International Journal of Audiology

ISSN: 1499-2027 (Print) 1708-8186 (Online) Journal homepage: https://www.tandfonline.com/loi/iija20

Development of the Continuous Number Identification Test (CNIT): feasibility of dynamic assessment of speech intelligibility

Erol J. Ozmeral, Eric C. Hoover, Patricia Gabbidon & David A. Eddins

To cite this article: Erol J. Ozmeral, Eric C. Hoover, Patricia Gabbidon & David A. Eddins (2020): Development of the Continuous Number Identification Test (CNIT): feasibility of dynamic assessment of speech intelligibility, International Journal of Audiology, DOI: 10.1080/14992027.2020.1718782

To link to this article: https://doi.org/10.1080/14992027.2020.1718782

Published online: 31 Jan 2020.

INTERNATIONAL JOURNAL OF AUDIOLOGY
https://doi.org/10.1080/14992027.2020.1718782

ORIGINAL ARTICLE

Development of the Continuous Number Identification Test (CNIT): feasibility of dynamic assessment of speech intelligibility

Erol J. Ozmeral[a], Eric C. Hoover[b], Patricia Gabbidon[a] and David A. Eddins[a,c]

[a]Department of Communication Sciences and Disorders, University of South Florida, Tampa, FL, USA; [b]Department of Hearing and Speech Sciences, University of Maryland, College Park, MD, USA; [c]Department of Chemical and Biomedical Engineering, University of South Florida, Tampa, FL, USA

ABSTRACT

Objective: The present study was motivated by a need for a speech intelligibility test capable of indexing dynamic changes in the environment and adaptive processing in hearing aids. The Continuous Number Identification Test (CNIT) was developed to meet these aims.

Design: From one location in the free field, speech was presented in noise (~2 words/s) with a 100-ms inter-word interval. On average, every fourth word was a target digit and all other words were monosyllabic words. Non-numeric words had a fixed presentation level such that the dominant signal-to-noise-ratio (SNR) was held at +6 dB SNR relative to background maskers. To prevent ceiling effects, however, targets were presented at a user-specific SNR, determined by an initial adaptive-tracking procedure that estimated the 79.4% speech reception threshold.

Study sample: Ten normal-hearing listeners participated.

Results: The CNIT showed comparable psychometric qualities of other established speech tests for long time scales (Exp. 1). Target-location changes did not affect performance on the CNIT (Exp. 2), but the test did show high temporal resolution in assessing sudden changes to SNR (Exp. 3).

Conclusions: The CNIT is highly customisable, and the initial experiments tested feasibility of its primary features which set it apart from currently available speech-in-noise tests.

ARTICLE HISTORY
Received 5 April 2019
Revised 9 January 2020
Accepted 16 January 2020

KEYWORDS
Speech-in-noise test; speech recognition; hearing aid assessment; spatial hearing

Introduction

Hearing aids and other hearing enhancement technologies are designed to improve communication and awareness of environmental sound for persons with hearing deficits. The primary complaint of such individuals most often is difficulty understanding speech when competing background sound is present (Kochkin 2009). As such, a primary outcome measure used to gauge the successful use of such rehabilitative technology, and potentially steer treatment, is a measure of speech understanding in competing backgrounds. The choice of test materials and procedures for measuring speech understanding depends on the specific goals of the evaluation or the research questions to be addressed. For example, a common goal for scientists focussed on hearing devices is to evaluate the effect of a specific signal processing feature or set of features on speech perception in background competition. When such features are inherently static or are (known or presumed to be) static under the chosen test conditions, the test of choice might be whole sentence recognition or sentence key-word identification. However, if the features under study are dynamic in their engagement, disengagement, or function, then short sentences may not provide a very useful assay of performance, as the dynamics associated with the features may be on a time scale that is considerably shorter or longer than a single sentence. Tests involving speech passages are not suitable for the validation of dynamic signal processing features because of the long time delay necessary for the listener to hear the sentence, comprehend the information, and recall an appropriate response. The goal of this study is to evaluate the psychometric properties of a novel speech test that is suitable for evaluating dynamic signal processing features in modern hearing enhancement devices.

To identify the specifications required for a test of dynamic signal processing features, it is important to consider the environments that trigger them. The engagement of many hearingaid features, first requires analysis and classification of the acoustic environment. Such a process usually takes several to tens of seconds, depending on the stability of the environment and the function of the specific classifier. Once the classification is complete, the decision to enable or disable a feature or set of features must be made. This usually can be done very quickly, on a time scale of milliseconds or tens of milliseconds. Finally, the time required to fully engage or disengage a feature can vary widely, based on a host of variables considered important by the manufacturer, and certainly including the desire to avoid intrusive or bothersome transitions. It is also the case that the environmental events that trigger changes in signal processing can vary in their temporal characteristics. For example, in a turn-taking conversation, switching of talkers may be nearly instantaneous or with long-duration pauses between talkers. Likewise, the addition of a competing sound source may be gradual or increase suddenly, as when a machine such as a blender is suddenly turned on. Based on these, and other real-life considerations, it is desirable to have a test that has good temporal resolution (e.g., 1–2 s) but that can assess changes in

CONTACT Erol J. Ozmeral eozmeral@usf.edu Department of Communication Sciences and Disorders, University of South Florida, 4202 E Fowler Ave, PCD 1017 Tampa, FL 33620, USA

performance with this resolution over long periods of time (e.g., tens of seconds).

Apart from the time scale, the target-to-interferer signal-to-noise ratio (SNR) is another critical environmental parameter for several reasons. First, it is important that the test operates within a relevant range of SNR values. We know that in common real-world listening scenarios with background noise, the SNRs frequently range from −6 dB to +12 dB (Smeds, Wolters, and Rung 2015; Wu et al. 2018). Above that range, performance is not very different from listening in quiet, while below that range most people are not able or willing to listen in communication scenarios. Second, it is between this range of SNR values that many hearing enhancement devices engage and disengage various signal processing features. Third, to gauge performance behaviourally, it is important to avoid possible ceiling or floor effects. Unfortunately, for many existing speech tests, aided listener performance is at or near floor or ceiling within that operating range (e.g., Naylor 2016). As Naylor points out, issues related to internal and external validity associated with measuring aided speech perception, and individual differences in SNR operating characteristics, can be substantial impediments to the evaluation of aided listening. These factors are considered in the context of the current test in the Discussion below.

Finally, when evaluating the potential benefit of signal processing features using speech perception tests in an acute listening situation (such as in the laboratory), it is convenient to use closed-set responses that minimise the effects of familiarity and the need for novel stimuli associated with open-set word lists or sentences. Closed-set corpora have the benefit of allowing for tests of longer durations and can be repeated across multiple sessions with less concern for recognition memory. Closed-set corpora also reduce between-subject variability associated with cognitive contributions to sentence recognition relative to open-set recognition (Clopper, Pisoni, and Tierney 2006). This is especially beneficial for evaluating the contribution of bottom-up auditory cues though it may also limit the ability to index the effect of dynamic features on cognitive load, ease of listening, and related top-down processes that contribute to the benefit of amplification. Although many currently available speech tests address complex linguistic and cognitive aspects of hearing or static signal processing features, there are no speech tests currently available that are suitable for the evaluation of dynamic signal processing features given these constraints.

Design goals for new test

- To track performance over short and long time scales.
- To have limitless speech materials.
- To yield performance indices over a wide range of SNR values to accommodate different listener and stimulus characteristics.
- To have psychometric properties that are compatible with SNR values within the operating range of hearingaid feature dynamics while also within the measureable operating range of the human listener.
- To be capable of accommodating adaptive psychophysics methods (e.g., titrating on SNR), when useful, as well as of a constant stimulus method.
- To have rapid responses and automated scoring.

To meet each of these design criteria, we developed a unique number-identification test that can be presented in a variety of background competition scenarios. The target stimuli consist of the nine monosyllabic numbers between one and ten (i.e., excluding the two-syllable number seven). In most competing background sounds, performance on a number identification task is quite good even at poor SNRs that are not typical of real-world communication and are not typical of the SNRs for which hearingaid signal processing features are designed and work well. Therefore, for this test, the monosyllabic numbers are embedded in a continuous stream of monosyllabic, non-numeric words with these occurring in a stream more frequently than the targets. The sequence of non-numeric words interspersed with numeric targets can be presented continuously, with the listener focussing on and responding to the numeric targets. An essential feature of this test is that the non-numeric words can be presented at a relatively high SNR, typical of realistic conversational SNRs, while the numeric targets can be presented at a lower SNR that challenges the limits of the listener's performance on the speech identification task. In this case, the higher SNR of the more frequent non-numeric stimuli would effectively drive signal processing of a hearing instrument whereas the lower SNR of the targets indexes human subject performance. Characteristics of the background competition can be chosen by the experimenter based on the research questions to be addressed, though in this study we will use background competition consisting of eight turn-taking conversations presented simultaneously from eight separate locations in space, effectively simulating an urban conversational scenario, such as a train station café.

Materials and methods

Acquiring the speech materials

Monosyllabic English words were chosen from a list of 2938 words used in a lexical decision study (Balota, Pilotti, and Cortese 2001). Items removed included homonyms, plural forms of another word, proper nouns, words that obviously could be offensive, words that may be confused with digits (e.g., "five" and "hive"), and those with a familiarity value < 2, corresponding to a subjective rating of exposure of at least "once a year" (Balota, Pilotti, and Cortese 2001). The final speech corpus consisted of 2535 such monosyllabic items. In addition to this large list, the numbers "one" through "ten" were included.

All recordings were made in a sound attenuating booth ($10 \times 9.4" \times 6'6"$, single-walled) using an Audio-Technica (AT) 835ST microphone with analog-to-digital conversion via MOTU UltraLite mk3 USB interface. The microphone was aligned to be 12 inches from the talker's mouth such that average digital peak levels were kept constant throughout all recording sessions. A mesh "pop filter" fitted to the microphones blocked most unwanted plosives from entering microphones. Recording software (Sequoia DAW) stored recordings as mono 16-bit files in .WAV format sampled at 44.1 kHz.

Four talkers (two male [M1, M2], two female [F1, F2]) were recruited from the community (M2 was the author EJO), and they were chosen because their dialect was regionally neutral. Talkers were instructed and coached to produce speech with minimal pitch contour and minimal fluctuation of intensity; however, due to the length of recording sessions, and the necessity of breaking the sessions up over a period of several weeks, some expected fluctuation occurred to varying degrees with each talker.

Recording sessions were administered by two trained researchers who monitored each speech utterance over circumaural headphones (Sennheiser HD580 Pro) from outside the testing

booth. Recordings were made in blocks of 25 randomly chosen words in order to prevent talker fatigue. A custom MATLAB interface (The Mathworks, Inc; Natick, MA) prompted the talker for each word while simultaneously prompting the experimenters to accept or reject the talker's utterance for that word. If the experimenters agreed to accept the recording, the word was removed from the block, otherwise if either experimenter rejected the recording, the word returned to the block until an utterance was accepted. Following the block, experimenters listened to each accepted utterance twice and were allowed the opportunity to reject the recording and have the word recorded again.

Following the completion of all blocks (106 total), each recorded word was manually cropped to eliminate any silence in the recording either before or after a minimally audible speech sound. It was important to establish uniformity across the speech samples. The challenge was to determine an anchor parameter that could be applied to each speech utterance. In the end, we chose to normalise to a reference VU (volume unit) value using an algorithm in MATLAB (see VUSOFT; Lobdell and Allen 2007). The VU is distinctly different from RMS or peak power value calculations and normalising to either of those, and it also yields results that were judged to be more consistent in loudness by human subjects than other loudness normalisation algorithms, like R128 (EBU 2014). These other algorithms, specifically designed to measure perceptual loudness, were highly variable, most likely because they were designed for samples far exceeding the short duration of our monosyllabic words. The ultimate reference VU value was determined by computing the average VU value for the entire sample set. Some words, when scaled up to match the reference VU value, had digital peaks clipped at 0 dBFS, so the reference value was adjusted by −6 dB to have enough headroom and to avoid any clipping. The recorded numerals, which form a special subset to be used in the number identification test, were singled out for special attention to ensure absolute consistency in perceptual loudness. The numerals one through ten were manually adjusted by an experienced listener and then validated by a second experienced listener. After manual adjustment, these samples had VU values which deviated from the reference value. Their new average VU value was computed, and then the gain on each numeral was adjusted by a constant value such that the average VU value was equal to the reference value. Among alternative approaches to VU adjustment, one could homogenise digit intelligibility by measuring the intelligibility of each digit across a span of SNRs, then adjust levels to equate thresholds corresponding to a chosen point on the psychometric function (e.g., Houben et al. 2014). This method can be desirable for a large closed-set of targets and especially an open set would require homogenisation across stimuli (Clopper, Pisoni, and Tierney 2006).

Test environment and apparatus

The Continuous Number Identification Test (CNIT) was originally designed to evaluate aided speech-in-noise performance in the context of a number of signal processing features, including the binaural hearing systems deployed in hearing aids. Thus, we chose an array of spatially distributed loudspeakers to simulate free field environments. In the experiments described in this study, the spatial array consisted of 24 KEF Q100 loudspeakers equally-spaced in the horizontal plane at 15 degree intervals on centre. The loudspeakers were fixed atop a 360° aluminium ring with a radius of 41" and height of 51". The array was housed in a sound attenuating booth (10' × 9'4" × 6'6", double-walled) with a height adjustable chair at the centre of the ring. The 24 loudspeakers were powered by three 8-channel amplifiers (Ashly ne8250) with digital-to-analog conversion by a 24-channel external soundcard (MOTU 24ao) controlled via custom MATLAB interface.

Each loudspeaker was calibrated by recording the output at the centre of the array using a $^1/_2$-inch free-field microphone (Brüel & Kjaer Type 4191) with a preamplifier (Brüel & Kjaer Type 2669) and conditioning amplifier (G.R.A.S. 12AA), digitised (MOTU 24ao) for analysis with a custom MATLAB script. A reference voltage for microphone calibration was established by presenting a 1 kHz calibration tone from a calibrator (Brüel & Kjaer Type 4230) at 94 dB SPL. A continuous frequency sweep was presented from each speaker from 0.1 kHz to 20 kHz in three repetitions. The deviation from mean magnitude spectrum was ±1 dB. The relative difference in output level between each loudspeaker was minimised through manual adjustment of the amplifier gain to within ±6 dB RMS. Following this procedure, correction values were recorded and the output to each loudspeaker was adjusted digitally at playback to match output level across the array.

Test materials

Three classes of stimuli make up the CNIT: *non-target words* (monosyllabic, non-numeric words), *target words* (monosyllabic numbers), and *background interferers* (multi-talker babble). The long-term average spectrum for all three classes of test materials is displayed in Figure 1(A). Background interferers were made up of eight separate non-English, turn-taking conversations (Spanish, Italian, Hungarian, French, Japanese, German, Chinese and Danish) between one male and one female. The use of non-English interferers was chosen to avoid potential effects of informational masking (Rhebergen, Versfeld, and Dreschler 2005; Van Engen and Bradlow 2007); however, considering the simultaneous mixture of eight background talkers is generally considered an energetic masker (Miller 1947; Simpson and Cooke 2005; Rosen et al. 2013), it is conceivable that future iterations of the CNIT would not require the same 8-talker babble. Each conversation recording ranged in duration between 45 and 65 s, and when necessary, were looped independently to fulfil the duration requirements of the test. Masker recordings were provided by corporate partners. Non-target words were taken from the recorded speech corpus detailed above, not including the numerals. Target words were the spoken numerals, "one" through "ten", but not including "seven". In the present feasibility experiments, target words and non-target words were always from the same male talker, M1.

General presentation and procedure

All eight maskers were presented simultaneously and randomly assigned on a per-trial basis to positions of ±15°, ±75°, ±105°, ±165° azimuth relative to the nose of the listener. Overall combined background level was fixed at 70 dB SPL. Non-target and target words were presented from a single loudspeaker position as a non-overlapping stream. Speaker position for the speech stream could vary depending on the experiment. For example, in Experiment 2 below, the speech stream begins at a lateral position and switches location halfway through a given trial. Figure 1(B) displays an example trial (used in Experiment 2) by highlighting the loudspeaker position (in degrees) for each masker

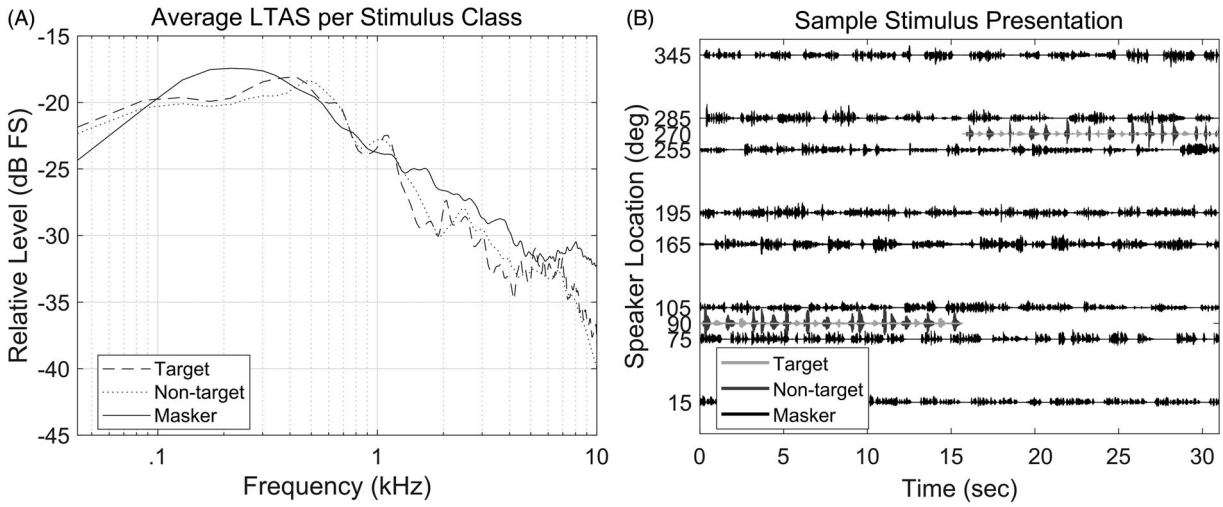

Figure 1. (A) Long-term average spectra of targets, non-targets, and background maskers. (B) Stimulus example from Experiment 2 in which the speech stream switches from the loudspeaker at 90° right-of-centre to 90° left-of-centre (RL). Target words are represented in light grey, non-targets are represented in medium grey, and masker conversations are represented in black. Non-targets were presented at +6 dB SNR relative to the overall level of the maskers, whereas targets were presented at a different SNR relative to the background (e.g., −3 dB).

stimulus and the single speech stream. Maskers consisted of turn-taking between a male and female speaker (black lines) in non-English languages, and the speech stream consisted of target digits (light grey) and non-target words (medium grey). Target and non-target words were chosen randomly with replacement and presented with an inter-stimulus interval (ISI) of 100 ms, an average of two words per second. Due to the large size of the non-target corpus and randomised presentation, there was an assumption that stimulus intelligibility was consistent on average. On average, a target word was selected every fourth word (minimum of at least two non-targets between targets and a maximum of four words between). Non-target and target word intensities were independent of one another, such that the non-target words were fixed at 76 dB SPL (+6 dB SNR *re* background maskers), and target words were allowed to vary adaptively (see Adaptive Test Procedure) or were held at a single fixed, listener-dependent SNR *re* background makers (see Fixed-Level Test Procedure). The advantage of having two independent SNRs for target and non-target words was that the more-frequent non-target words were at an intensity that could engage any hearing aid's advanced processing, while the often lower intensity target words could probe behavioural performance away from ceiling effects. This feature was a key criterion as laid out above (see Introduction). It is important to note, that due to the recording methods used to acquire the speech materials, as well as a constant ISI in the speech stream, the test materials do not aim to mirror natural prosody in speech; rather, the materials are intended to probe singular moments of speech identification.

Adaptive test procedure

An adaptive testing procedure was designed for two reasons: (1) to acquire a quick speech-in-noise threshold; and (2) to acquire a performance baseline for subsequent fixed-level testing. For the adaptive test procedure, target and non-target words were presented from the front facing (0°) speaker. Non-target words were fixed in SNR re background (+6 dB) while target words adaptively changed SNR starting at the same level as the non-target words. The adaptive test estimated the target word presentation level corresponding to 79.4% correct using a 3-down, 1-up staircase procedure (Levitt 1971). The initial step size was 5 dB; after two reversals, step size was reduced to 2 dB, and after two more reversals, the step size was reduced to 1 dB. Termination of the test occurred after the eighth reversal. Threshold was taken as the average of the final 4 of 8 reversals. On average, each threshold estimate took less than 4 min. Three tracks were completed and thresholds were averaged to provide the final estimated threshold value.

Fixed-level test procedure

The fixed-level test was distinct from the adaptive test in that target words were fixed throughout a set of trials at a single presentation level, typically guided by the initial adaptive test. The primary purpose of the fixed-level test was to take advantage of the multiple time scales that the CNIT affords the experimenter – another key goal of the test (see Introduction). That is, because a target is presented every 2 s, on average every four words, speech identification can be tracked over time with a resolution of one sample every 2 s. Analysis can focus on overall behavioural performance averaged by block or session, as many other clinical tests of speech in noise are presently constructed (e.g., HINT; Nilsson, Soli, and Sullivan 1994), or if more granular information is warranted, analysis can be broken down by smaller temporal windows (e.g., every 2 s) across multiple blocks. This is achieved by limiting the duration of a block of trials to a short period (e.g., 30 s) and presenting multiple blocks of trials (e.g. 10) per testing condition. These values can be scaled up or down to accommodate power analyses and specific hypothesis testing. Below, Experiments 2 and 3 take advantage of the temporal precision of the CNIT to investigate the near instantaneous behavioural effect of changing target location (Experiment 2) or target SNR (Experiment 3).

Scoring performance on CNIT

Whether the CNIT is administered in its fixed-level design or with the adaptive variant, participants have the same task and method of response. As each target word is presented, participants identify the target number and respond on a keypad. The keypad was a graphical user interface controlled in MATLAB and presented via touchscreen monitor (GeChic 1503I)

positioned at lap level to the participant. Scoring was performed by evaluating whether the correct number was selected by the listener in a window following the offset of the target word. For the adaptive test, a response was considered correct if it occurred between the offset of the target word and the end of the temporal bin. Given potential limitations with this scoring method, post-hoc analyses were performed on the data from the fixed-level tests and data were scored as follows. Histograms of response behaviour ($n = 10$) were constructed for many target SNRs, from which the probability density functions (PDFs) were computed for a response to a target or a non-target word (Figure 2). The difference in PDFs for responses following target and non-target words were used to compute the log likelihood of a response given a target word relative to the background response rate. This revealed a window from 0.625 to 1.25 s following the offset of the target word in which a response was most likely to be the related to that target word. This window was used to score responses in the fixed-level tests as either correct or incorrect for each target presented. Individual responses were grouped into proportion correct scores in three ways: the proportion correct within a single block, across blocks in each 2-s temporal bin, or across all trials and all temporal bins. Each of these scoring methods provide answers to particular questions that will be addressed in the experiments below.

Demonstration of feasibility of the CNIT

Participants

There were a total of 10 young, normal-hearing participants (8 F, 2 M) ranging in age from 19 to 28 years ($\mu = 23.5$; s.d. $= 2.9$). Normal hearing was defined as having pure tone air- and bone-conduction threshold less than or equal to 25 dB at octave frequencies between 0.25 and 8 kHz. Other exclusion criteria were excessive cerumen, compromised middle ear system, or a score of less than 26 on the Montreal Cognitive Assessment (MoCA; Nasreddine et al. 2005) screening instrument. Written informed consent was provided as approved by the university Institutional Review Board. Participants were compensated for their time.

Feasibility procedure

Three feasibility experiments were conducted with the same listeners. Testing for each participant spanned two sessions that were no longer than two hours each. Prior to the start of Experiment 1, the adaptive CNIT was administered to identify the appropriate range of stimulus levels for the psychometric function as well as the fixed target SNR for Experiments 2 and 3. Listeners were instructed to maintain head position fixated at the front speaker at all times.

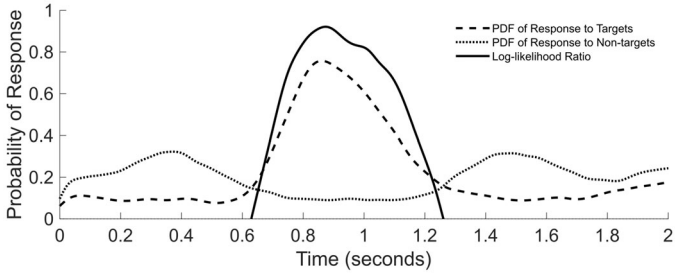

Figure 2. Probability density functions for a response to targets (dashed line) and non-targets (dotted line), and the subsequent log-likelihood difference for probabilities above zero. The range of responses likely to represent response to a target word is indicated by the zero-crossing of the log-likelihood curve (0.625–1.25 s).

In Experiment 1, a constant stimulus method was used to generate a psychometric function characterising the performance of each individual on the CNIT. Individual psychometric functions were obtained using a range of SNRs above and below the participant's adaptive threshold (+4, +2, 0, −2, −4, −6, and −8 dB relative to threshold SNR) with the speech stream coming from the front loudspeaker. For each SNR (seven total), listeners received eight blocks with each block including 16 targets and spanning a duration of 32 s. Across the eight blocks, a total of 128 target words were presented, and testing took close to 4.25 min per SNR.

The two subsequent experiments assessed the temporal sensitivity to either a change in target speaker location (Experiment 2) or target SNR (Experiment 3). In Experiment 2, listeners were tested on two dynamic conditions in which the speech stream started at one location in the free field and then abruptly shifted 180°. In one condition, the speech stream was presented from a loudspeaker at 270° azimuth relative to the listener then after roughly 16 s switched to 90° azimuth (condition LR). In the second condition, the reverse was tested (condition RL). There were a total of 16 blocks for each of these two conditions. Signal levels were fixed to the intensity of the adaptive threshold measured earlier.

Experiment 3 investigated the sensitivity of the CNIT to changes in SNR of the targets. The speech stream was always presented from the front-facing loudspeaker. Practically, such SNR changes could be the result of changing signal or background level, as well as for aided listeners, the engagement of effective noise reduction technology such as directional microphones, etc. The SNR was manipulated in two separate fashions, either one large step or three consecutive smaller steps. In the first, the SNR was reduced (worsened) by 6 dB after the midpoint of the run (i.e., roughly 16 s). In the second condition, the SNR was reduced in three consecutive steps of 2 dB with the first step occurring after the midpoint (after 16 s) and each additional step occurring 2 and 4 s later. There were a total of 16 blocks for each condition, and the initial SNR was set to the individual's adaptive threshold measured earlier.

Results

Experiment 1: psychometric function

Individual and mean psychometric functions are plotted in separate panels of Figure 3. To test whether the adaptive track was indeed estimating 79.4% identification, the adaptive CNIT thresholds (plotted in Figure 3 as circles instead of x's) were submitted to means testing with a hypothesised value of 0.794. The mean proportion of correctly identified numbers in the fixed-level CNIT using the SNR derived from the adaptive CNIT was 0.76 (median = 0.78), which was not statistically different from the hypothesised result ($t[9] = -1.06$, $p = 0.32$).

Each psychometric function was fit with a four-parameter logistic function (solid black lines in Figure 3; Prins and Kingdom 2018) described by

$$y = \gamma + \frac{1-\gamma-\lambda}{1+e^{-\beta(x-\alpha)}} \quad \text{(Eq. 1)}$$

where α is the intercept coefficient, β is the slope, γ represents the guess-rate, and λ represents the lapse rate. The latter two

coefficients were fixed to 0.11 (chance level) and 0, respectively, while the first two terms were free parameters. From visual inspection of Figure 3, the individual data are well-characterized by the monotonic logistic fits, with slope coefficients ranging 0.26–0.44, and a mean of 0.34. Figure 3 also shows each of the fitted logistics (grey curves) as well as the resulting curve when the mean of the fitted parameters is computed (black curve) in the far right panel. When translated to percent correct, the peak slope of the mean function was 7.5%/dB at 50% proportion correct. Further analyses were conducted which showed that on a per-digit basis, slope of the psychometric curve ranged from 4.1%/dB to 10.3%/dB with a mean of 7.5%/dB and standard deviation of 1.9%/dB.

Experiment 2: effect of abrupt changes to target location

Trials were analysed in 2-s bins because a target was presented every 2 s on average. In Figure 4, the proportion of correctly identified numbers is presented per temporal bin for each condition (LR: dark grey; RL: light grey). Note that after the eighth bin, the switch in location occurred, such that any effect of switching location would be expected to occur after the eighth bin and as early as at the ninth bin (dashed box). It is clear, however, that switching location did not affect the identification performance overall. In a two-way repeated-measures ANOVA with factors of condition (2 levels) and temporal bin (16 levels), neither the effect of condition nor the interaction between bin and condition was significant, indicating the direction of the switch did not play a role in the results. A main effect of bin was found ($F[1, 16] = 6.76$, $p = 0.02$; $\eta_p^2 = 0.30$); however, this was entirely driven by the low scores seen in the final (16th) bin, mostly related to fewer responses to targets occurring late in the final 2 s. In all, Bonferroni-corrected posthoc measures indicated that the 16th bin significantly differed from 9 of the other 15 bins, and approached significance on three others. The eighth and ninth bins, on the other hand, had a mean difference of 0.008 ($p = 1.00$), and no other bins were significantly different from one another.

Experiment 3: effect of abrupt changes to SNR

As in Experiment 2, the proportion of digits correctly identified was averaged across trials for each listener and analysed in 2-s bins. Figure 5 displays the overall results. The regions of interest are highlighted in the dashed boxes: (1) the first transition after the eighth bin, and (2) the subsequent SNR changes in the second condition from the 9th to the 10th to the 11th bin. In the 1-step condition (dark grey bars), it is clear that the 6-dB attenuation led to an overall decrease in performance from a mean of 0.85 to 0.5 (tested using *t*-tests and corrected for repeated measures [$\alpha = 0.025$]; $t[9] = 7.2$, $p < 0.001$). In the three-step condition (light grey bars), a predictable reduction in performance is observed at each −2-dB change in SNR from 0.83 to 0.45 by the 11th bin. At the 11th bin, both conditions had stimuli presented at −6 dB relative to the adaptive threshold, and no statistical difference was measured ($t[9] = .098$, $p = 0.35$) between conditions for that bin. For the three-step condition, posthoc measures (paired-samples *t*-tests per comparison) indicated that there was not a significant difference in performance between the eighth and ninth bin ($p = 0.70$); however, this is not unsurprising given the psychometric functions presented earlier. That is, the steepest portion of the psychometric function occurs more than 2 dB below the participant's adaptive threshold. There was however a statistically significant difference in performance between the 9th and 10th bin ($p = 0.01$; Bonferroni correct for three comparisons [$\alpha = 0.016$]) and the 10th and 11th bin only tended towards significance when corrected for multiple comparisons ($p = 0.03$).

Discussion

The CNIT is a speech-in-noise test that was developed for the purposes of assessing speech perception performance that is dynamic and that operates at relatively high SNRs. One potential use of the CNIT is for assessing hearing instrument processing, and future work will be geared towards assessing the psychometric properties of the CNIT for hearing-impaired listeners with and without hearing instruments. Standard speech-in-noise tests, such as the HINT (Nilsson, Soli, and Sullivan 1994), WIN (Wilson 2003), and BKB-SIN (Bench, Kowal, and Bamford 1979), have their limitations that prevent a full evaluation of the benefits of aided listening and are susceptible to issues of validity given the operational SNRs (e.g., Naylor 2016). The properties of the CNIT allow for longitudinal evaluation without exhausting finite speech materials, and it measures behaviour at short and long temporal scales. If it were adopted for use in clinical populations, it would have the advantage of engaging hearing instruments at suprathreshold levels while also avoiding floor or ceiling effects when measuring speech-in-noise performance.

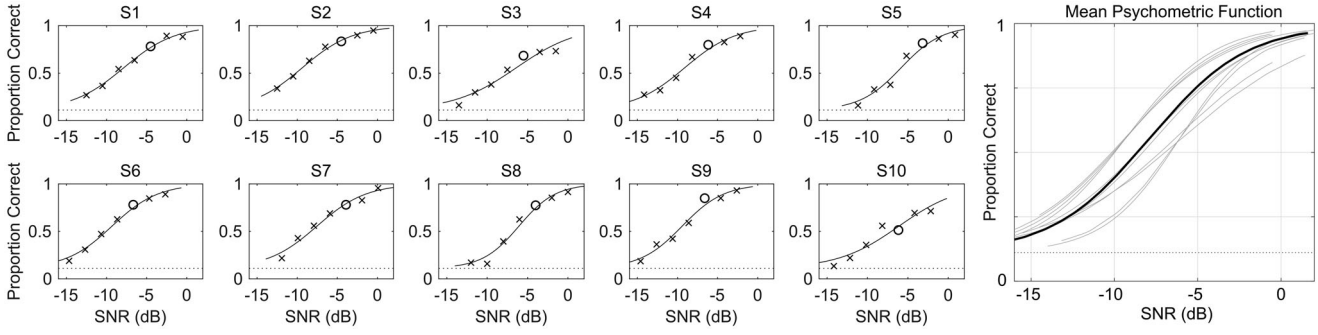

Figure 3. Proportion correct identification of target digits as a function of SNR relative to background for each of the ten subjects is shown in the small panels. The mean psychometric curves per subject (grey curves) and the aggregate mean logistic function (black curve) is shown in the far right panel. Subject's initially received the adaptive CNIT to determine threshold (indicated by a circle) before being tested at each of 7 SNRs −8 to +4 dB in 2-dB steps (indicated by x's excluding 0 dB) relative to the adaptive threshold. The solid line indicates the 4-parameter fitted logistic function as described by Prins and Kingdom (2018). Dotted horizontal lines indicate chance level (0.11).

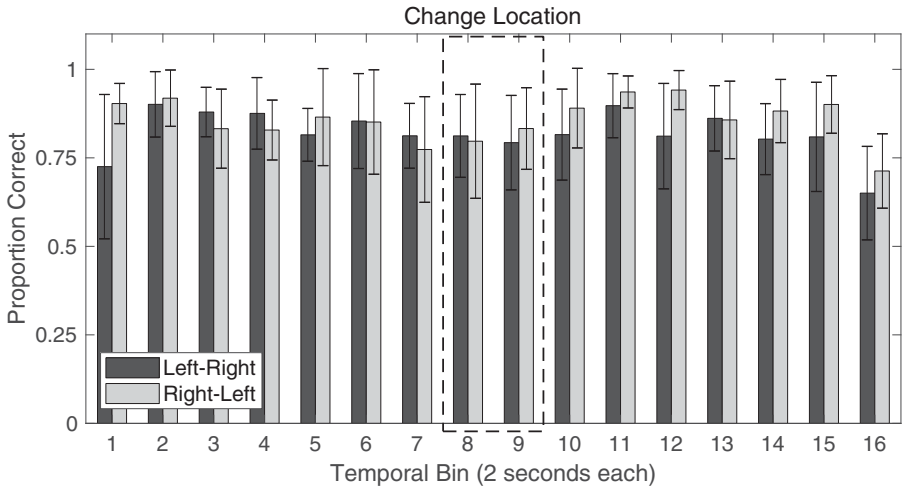

Figure 4. Proportion of correct scores per 2-s temporal window across multiple blocks. In Experiment 2, the location of the speech stream switched from right-to-left (RL; grey bars) or left-to-right (LR; black bars) after 16 s into the 32-s block. The dashed box indicates the transition interval, and shows no change in performance after the switch. Error bars are in standard deviation.

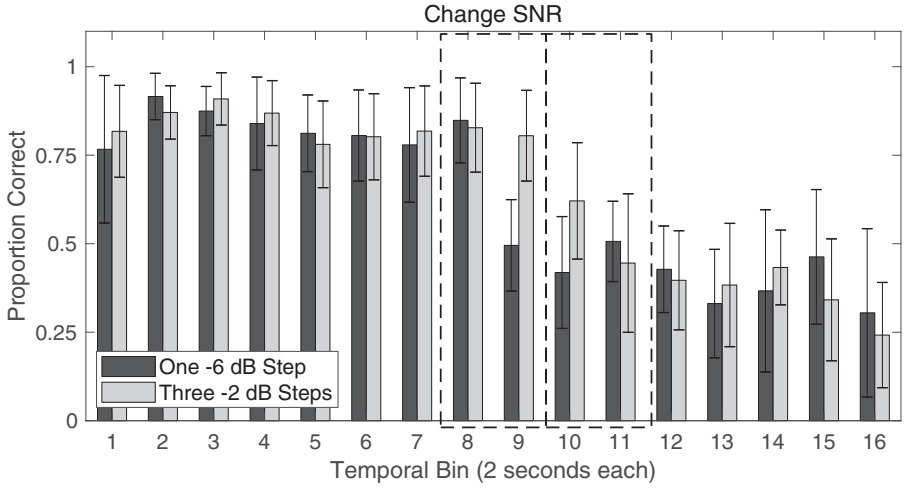

Figure 5. Proportion of correct scores per 2-s temporal window across multiple blocks. In Experiment 3, the SNR of the target words either abruptly changed by −6 dB (black bars) or change by −2-dB steps (grey bars) in three consecutive 2-s intervals. Performance changes were observed in both the single-step and three-step conditions, providing evidence for the temporal sensitivity of the CNIT. Error bars are in standard deviation.

In the feasibility phase of development, Experiment 1 revealed that the CNIT produces a shallow monotonic psychometric function (7.5%/dB) as illustrated not only by mean data, but also at the individual level (Figure 3). With respect to other common clinical tests with multi-talker background babble, the average slope of the psychometric functions in the present study was shallower than digit pair and digit triplet identification (e.g., 9.1%/dB and 10.2%/dB, respectively; Wilson, Burks, and Weakley 2005), and it was shallower than the BKB-SIN (11.9%/dB) and the Quick-SIN (10.8%/dB) but slightly steeper than the WIN (6.3%/dB; slopes for each test found in Wilson, McArdle, and Smith 2007), as well as digit triplets in babble (6.5%/dB; McArdle, Wilson, and Burks 2005). Nevertheless, the slope of the function is not widely different from other standard speech-in-noise tests, which is promising if it is to be adopted for future speech-in-noise evaluation.

When choosing among speech-in-noise tests, it may be important for the test to differentiate populations of listeners, such as normal-hearing (NH) and hearing-impaired (HI) listeners. Wilson, McArdle, and Smith (2007) showed that, among the tests mentioned above, the WIN produced the greatest separation between those groups (10.1 dB SNR), followed by the Quick-SIN (7.9 dB SNR), the HINT (5.6 dB SNR), and the BKB-SIN (4.2 dB SNR). If the corpus is a closed set and avoids potential factors other than audibility (e.g., cognition, memory, etc.), then tests such as the WIN and the CNIT should have better success at showing differences between NH and HI listeners.

Experiment 2 was originally included to provide benchmark data for an ongoing investigation of spatial hearing systems used in hearing aids (i.e., directional microphones or "beamforming"; Amlani 2001). The two conditions reported here demonstrate that the change in location, per se, does not lead to a change in performance, at least at the 2-s resolution that the CNIT can measure. On the one hand this was somewhat surprising, but is likely explained by the favourable SNRs and NH listening group. The selected data reported here only pertains to the left and right target location change and vice versa. Performance for other switch pairs remains to be tested, though one could view the current conditions, at the extremes, as being a worse-case scenario.

Finally, the goal of signal processing features such as those involving directional microphones or digital noise reduction features are to improve SNR. In natural environments, for example, adaptive directional microphone technology can improve SNR by 2–3 dB (Ricketts 2005; Ricketts and Hornsby 2006). Experiment 3 demonstrated that the CNIT can be sensitive to 2-dB changes in SNR within the 2-s temporal frame, but there are caveats to this observation which rely on the slope of the underlying psychometric function. There are two ways intensity sensitivity could change with different stimuli or listener groups. First, whereas the adaptive procedure appears to be capable of locating a consistent percent correct point for normal hearing listeners, it may be less accurate for different maskers or listener groups with shallower psychometric functions. Second, more trials can be used to change the Type II error rate for tasks with a smaller change in intensity. At the 2-dB intensity resolution, the CNIT may be able to assess expected benefit from hearingaid signal processing features such as directional microphone systems and other SNR-improving algorithms, which is not possible at short time scales in common sentence-based tests. Sentence-based tests also contain potential memory confounds (e.g., recency and primacy), and it is difficult to assess *when* performance deviates due to external or internal changes. As with the other experiments, further work must be completed in HI listeners to demonstrate similar resolutions.

As stated above, the motivation for the development of the CNIT was to create a test that was capable of providing meaningful comparisons during aided speech identification, and to avoid potential flaws associated with the presently available tests, which are often threatened by internal or external validity issues (Naylor 2016). Internal validity, or the potentially variable effects of hearingaid processing at various SNRs, is mitigated in the CNIT because of the constant and dominant SNR between the non-target speech and background maskers. That is, if comparing across hearing aids, the benefits or disadvantages of device-specific processing will not vary as the adaptive CNIT titrates on the SNR of the target words. External validity, or the potentially misguided effects at unrealistic SNRs, is mitigated again by the constant and favourable SNR between the non-targets and the background. Though the target words are usually attenuated and reach uncharacteristic SNRs from the "real world," the hearing aid processing would be tested in their natural state due to the favourable SNR of the non-targets. Finally, Naylor warns of the inherent difference among participants and their SNR requirements, and again, the CNIT can overcome issues related to individual variability because of the consistent operating range of the dominant audio in the test (i.e., the non-target words fixed at +6 dB SNR).

Some further caution is worth considering. Though this test was developed for future use with aided HI listeners, there are potential pitfalls associated with hearing instruments and clinical populations that may interact with the CNIT's current design and results of the feasibility tests which used only NH listeners. First, certain instrument processing may address the occurrence of having soft sounds interspersed with otherwise louder speech signals, so it is important to follow up with some reference signal processing algorithms to determine whether device-specific approaches are comparable. Second, though digit tasks tend not to be cognitively demanding, it remains to be seen whether cognitive factors such as processing speed or executive function will differentially affect elderly listeners with or without hearing loss. Third, the non-effect of spatial changes seen in Experiment 2 may not be observed with HI and/or elderly listeners. Finally, though the CNIT can be run efficiently and scaled to potentially short time windows, it nevertheless makes use of resources not always found in a clinic and stable measures may require more time than a clinician has available; therefore, the CNIT in its current form is only intended as a laboratory research tool.

Summary and conclusions

A new speech-in-noise test was developed and evaluated for feasibility. The test was originally developed to meet a practical need to measure the impact of hearing instrument digital signal processing on speech intelligibility. No existing tests were available that could probe behaviour at multiple temporal scales, including every 2 s, and also engage appropriate processing by presenting independent device and listener SNRs. Experiment 1 demonstrated that the CNIT has a comparable psychometric behaviour to standard clinical speech in noise tests. Experiment 2 showed that instantaneous changes in target position do not result in a significant change in performance on the CNIT task, which may demonstrate the potential for assessing spatial hearing algorithms. Finally, Experiment 3 demonstrated that the CNIT is sensitive to sudden changes in SNR as small as 2 dB under the right circumstances. This sensitivity to small SNR changes indicates that this test could be suitable for use in assessing dynamic hearing instrument signal processing features, which operate on a similar intensity scale. Future studies will aim to measure CNIT feasibility in older listeners with normal hearing and with hearing impairment, both aided and unaided.

Acknowledgements

EJO, ECH, and DAE contributed equally to the design, development, and implementation of the test. ECH and PG performed the experiments and analysed the data. EJO wrote the paper with significant contributions from PG, ECH, and DAE. Aspects of development were inspired by conversations with Dr. Rahul Shrivastav and subsequent conversations with Drs. David Pisoni and Richard Wilson. We appreciate technical expertise in the recording process provided by Dr. Luke Wasserman.

Disclosure statement

No potential conflict of interest was reported by the authors.

Funding

This work was supported in part by funding from Unitron, Inc., and the Sonova Corporation.

ORCID

Erol J. Ozmeral http://orcid.org/0000-0001-9529-001X
Eric C. Hoover http://orcid.org/0000-0002-1187-9925

References

Amlani, A. M. 2001. "Efficacy of Directional Microphone Hearing Aids: A Meta-Analytic Perspective." *The Journal of the American Academy of Audiology* 12 (4): 202–214.

Balota, D. A., M. Pilotti, and M. J. Cortese. 2001. "Subjective Frequency Estimates for 2,938 Monosyllabic Words." *Memory & Cognition* 29 (4): 639–647. doi:10.3758/BF03200465.

Bench, J., A. Kowal, and J. Bamford. 1979. "The BKB (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children." *British Journal of Audiology* 13 (3): 108–112. doi:10.3109/03005367909078884.

Clopper, C. G., D. B. Pisoni, and A. T. Tierney. 2006. "Effects of Open-Set and Closed-Set Task Demands on Spoken Word Recognition." *Journal of the American Academy of Audiology* 17 (5): 331–349. doi:10.3766/jaaa.17.5.4.

European Broadcast Union (EBU). 2014. *R128: Loudness Normalization and Permitted Maximum Level of Audio Signals* Geneva.

Houben, R., J. Koopman, H. Luts, K. C. Wagener, A. van Wieringen, H. Verschuure, and W. A. Dreschler. 2014. "Development of a Dutch Matrix Sentence Test to Assess Speech Intelligibility in Noise." *International* Journal of Audiology 53 (10): 760–763. doi:10.3109/14992027.2014.920111.

Kochkin, S. 2009. "MarkeTrack VIII: 25-Year Trends in the Hearing Health Market." *Hearing Review* 16: 12–31.

Levitt, H. 1971. "Transformed Up-Down Methods in Psychoacoustics." *The Journal of the Acoustical Society of America* 49 (2B): 467–477. doi:10.1121/1.1912375.

Lobdell, B. E., and J. B. Allen. 2007. "A Model of the VU (Volume-Unit) Meter, with Speech Applications." *The Journal of the Acoustical Society of America* 121 (1): 279–285. doi:10.1121/1.2387130.

McArdle, R. A., R. H. Wilson, and C. A. Burks. 2005. "Speech Recognition in Multitalker Babble Using Digits, Words, and Sentences." *Journal of the American Academy of Audiology* 16 (9): 726–739. quiz 763-724. doi:10.3766/jaaa.16.9.9.

Miller, G. A. 1947. "The Masking of Speech." *Psychological Bulletin* 44 (2): 105–129. doi:10.1037/h0055960.

Nasreddine, Z. S., N. A. Phillips, V. Bédirian, S. Charbonneau, V. Whitehead, I. Collin, J. L. Cummings, and H. Chertkow. 2005. "The Montreal Cognitive Assessment, MoCA: A Brief Screening Tool for Mild Cognitive Impairment." *Journal of the American Geriatrics Society* 53 (4): 695–699. doi:10.1111/j.1532-5415.2005.53221.x.

Naylor, G. 2016. "Theoretical Issues of Validity in the Measurement of Aided Speech Reception Threshold in Noise for Comparing Nonlinear Hearing Aid Systems." *Journal of the American Academy of Audiology* 27 (7): 504–514. doi:10.3766/jaaa.15093.

Nilsson, M., S. D. Soli, and J. A. Sullivan. 1994. "Development of the Hearing in Noise Test for the Measurement of Speech Reception Thresholds in Quiet and in Noise." *The Journal of the Acoustical Society of America* 95 (2): 1085–1099. doi:10.1121/1.408469.

Prins, N., and F. A. A. Kingdom. 2018. "Applying the Model-Comparison Approach to Test Specific Research Hypotheses in Psychophysical Research Using the Palamedes Toolbox." *Frontiers in Psychology* 9: 1250. doi:10.3389/fpsyg.2018.01250.

Rhebergen, K. S., N. J. Versfeld, and W. A. Dreschler. 2005. "Release from Informational Masking by Time Reversal of Native and Non-Native Interfering Speech." *The Journal of the Acoustical Society of America* 118 (3): 1274–1277. doi:10.1121/1.2000751.

Ricketts, T. A. 2005. "Directional Hearing Aids: Then and Now." *The Journal of Rehabilitation Research and Development* 42 (4 Suppl 2): 133–144.

Ricketts, T. A., and B. W. Hornsby. 2006. "Directional Hearing Aid Benefit in Listeners with Severe Hearing Loss." *International Journal of Audiology* 45 (3): 190–197. doi:10.1080/14992020500258602.

Rosen, S., P. Souza, C. Ekelund, and A. A. Majeed. 2013. "Listening to Speech in a Background of Other Talkers: Effects of Talker Number and Noise Vocoding." *The Journal of the Acoustical Society of America* 133 (4): 2431–2443. doi:10.1121/1.4794379.

Simpson, S. A., and M. Cooke. 2005. "Consonant Identification in N-Talker Babble is a Nonmonotonic Function of N." *The Journal of the Acoustical Society of America* 118 (5): 2775–2778. doi:10.1121/1.2062650.

Smeds, K., F. Wolters, and M. Rung. 2015. "Estimation of Signal-to-Noise Ratios in Realistic Sound Scenarios." *Journal of the American Academy of Audiology* 26 (2): 183–196. doi:10.3766/jaaa.26.2.7.

Van Engen, K. J., and A. R. Bradlow. 2007. "Sentence Recognition in Native- and Foreign-Language Multi-Talker Background Noise." *The Journal of the Acoustical Society of America* 121 (1): 519–526. doi:10.1121/1.2400666.

Wilson, R. H. 2003. "Development of a Speech-in-Multitalker-Babble Paradigm to Assess Word-Recognition Performance." *Journal of the American Academy of Audiology* 14: 453–470. doi:10.3766/jaaa.16.8.11.

Wilson, R. H., C. A. Burks, and D. G. Weakley. 2005. "A Comparison of Word-Recognition Abilities Assessed with Digit Pairs and Digit Triplets in Multitalker Babble." *The Journal of Rehabilitation Research and Development* 42 (4): 499–510. doi:10.1682/JRRD.2004.10.0134.

Wilson, R. H., R. A. McArdle, and S. L. Smith. 2007. "An Evaluation of the BKB-SIN, HINT, QuickSIN, and WIN Materials on Listeners with Normal Hearing and Listeners with Hearing Loss." *Journal of Speech, Language, and Hearing Research* 50 (4): 844–856. doi:10.1044/1092-4388(2007/059).

Wu, Y. H., E. Stangl, O. Chipara, S. S. Hasan, A. Welhaven, and J. Oleson. 2018. "Characteristics of Real-World Signal to Noise Ratios and Speech Listening Situations of Older Adults with Mild to Moderate Hearing Loss." *Ear Hear* 39 (2): 293–304. doi:10.1097/AUD.0000000000000486.