ANL outcome depend on the meaningfulness of the speech material?

The results of the statistical model (cf. Table 4) showed that ANLs for the meaningful materials (SENT, CONV) were significantly different from those for the non-meaningful ISTS material ($\chi^2(1, N = 341) = 17.98, p < .001$). Participants showed 1.46 dB higher ANLs and thus less noise acceptance for the ISTS signal in comparison with the meaningful materials. The observed effect direction matched our a-priori hypothesis that participants would accept less noise for the non-semantic ISTS material than for the meaningful materials. Block order of presentation did not influence ANL, nor did session number. These control variables also did not interact with the meaningfulness of the test material. The absence of a significant effect of session number on ANL suggests that ANL was stable over sessions and that no semantic priming occurred between sessions. This absence of priming held across material types as the meaningfulness × session number interaction was insignificant. Block order did not affect the ANL outcome, which suggests that our randomization procedure was adequate. For reasons of brevity block order is left out in the model presented below (the variable having six levels) ($\chi^2(5, N = 341) = 2.13, p > .1$).

We also investigated the effect of meaningfulness including all conversational materials (this implies that it can only be assessed for session I). To that end, we averaged ANLs per participant over the conversational materials (CONV1–CONV4). In line with the results presented in Table 4, this analysis showed an effect of meaningfulness on ANL with less noise acceptance for the non-meaningful ISTS material compared to the two types of meaningful materials ($\chi^2(1, N = 170) = 18.47, p < .001$).

[TABLE 4 about here]

Research Question 1B: Does ANL outcome depend on the semantic coherence of the speech material?

A significant effect of coherence was observed with higher ANLs for the material with coherence on discourse level, i.e. the conversational material ($\chi^2(1, N = 227) = 6.04, p < .05$) than for the concatenated sentences (cf. Table 5). Thus, for the conversational test material participants accepted less background noise. The size of the effect was 1.05 dB. The observed direction of the effect matched the hypothesis that participants would accept less noise for the conversational material, which was coherent at the discourse level, but may have been more difficult in terms of speech rate and speaking style than the concatenated sentences. Again, neither simple nor interaction effects (with the variable of interest, i.e., coherence) were found for the predictors session number and block order suggesting that the randomization procedures were appropriate and that there was no semantic priming from the first to the second session. The control variable block order is not included in the model below for reasons of brevity ($\chi^2(5, N = 227) = 2.62, p > .1$).

We also investigated whether the coherence effect can be generalized to different conversational speech fragments by replacing the conversational ANL values in the analysis above (CONV4) by the average ANL over the four conversational speech materials (CONV1–CONV4) per participant (for the first session only). The results of this alternative analysis did not replicate the previous finding of a coherence effect on ANL ($\chi^2(1, N = 113) = 1.41, p > .1$). Thus, there is no clear evidence for a coherence effect on ANL in our data. We raised the possibility that speech rate may affect ANL outcomes and that the difference between the conversational and concatenated sentences material is not just about discourse coherence, but also about speech rate. To follow up on that, we tested whether speech rate differences between the four conversational fragments affected ANL outcome by setting up a linear mixed-effect model with speech rate as a continuous predictor of ANL (first session measurements only, only conversational fragments). Speech rate turned out not to be a significant predictor of ANL in this subset analysis ($\chi^2(1, N = 228) = 0.33, p > .1$).

[TABLE 5 about here]

Research Question 2: Does ANL repeatability differ across speech material types?

The mixed-model analysis did not show a significant speech material effect on repeatability of the ANL, quantified as the difference between the ANLs per participant for the two test sessions ($\chi^2(2, N = 169) = 0.57, p > .1$). In an additional analysis on repeatability across material types we used the statistical approach of the *coefficient of repeatability* (CR). Figure 3 displays the Bland-Altman plots for the three materials for which two test sessions had been run.

[FIGURE 3 about here]

The highest coefficient of repeatability and thus the lowest repeatability was found for the ISTS material (CR = ±6.65 dB). Both the concatenated sentences material (SENT) as well as the conversational material showed lower coefficients of repeatability and thus numerically slightly better repeatability. For the concatenated sentences material (SENT) the CR was ±6.40 dB. The best repeatability (numerically) was found for the conversational test material with a CR of ±6.14 dB. The combination of these two analyses suggests comparable repeatability across the speech materials.

In an additional step we calculated the coefficients of repeatability for all test materials over subsequent repetitions within test sessions. Table 6 shows that ANL repeatability increased numerically (i.e., CRs decreased) within test session I for all test materials except for CONV3. The same pattern of improved repeatability is seen for the CRs within test session II except for the SENT material. Overall, the repeatability in test session II does not seem to be numerically different from the repeatability in test session I. Note that repeatability seems to be most stable for the CONV4 material both within and across test sessions.

Research Question 3: Are ANLs differentially associated with self-report measures of listening effort and of hearing-related activity limitations for the different speech materials?

We first tested whether the first *subscale* of the SSQ self-report questionnaire (‘Speech hearing’) would be associated with ANL outcomes. The model showed significant material effects ($\chi^2(2, N = 341) = 21.39$, p < .001) with highest ANLs found for the ISTS material and lowest ANLs for the sentence material (SENT). Importantly, this model showed a significant effect of the subjective questionnaire predictor SSQ (subscale ‘Speech hearing’) on ANL ($\chi^2(1, N = 341) = 4.62$, $p < .05$, see Table 7). Higher scores on the SSQ subscale (i.e., fewer self-reported limitations due to hearing problems) were associated with more noise acceptance and thus lower ANLs. For an increase of 1 point on the SSQ ‘Speech hearing’ subscale the model predicted an ANL decrease of approximately 1 dB, which corresponds to an overall effect size of 4.4 dB (with the SSQ ‘Speech hearing’ subscale ranging from: 4.86 to 9.36). However, the model did not show differential SSQ subscale effects on ANL for the three materials ($\chi^2(2, N = 341) = 0.74$, p > .1).

We also investigated the association between the third *subscale* of the SSQ self-report questionnaire (‘Qualities of hearing) and ANL. The model showed significant material effects with lowest ANLs for the sentence material ($\chi^2(2, N = 341) = 21.31$, p < .001). However, we did not find an association between ANL and the third *subscale* of the SSQ self-report ($\chi^2(1, N = 341) = 0.43$, $p > .1$), nor differential SSQ ‘Qualities of hearing’ effects on ANL for the three materials ($\chi^2(2, N = 341) = 1.56$, $p > .1$).

In a third step we analyzed the association between the factor ‘Effort and concentration’ (questions number 15 and 18 of the ‘Qualities of hearing’ *subscale* of the SSQ) and ANL. As for the analyses above, the model showed significant material effects with lowest ANLs for the sentence material ($\chi^2(2, N = 341) = 21.32$, p < .001). Yet, neither an association of ANL with the factor ‘Effort and concentration’ ($\chi^2(1, N = 341) = 1.80$, $p > .1$) nor differential ‘Effort and concentration’ effects on ANL for the three materials were found ($\chi^2(2, N = 341) = 1.30$, $p > .1$).

[TABLE 7 about here]

Additionally, we explored the strength of the association between the SSQ self-report measures (subscale ‘Speech hearing’) and the ANLs (pooled over sessions) separately for the three materials by running correlation analyses. Only for the conversational material (CONV) a marginally significant correlation ($r = -0.23$, $p = .082$, Pearson’s r) was found.

Research Question 4: Do participant characteristics such as working memory (4A), and age, hearing thresholds, speech perception in noise, and self-control abilities predict ANL (4B)?

Again, ANLs were pooled over the two test sessions for each of the three materials. Working memory was not correlated with ANL ($p > .1$). Likewise, none of the other correlations ($N = 15$) were statistically significant at an alpha level of .05 (i.e., not even before application of any correction required for multiple testing). Similarly, adding participant characteristics as continuous variables to either of the linear mixed-effect models discussed above (for research questions 1A and 1B) did not yield any significant effects of these participant-related variables.

Discussion

The clinical purpose of the acceptable noise level test (ANL) is to predict self-reported hearing problems and future hearing aid success as reliably as possible. Therefore, it is crucial to know whether and how its clinical applicability depends on what speech material listeners are presented with and how the test is administered. Material effects on the outcome of the ANL test have been addressed in numerous studies (von Hapsburg and Bahng, 2006; Gordon-Hickey and Moore, 2008; Ho et al., 2013; Olsen et al., 2012a, Olsen et al, 2012b, Olsen and Brännström, 2014). In a number of recent publications (Olsen et al., 2012a; Olsen et al., 2012b; Brännström et al., 2012; Brännström et al., 2014a; Brännström et al., 2014b) – the International Speech Test Signal (ISTS, Holube et al., 2010) has been used, which is non-meaningful by definition. However, the original ANL test fragment used by Nábělek et al. (2006), in which ANL outcome was shown to be predictive of hearing aid uptake, was a meaningful and coherent read story, and thus linguistically different from the ISTS material. With the present study we investigated material effects on ANL to find out whether meaningfulness and coherence affect ANL (RQ1). In addition, we evaluated the repeatability of the ANL test across a range of test materials to check whether ecologically more valid materials yield a comparable repeatability as more standard audiology materials and the ISTS signal (RQ2). Further, we analyzed the association between ANLs and the outcome of a questionnaire that measures activity limitations due to hearing problems to elaborate on the connection between listening effort and ANLs. We also re-examined the association of working memory and self-control abilities and ANLs (RQ4) found in previous studies (Brännström et al., 2012; Nichols and Gordon-Hickey, 2012).

As expected, ANLs were higher for the ISTS material in comparison with the meaningful materials. Our interpretation of this effect is that the available redundancy for the meaningful materials facilitated speech processing (via top-down processing) and thus led participants to choose higher levels of acceptable noise (i.e., lower ANLs) than for the non-meaningful material. The unintelligible ISTS signal might have led participants to still want to hear as much as possible (i.e., relying more heavily on bottom-up processing). Furthermore, contrasting conversational ANL test materials with a passage of concatenated standard audiology sentences, we have not found convincing evidence for a semantic coherence effect on ANL. Possibly, the faster and more casual speaking style in the conversational material made listening more difficult, but this speaking style effect may have been offset by greater semantic coherence in the conversation, providing a form of discourse redundancy. The data did not provide clear evidence for priming effects across tests sessions (but note that Table 6 shows that coefficients of repeatability were largest between the first and second measurement within test session I). All in all, these results provide some evidence that top-down processing plays a role in ANL performance.

An important question was whether repeatability differs across the three speech materials. Neither the statistical modelling approach nor the analysis of the coefficient of repeatability (CR) showed statistically differential repeatability. Rather, repeatability was comparable for the three speech material types with CR values ranging between ±6.14 dB for the conversational material and ±6.65 dB for the ISTS material. Crucially, a coefficient of repeatability lower or equal to ±6 dB ensures that measurement error is lower than the distance between the two thresholds used to categorize hearing aid users as either successful or unsuccessful (≤ 7 and > 13 dB, cf. Nábělek et al., 2006). Across test sessions, all three speech material types yielded CRs just above the critical ±6 dB threshold. With respect to ANL repeatability within test sessions, the conversational material (CONV4) yielded most stable CRs with values below ±6 dB. Our interpretation of the relatively high CR values across sessions is that listeners' internal criteria for MCL and BNL may be somewhat variable over time, particularly if they are engaged in other activities in-between test and retest measurements. As suggested by Brännström et al. (2014b), noise acceptance while following speech may best be considered a range (Acceptable Noise Range), rather than a specific level (ANL). The relatively poor repeatability of ANL may raise concerns about the clinical value of the ANL as an indicator for hearing aid use and success. However, if the ANL is used to compare two hearing aid conditions within one session, within-session reliability seems to be sufficient. For example, the ANL has been used successfully to show the effect of a noise reduction algorithm (Mueller et al. 2006; Peeters et al. 2009, Dingemanse and Goedegebure, 2015). Further research would be required to investigate whether Acceptable Noise Range may be a more reliable predictor of hearing problems and future hearing aid success than ANL.

Our analysis on the association of ANLs and the outcome of a subjective hearing-related questionnaire (RQ3) relates to recent discussion about the clinical meaning of concepts such as listening effort and fatigue in hearing-impaired individuals (McGarrigle et al., 2014). Our data showed a significant effect of participants' score on the subscale 'Speech hearing' of the Speech, Spatial, and Qualities of Hearing self-report (SSQ, Gatehouse and Noble, 2004) on ANL, particularly when listening to conversational speech. Participants who reported fewer listening problems also tolerated more noise while listening to speech (i.e., lower ANLs). Most questions of the 'Speech hearing' subscale are about conversation in noise. Both measurements (SSQ and ANL) are subjective judgements, where SRT measurements are not. This makes an association between ANL and SSQ more likely than an association between SRT and SSQ. The subscale 'Qualities of Hearing' was not significantly correlated with ANL. The between-participant differences of the 'quality of sound rating' were relatively small in this group of nearly normal-hearing participants. Possibly, perceived sound quality and ANL may be associated among hearing-impaired participants. No association was found between ANL and the subscale 'Effort and Concentration'. This suggests that noise tolerance (as one aspect of listening comfort), is a different concept than the listening effort concept as formulated in these specific questionnaire questions. Further research should clarify differences and commonalities of both concepts.

The association between self-reported listening difficulties in noise and noise acceptance (i.e., ANL) only becomes evident when such an ANL test relates to everyday experiences. We think this result clearly makes a case for the use of ecologically valid conversational materials in clinical testing. Audiologists and speech researchers should think about how representative the type of noise and noise levels are of everyday listening, but they should also care about differences between read aloud speech and spontaneous conversation.

Further, the attempt to replicate working memory effects on ANL was unsuccessful. This suggests that noise tolerance, as one aspect of listening comfort, is not related to individual working memory capacity. Importantly, in line with previous studies (cf. Akeroyd, 2008), working memory was considerably correlated with speech perception in noise (cf. Table 2), with higher working memory relating to better speech perception. The failure to replicate working memory effects on ANL in our study can be accounted for in two ways. First, it may be due to the use of different test materials and test procedures to quantify working memory. The test that Brännström et al. (2012) used to quantify working memory was an *auditory* version of the reading span task in which the examiner presented the sentences orally, which may have increased the contribution of hearing. Alternatively, the lack of a correlation between ANL and working memory can be taken to underline that ANL and speech perception in noise are different in nature. The latter account ties in with our observation that ANLs did not relate to age, hearing thresholds, and speech-in-noise perception abilities. This held in the relatively good-hearing adult sample as tested here, but was also found by Nábělek et al. (1991, 2004), Moore et al. (2011), Freyaldenhoven et al. (2007) and Plyler et al. (2007) for both normal-hearing and hearing-impaired participants. Moreover, we have not found evidence for an association between ANL and self-control abilities reported in Nichols and Gordon-Hickey (2012). However, the latter study used a self-control scale containing 36 items in contrast to the Brief Self-Control Scale with 13 items that we asked our participant to fill in.

The combined pattern of results converges on material effects being present for the acceptable noise level test with better noise tolerance and slightly better and more stable repeatability, at least numerically, for meaningful stimuli. We have also shown that activity limitations due to hearing problems and ANLs are related, especially if conversational materials are used as ANL test material. More natural speech materials can thus be used in a clinical setting as repeatability is not reduced compared to more standard materials. We aim to conduct follow-up research to investigate whether ecologically valid test materials – such as the conversational speech material used in this study – can be used to improve the predictive power of the ANL test for hearing aid success, relative to more standardized speech materials.

References

Akeroyd, M. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *Int.J.Audiol.* 47, Suppl. 2, S53–S71. doi: 10.1080/14992020802301142

Akeroyd, M., Guy, F.H., Harrison, D.L., and Suller, S.L. (2014). A factor analysis of the SSQ (Speech, Spatial, and Qualities of Hearing Scale). *Int.J.Audiol.* 53, 101-114. doi: 10.3109/14992027.2013.824115

Bates, D., Maechler, M., Bolker, B., and Walker, S. (2013). lme4: Linear mixed-effect models using Eigen and S4 (R package version 1.0-5), https://cran.r-project/package=lme4

Baumeister, R.F., Stillwell, A.M., and Heatherton, T.F. (1994). Guilt - an interpersonal approach. Psychological Bulletin 115, 243–267. doi: 10.1037/0033-2909.115.2.243

Bender, R.A., Niebuhr, D.P., Getta, J.P., and Anderson, C.V. (1993). Longitudinal-study of hearing-aid effectiveness. I: Objective measures. J.Speech Hear.Res. 36, 808–819. doi: 10.1044/jshr.3604.808

Besser, J., Koelewijn, T., Zekveld, A.A., Kramer, S.E., and Festen, J.M. (2013). How linguistic closure and verbal working memory relate to speech recognition in noise - A review. *Trends in Amplification* 17, 75–93. doi: 10.1177/1084713813495459

Besser, J. (2015). The connected ear. Influences of cognitive and auditory-temporal processing on speech understanding in adverse conditions. Unpublished PhD thesis, 214 pages.

Bland, J.M., and Altman, D.G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. Lancet 1, 307–310. doi: 10.1016/S0140-6736(86)90837-8

Bosman A.J., and Smoorenburg G.F. (1992). Woordenlijst voor spraakaudiometrie (Compact Disc). Gouda, The Netherlands: Electro Medical Instruments bv & Weenhuis Medical Audio bv.

Bosman A.J., and Smoorenburg G.F. (1995). Intelligibility of Dutch CVC syllables and sentences for listeners with normal hearing and with three types of hearing impairment. *Audiology* 34(5), 260–284. doi: 10.3109/00206099509071918

Brännström, K.J., Lantz, J., Holme Nielsen, L, and Olsen, S. Ø. (2012a). Acceptable noise level with Danish, Swedish, and non-semantic materials. *Int.J.Audiol.* 51, 146–156. doi: 10.3109/14992027.2011.609183

Brännström, K.J., Zunic, E., Borovac, A., and Ibertsson, T. (2012b). Acceptance of background noise, working memory capacity, and auditory evoked potentials in subjects with normal hearing. *J.Am.Acad.Audiol.* 23, 542–552. doi: 10.3766/jaaa.23.7.6

Brännström, K.J., Holm, L., Kastberg, T., and Olsen, S. Ø. (2014a). The acceptable noise level: The effect of repeated measurements and working memory on the most comfortable level in the ANL test. *Int.J.Audiol.* 53, 787–795. doi: 10.3109/14992027.2014.938781

Brännström, K.J., Holm, L., Kastberg, T., and Olsen, S. Ø. (2014b). The acceptable noise level: The effect of repeated measurements. *Int.J.Audiol.* 53, 21–29. doi: 10.3109/14992027.2013.824116

Byrne, D., Dillon, H., Tran, K., Arlinger, S., Wilbraham, K., Cox, R., Hagerman, B., Hetu, R., Kei, J., Lui, C., Kiessling, J., Kotby, M.N., Nasser, N.H.A., Elkholy, W.a.H., Nakanishi, Y., Oyer, H., Powell, R., Stephens, D., Meredith, R., Sirimanna, T., Tavartkiladze, G., Frolenkov, G.I., Westerman, S., and Ludvigsen, C. (1994). An international comparison of long-term average speech spectra. *J.Acoust.Soc.Am.* 96, 2108–2120. doi: 10.1121/1.410152

Cord, M.T., Surr, R.K., Walden, B.E., and Dyrlund, O. (2004). Relationship between laboratory measures of directional advantage and everyday success with directional microphone hearing aids. *J.Am.Acad.Audiol*. 15, 353–364. doi:10.3766/jaaa.15.5.3

Crowley, H.J., and Nabelek, I.V. (1996). Estimation of client-assessed hearing aid performance based upon unaided variables. *J.Speech Hear.Res.* 39, 19–27. doi: 10.1044/jshr.3901.19

Daneman, M., and Merikle, P.M. (1996). Working memory and language comprehension: A meta-analysis. *Psychon.B.Rev.* 3, 422–433. doi: 10.3758/bf03214546

Dingemanse, J.G., and Goedegebure, A. (2015). Application of noise reduction algorithm ClearVoice in cochlear implant processing: Effects on noise tolerance and speech intelligibility in noise in relation to spectral resolution. *Ear Hear.* 36, 357–367. doi: 10.1097/aud.0000000000000125

Eddins, D.A., Arnold, M., Klein, A., Ellison, J. (2013). Individual variability in unaided and aided measurements of the acceptable noise level. *Semin.Hear.* 34, 118–127. doi: 10.1055/s-0033-1341348

Francis, A.L., Fuellgrabe, C. (2015). Research on listening effort: History and methods, theory and practice. *J.Acoust.Soc.Am.* 137, 2209. doi: 10.1121/1.4920043

Freyaldenhoven, M.C., Plyler, P.N., Thelin, J.W., and Hedrick, M.S. (2007). The effects of speech presentation level on acceptance of noise in listeners with normal and impaired hearing. *J.Speech Lang.Hear.Res.* 50, 878–885. doi: 10.1044/1092-4388(2007/062)

Gatehouse, S., and Noble, W. (2004). The speech, spatial and qualities of hearing scale (SSQ). *Int.J.Audiol.* 43, 85–99. doi: 10.1080/14992020400050014

Gordon-Hickey, S., and Moore, R.E. (2008). Acceptance of noise with intelligible, reversed, and unfamiliar primary discourse. *Am.J.Audiol.* 17, 129–135. doi: 10.1044/1059-0889(2008/06-0018)

Gordon-Hickey, S., and Morlas, H. (2015). Speech Recognition at the Acceptable Noise Level. *J.Am.Acad.Audiol.* 26, 443–450. doi: 10.3766/jaaa.14079

Greenberg, S. (1998). Speaking in shorthand – A syllable-centric perspective for understanding pronunciation variation. Proceedings of the ESCA Workshop on Modelling Pronunciation Variation for Automatic Speech Recognition, Kekrade (The Netherlands), pp. 47–56.

Ho, H.C., Wu, Y.H., Hsiao, S.H., Stangl, E., Lentz, E.J., and Bentler, R.A. (2013). The equivalence of acceptable noise level (ANL) with English, Mandarin, and non-semantic speech: A study across the US and Taiwan. *Int.J.Audiol.* 52, 83–91. doi: 10.3109/14992027.2012.733422

Hochberg, I. (1975). Most comfortable listening for loudness and intelligibility of speech. *Audiology* 14, 27–33. doi: 10.3109/00206097509071720

Holube, I., Fredelake, S., Vlaming, M., and Kollmeier, B. (2010). Development and analysis of an International Speech Test Signal (ISTS). *Int.J.Audiol.* 49, 891–903. doi: 10.3109/14992027.2010.506889

Humes, L.E., Halling, D., and Coughlin, M. (1996). Reliability and stability of various hearing-aid outcome measures in a group of elderly hearing-aid wearers. *J.Speech Hear.Res.* 39, 923–935. doi: 10.1044/jshr.3905.923

Killion, M.C., Niquette, P.A., Gudmundsen, G.I., Revit, L.J., and Banerjee, S. (2004). Development of a quick speech-in-noise test for measuring signal-to-noise ratio loss in normal-hearing and hearing-impaired listeners. *J.Acoust.Soc.Am.* 116, 2395–2405. doi: 10.1121/1.1784440

Kim, J., Nam, K.W., Yook, S., Hong, S.H., Jong, D.P., and Kim, I.Y. (2015). Effect of the degree of sensorineural hearing impairment on the results of subjective evaluations of a noise-reduction algorithm. *Speech Commun.* 68, 1–10. doi: 10.1016/j.specom.2015.01.001

Koelewijn, T., Zekveld, A. A., Festen, J. M., Rönnberg, J., and Kramer, S. E. (2012). Processing load induced by informational masking is related to linguistic abilities. *Int.J.Otolaryngol.*, 1-11. doi:10.1155/2012/865731.

Kuijer, R., De Ridder, D., Ouwehand, C., Houx, B., and Van Den Bos, R. (2008). Dieting as a case of behavioural decision making. Does self-control matter? *Appetite* 51, 506–511. doi:10.1016/j.appet.2008.03.014

Lecumberri, G.M.L., and Cooke, M. (2006). Effect of masker type on native and non-native consonant perception in noise. *J.Acoust.Soc.Am.* 119(4), 2445–2454. doi: 10.1121/1.2180210

McCormack A., and Fortnum H. (2013). Why do people fitted with hearing aids not wear them? *Int.J.Audiol.* 52(5), 360–368. doi: 10.3109/14992027.2013.769066

McGarrigle, R., Munro, K.J., Dawes, P., Stewart, A.J., Moore, D.R., Barry, J.G., and Amitay, S. (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. doi: 10.3109/14992027.2014.890296

Merikle, P.M., and Daneman, M. (1996). Memory for unconsciously perceived events: Evidence from anesthetized patients. *Conscious.Cogn.* 5, 525–541. doi: 10.1006/ccog.1996.0031

Moore, R., Gordon-Hickey, S., and Jones, A. (2011). Most Comfortable Listening Levels, Background Noise Levels, and Acceptable Noise Levels for Children and Adults with Normal Hearing. *J.Am.Acad.Audiol.* 22, 286–293. doi: 10.3766/jaaa.22.5.5

Mueller, G.H., Weber, J., and Hornsby, B.W. (2006). The effects of digital noise reduction on the acceptance of background noise. *Trends in Amplification* 10, 83–93. doi: 10.1177/1084713806289553

Nábělek, A. K., Tucker, F. M., and Letowski, T. R. (1991). Toleration of background noises: Relationship with patterns of hearing aid use by elderly persons. *J.Speech Hear.Res.* 34, 679–685. doi: 10.1044/jshr.3403.679

Nábělek, A.K., Tampas, J.W., and Burchfield, S.B.(2004). Comparison of speech perception in background noise with acceptance of background noise in aided and unaided conditions. J.Speech Lang. Hear. Res. 47, 1001–1011. doi.org/10.1044/1092-4388(2004/074)
Nábělek, A.K., Freyaldenhoven, M.C., Tampas, J.W., Burchfield, S.B., and Muenchen, R.A. (2006). Acceptable noise level as a predictor of hearing aid use. J.Am.Acad.Audiol. 17, 626–639. doi: 10.3766/jaaa.17.9.2

Ng, E.H., Rudner, M., Lunner, T., Pedersen, M.S., and Rönnberg, J. (2013). Effects of noise and working memory capacity on memory processing of speech for hearing-aid users. *Int.J.Audiol.* 52, 433–441. doi: 10.3109/14992027.2013.776181

Ng, E.H., Classon, E., Larsby, B., Arlinger, S., Lunner, T., Rudner, M., and Rönnberg, J. (2014). Dynamic relation between working memory capacity and speech recognition in noise during the first 6 months of hearing aid use. *Trends Hear*. 18. 1–10. doi:10.1177/2331216514558688.

Nichols, A.C., and Gordon-Hickey, S. (2012). The relationship of locus of control, self-control, and acceptable noise levels for young listeners with normal hearing. *Int.J.Audiol.* 51, 353–359. doi: 10.3109/14992027.2011.645074

Olsen, S.O., Lanz, J., Holme Nielsen, L., Lantz, J., and Brännström, K.J. (2012a). Acceptable noise level: Repeatability with Danish and non-semantic speech materials for adults with normal hearing. *Int.J.Audiol.* 51(7), 557–563. doi: 10.3109/14992027.2012.666362

Olsen, S.O., Lanz, J., Holme Nielsen, L., and Brännström, K.J. (2012b). Acceptable noise level (ANL) with Danish and non-semantic speech materials in adult hearing-aid users. *Int.J.Audiol.* 51(9), 678–688. doi: 10.3109/14992027.2012.692822

Olsen, S.O., and Brännström, K.J. (2014). Does the acceptable noise level (ANL) predict hearing-aid use? *Int.J.Audiol.* 53, 2–10. doi: 10.3109/14992027.2013.839887

Peeters, H., Kuk, F., Lau, C.C., and Keenan, D. (2009). Subjective and objective evaluation of noise management algorithms. J.Am.Acad.Audiol. 20(2), 89–98. doi: 10.3766/jaaa.20.2.2

Plyler, P. N., Madix, S. G., Thelin, J. W., and Johnston, K. J. (2007). Contribution of high-frequency information to the acceptance of background noise in listeners with normal and impaired hearing. *Am.J.Audiol.* 16, 149–156. doi: 10.1044/1059-0889(2007/019)

Plyler, P.N., Bahng, J., and von Hapsburg, D. (2008). The acceptance of background noise in adult cochlear implant users. *J.Speech Lang.Hear.Res.* 51, 502–515. doi: 10.1044/1092-4388(2008/036)

Recker, K.L., McKinney, M.F., and Edwards, B.W. (2014). Loudness as a cue for acceptable noise levels. *J.Am.Acad.Audiol.* 25, 605–623. doi.org: 10.3766/jaaa.25.6.10

Rennies, J., Schepker, H., Holube, I., and Kollmeier, B. (2014). Listening effort and speech intelligibility in listening situations affected by noise and reverberation. *J.Acoust.Soc.Am.* 136, 2642–2653. doi: 10.1121/1.4897398

Rönnberg J., Rudner, M., Foo, C., Lunner, T. (2008). Cognition counts: a working memory system for ease of language understanding (ELU). *Int.J.Audiol.* 47(Suppl.2), S99–105 doi: 10.1080/14992020802301167.

Rönnberg, J., Lunner T., Zekveld, A., Sörqvist, P., Danielsson, H., Lyxell, B., Dahlström, Ö., Signoret, C., Stenfelt, S., Pichora-Fuller, M. K., and Rudner, M. (2013). The Ease of Language Understanding (ELU) model: theoretical, empirical, and clinical advances. *Front.Syst.Neurosc.* 7(00031). doi: 10.3389/fnsys.2013.00031

Rudner, M., Rönnberg, J., and Lunner, T. (2011). Working memory supports listening in noise for persons with hearing impairment. *J.Am.Acad.Audiol.* 22, 156–167. doi: 10.3766/jaaa.22.3.4

Scharenborg, O., Sanders, E., Cranen, B. (2014). Collecting a corpus of Dutch noise-induced 'Slips of the Ear'. Proceedings of Interspeech 2014.

Schulte, M., Krueger, M., Meis, M., Wagener, K.C. (2015). Subjective listening effort. *J.Acoust.Soc.Am.* 137, 2236. doi: 10.1121/1.4920159

Shi, L.F., Azcona, G., and Buten, L. (2015). Acceptance Noise Level: Effects of the Speech Signal, Babble, and Listener Language. *J.Speech Lang.Hear.Res.* 58, 497–508. doi: 10.1044/2015_jslhr-h-14-0244

Smeds, K., Wolters, F., and Rung, M. (2015). Estimations of signal-to-noise ratios in realistic sound scenarios. *J.Am.Acad.Audiol.* 14(2), 183–196. doi: 10.3766/jaaa.26.2.7

Tangney, J.P., Baumeister, R.F., and Boone, A.L. (2004). High self-control predicts good adjustment, less pathology, better grades, and interpersonal success. *J.Pers.* 72, 271–324. doi:10.1111/j.0022-3506.2004.00263.x

Van Son, R.J.J.H., Wesseling, W., Sanders, E., and van den Heuvel, H. (2008). The IFADV corpus: A free dialog video corpus. Sixth International Conference on Language Resources and Evaluation, LREC 2008, 501–508.

Von Hapsburg, D., and Bahng, J. (2006). Acceptance of background noise levels in bilingual (Korean-English) listeners. *J.Am.Acad.Audiol.* 17, 649–658. doi: 10.3766/jaaa.17.9.4

Vaz, S., Falkmer, T., Passmore, A.E., Parson, R., and Andreou, P. (2013). The case for using the repeatability coefficient when calculation test-retest reliability. *Plos One.* 8(0), 1–7. doi: 10.1371/journal.pone.0073990

Versfeld, N.J., Daalder, L., Festen, J.M., and Houtgast, T. (2000). Method for the selection of sentence materials for efficient measurement of the speech reception threshold. *J.Acoust.Soc.Am.* 107, 1671–1684. doi: 10.1121/1.428451

Walravens, E., Keidser, G., Hartley, D., and Hickson, L. (2014). An Australian version of the acceptable noise level test and its predictive value for successful hearing aid use in an older population. *Int.J.Audiol.* 53, S52–S59. doi: 10.3109/14992027.2013.862599

Table 1: Descriptives for the participant characteristics.

	M	*SD*	Range
Age (years)	60.72	11.04	30 – 77
PTA (dB HL)	16.05	8.16	0 – 31.67
PTA^{HF} (dB HL)	25.09	15.68	-1.25 – 56.25
Speech perception in noise (% correct)	88.22	6.79	67.88 – 96.36
Reading Span (% correct)	28.43	10.73	0 – 48.15
Self-Control Scale (% of maximum)	67.34	12.05	38.46 – 93.85
SSQ Part 1 ‘Speech hearing’ (mean score)	7.07	1.07	4.86 – 9.36
SSQ Part 3 ‘Qualities of hearing’ (mean score)	7.98	0.93	5.50 – 9.83
SSQ ‘effort and concentration’ (mean score)	6.55	1.71	3.00 – 9.50

Table 2: Correlation matrix with correlation coefficients and significance levels for participant characteristics (Spearman’s rank, uncorrected). Significance level notation: $^{***}p < .001$; $^{**}p < .01$; $^{*}p < .05$; $.p < .1$.

	Age	PTA^{HF}	Speech perception in noise SPIN	Reading Span RST	Self-Control Scale SCS	SSQ ‘Speech hearing’ SSQ^{1}	SSQ ‘Qualities of hearing’ SSQ^{3}	SSQ ‘effort and concentration’ SSQ^{EC}
Age								
PTA^{HF}	$.42^{**}$							
SPIN	$-.48^{***}$	$-.71^{***}$						
RST	$-.35^{**}$	$-.28^{*}$	$.51^{***}$					
SCS	.08	.07	.01	-.06				
SSQ^{1}	-.19	-.08	.22.	-.03	$.39^{**}$			
SSQ^{3}	-.17	.01	.21	-.06	$.39^{**}$	$.65^{***}$		
SSQ^{EC}	-.10	-.07	.17	-.02	$.34^{**}$	$.54^{***}$	$.64^{***}$	

877 Table 3: ANL descriptive statistics for the six speech materials and the two test sessions (in dB).

Test material	Test session I		Test session II	
	M	*SD*	*M*	*SD*
CONV1	4.06	4.59	–	–
CONV2	4.39	4.58	–	–
CONV3	5.50	4.29	–	–
CONV4	5.30	4.43	4.81	4.53
SENT	4.32	5.57	4.13	5.24
ISTS	6.25	4.90	5.84	5.25

878 Table 4: Model testing for the effect of meaningfulness on ANL. Significance level notation:
879 ***p < .001; **p < .01; *p < .05; ^{ns}p > .1.

	Estimate	*SE*	*p*
Intercept	4.79	0.62	
Meaningfulness	1.46	0.44	***
Session number	-0.32	0.34	ns
Meaningfulness × session number	-0.09	0.59	ns

880 Table 5: Model testing for the effect of semantic coherence on ANL. Significance level notation:
881 ***p < .001; **p < .01; *p < .05; ^{ns}p > .1.

	Estimate	*SE*	*p*
Intercept	4.25	0.72	
Coherence	1.05	0.46	*
Session number	-0.12	0.43	ns
Coherence × session number	-0.37	0.60	ns

Table 6: Coefficients of repeatability (in dB) for ANL for the six speech materials and the two test sessions contrasting subsequent repetitions.

Test material	Test session I		Test session II	
	repetition 1 vs 2	repetition 2 vs 3	repetition 1 vs 2	repetition 2 vs 3
CONV1	6.04	4.42	–	–
CONV2	6.87	5.29	–	–
CONV3	5.76	6.34	–	–
CONV4	4.98	4.75	5.50	5.07
SENT	6.38	4.65	4.32	6.06
ISTS	6.76	4.68	6.16	5.76

Table 7: Model testing for differential associations between SSQ subscale scores and ANLs for three speech materials (CONV, SENT, ISTS). Significance level notation: ***p<0.001; **p < .01; *p < .05; ^{ns}p > .1.

	Estimate	*SE*	*p*
Intercept (CONV material)	12.14	3.65	
SENT material	-2.73	2.36	ns
ISTS material	0.97	2.39	ns
SSQ Part 1 (‘Speech hearing’)	-0.98	0.51	*
Session number	-0.34	0.31	ns
SSQ (‘Speech hearing’) × SENT material	0.26	0.33	ns
SSQ (‘Speech hearing’) × ISTS material	0.003	0.33	ns

Figure 1: Mean audiometric pure-tone air conduction thresholds (for left and right ear) as a function of frequency. Error bars represent standard errors.

Figure 2: ANL test results per speech material and per test session. Note that the notch plots include a marker for the mean (diamond symbol).

Figure 3: Bland-Altman plots for repeated ANL tests using conversational (CONV), concatenated sentence (SENT) and ISTS material. Horizontal lines represent the mean of the differences over the two test sessions as well as the boundaries for the 95% confidence interval per material type.

Figure 1.TIFF

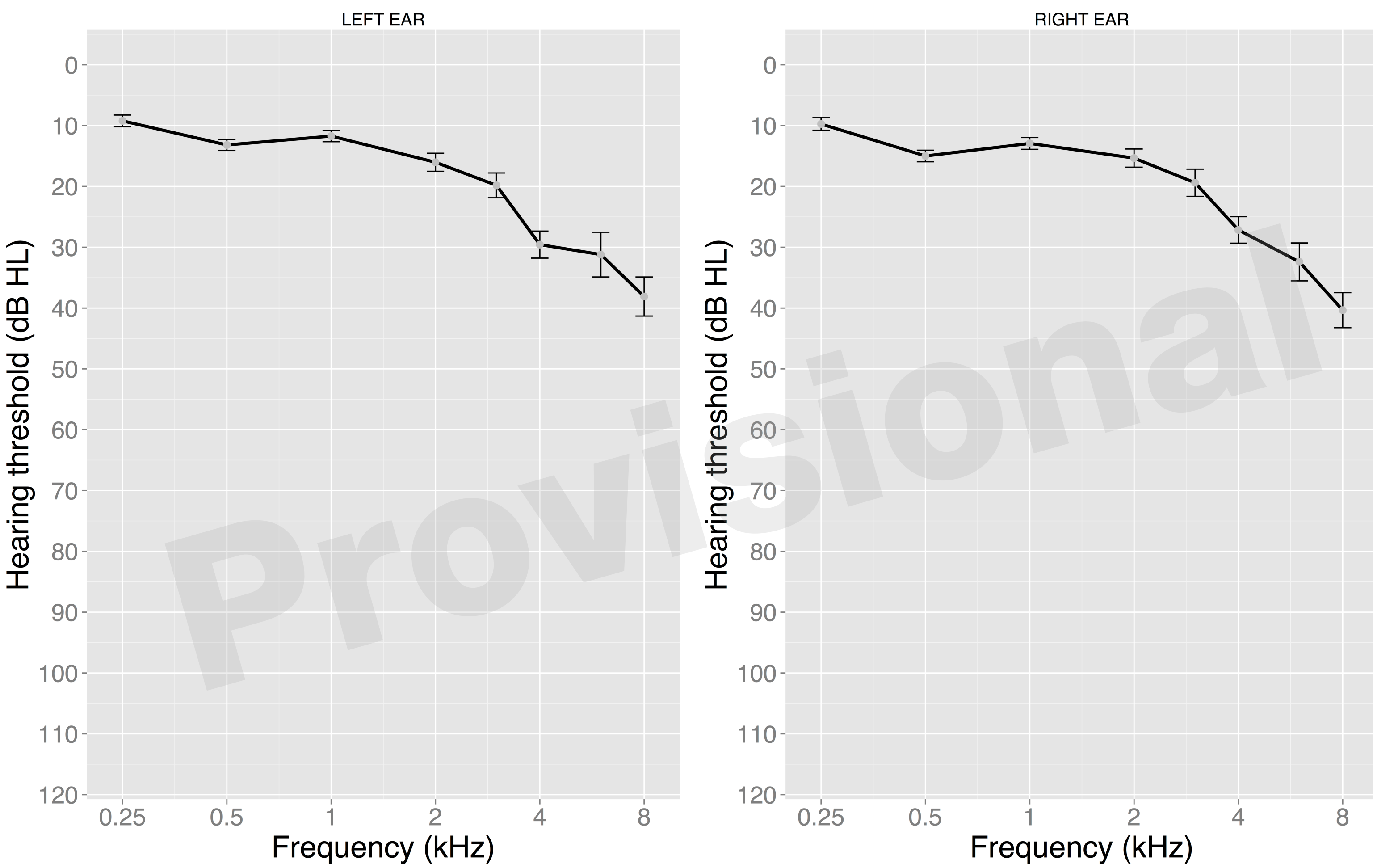

Figure 2.TIFF

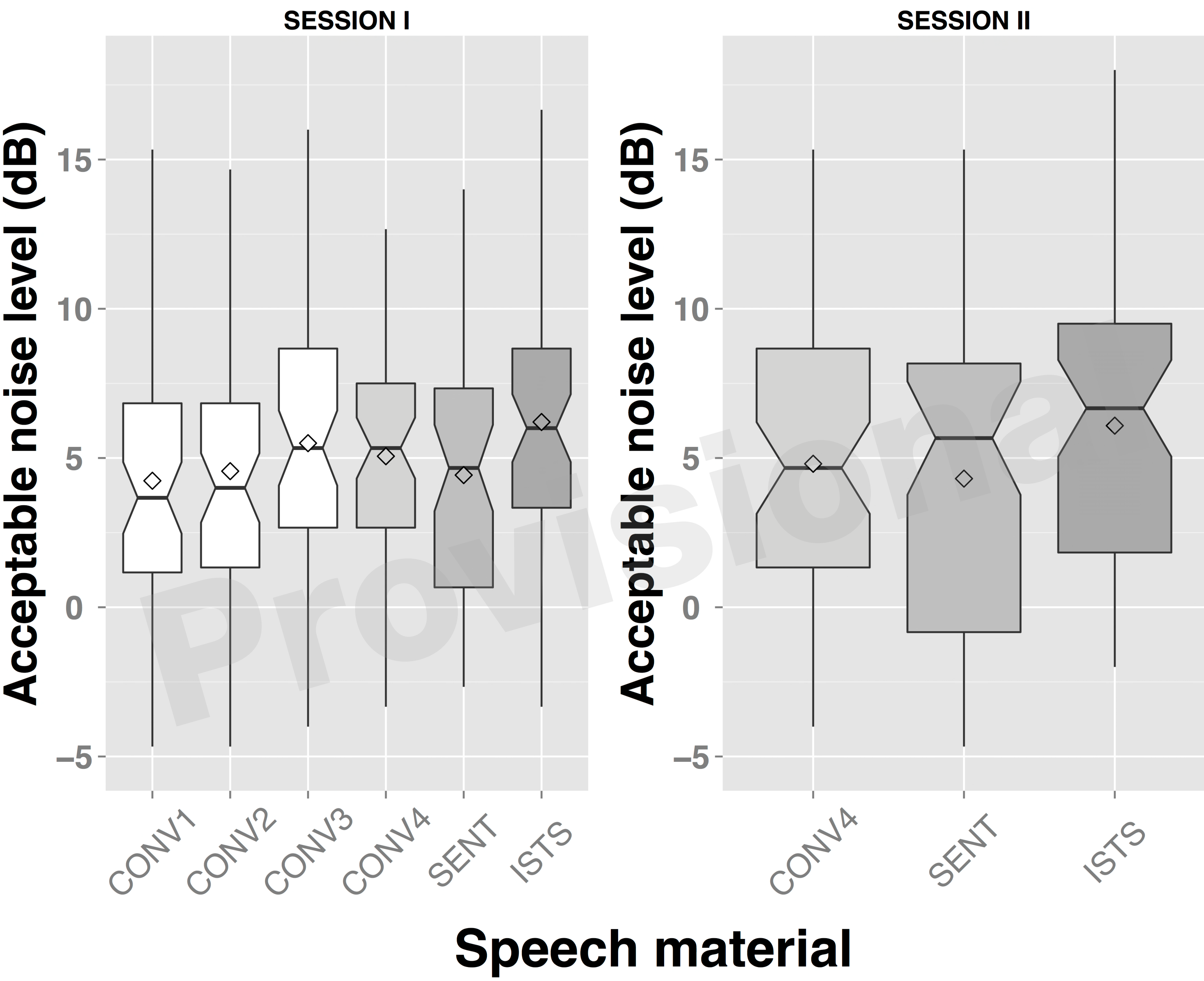

Figure 3.TIFF