International Journal of Audiology

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

Evaluation of the NAL Dynamic Conversations Test in older listeners with hearing loss

Virginia Best, Gitte Keidser, Katrina Freeston & Jörg M. Buchholz

To cite this article: Virginia Best, Gitte Keidser, Katrina Freeston & Jörg M. Buchholz (2017): Evaluation of the NAL Dynamic Conversations Test in older listeners with hearing loss, International Journal of Audiology, DOI: 10.1080/14992027.2017.1365275

To link to this article: http://dx.doi.org/10.1080/14992027.2017.1365275

Published online: 21 Aug 2017.

Taylor & Francis
Taylor & Francis Group

International Journal of Audiology 2017; Early Online: 1–9

International Journal of Audiology

Original Article

Evaluation of the NAL Dynamic Conversations Test in older listeners with hearing loss

Virginia Best[1], Gitte Keidser[2], Katrina Freeston[2] & Jörg M. Buchholz[2,3]

[1]Department of Speech, Language and Hearing Sciences, Boston University, Boston, MA, USA, [2]National Acoustic Laboratories, Sydney, NSW, Australia, and [3]Department of Audiology, Macquarie University, Sydney, NSW, Australia

Abstract

Objective: The National Acoustic Laboratories Dynamic Conversations Test (NAL-DCT) is a new test of speech comprehension that incorporates a realistic environment and dynamic speech materials that capture certain features of everyday conversations. The goal of this study was to assess the suitability of the test for studying the consequences of hearing loss and amplification in older listeners. *Design:* Unaided and aided comprehension scores were measured for single-, two- and three-talker passages, along with unaided and aided sentence recall. To characterise the relevant cognitive abilities of the group, measures of short-term working memory, verbal information-processing speed and reading comprehension speed were collected. *Study sample:* Participants were 41 older listeners with varying degrees of hearing loss. *Results:* Performance on both the NAL-DCT and the sentence test was strongly driven by hearing loss, but performance on the NAL-DCT was additionally related to a composite cognitive deficit score. Benefits of amplification were measurable but influenced by individual test SNRs. *Conclusions:* The NAL-DCT is sensitive to the same factors as a traditional sentence recall test, but in addition is sensitive to the cognitive factors required for speech processing. The test shows promise as a tool for research concerned with real-world listening.

Key Words: Speech comprehension, realistic tests, hearing loss, hearing aids

Introduction

Kiessling et al. (e.g. 2003) put forward a useful terminology to describe the hierarchy of auditory functions relevant to people in their daily lives. They distinguish *hearing*, a passive function by which we sense sounds and their properties, from *listening*, which is hearing with intention and attention. They then define extensions of hearing and listening in which information moves unidirectionally (*comprehending*) or bidirectionally (*communication*). In the same paper, Kiessling et al. also consider how hearing loss and other consequences of aging affect this cascade from hearing to communication. It is clear that deficits in the peripheral auditory system affect hearing, which can in turn make listening more difficult, and ultimately impair comprehension and communication. In addition, more central deficits may directly affect one's ability to listen selectively, assign meaning to speech in a limited time frame, or integrate the many cues provided by a conversational partner that enable smooth communication.

The ultimate goal of hearing rehabilitation is to enable those with hearing problems to communicate efficiently in real life. Unfortunately, we currently have no behavioural test to obtain a reliable measure of how effectively a person can communicate with (or without) amplification. Most audiological tests assess hearing, and to a lesser extent listening. We recently turned to the next step in the hierarchy, and developed a laboratory-based test to assess the ability of a person to comprehend speech under realistic conditions. This test incorporates natural conversational speech comprised of one, two or three target talkers and is implemented in a simulated noisy room containing competing conversations. The development of this test (the National Acoustic Laboratories Dynamic Conversations Test or NAL-DCT) is described in detail in a previous paper (Best et al. 2016). Also described in that paper are the results of a study in 30 listeners with normal hearing, which provided normative data and illustrated the basic psychometric properties of the test.

In the current paper, we describe an evaluation of this test in 41 older participants with hearing loss. The main goal of the study

Correspondence: Virginia Best, Department of Speech, Language and Hearing Sciences, Boston University, 635 Commonwealth Ave, Boston, MA 02215, USA. E-mail: ginbest@bu.edu

(Received 23 November 2016; revised 26 July 2017; accepted 27 July 2017)

ISSN 1499-2027 print/ISSN 1708-8186 online
DOI: 10.1080/14992027.2017.1365275

Abbreviations

BKB	Bamford-Kowal-Bench
NAL-DCT	National Acoustic Laboratories Dynamic Conversations Test
SRT	speech reception threshold
SNR	signal-to-noise ratio
NH	normally hearing
HI	hearing-impaired
4FAHL	four-frequency average hearing loss

was to assess the suitability of the test for studying the consequences of hearing loss and amplification in this population. In this study, we also wanted to understand what new information can be gained from the NAL-DCT over that gained from currently available speech tests. To this end, we collected data on a common sentence recognition test in the same group of listeners, and focussed our analysis on differences between the two tests. One of the key differences between the NAL-DCT and the sentence test, we believe, is the increased dependence on cognitive processing in the NAL-DCT. Hence, we collected several standard cognitive measures from our listeners, to allow statistical comparisons that might help us demonstrate the reliance of the NAL-DCT on cognitive processes that are also relevant in real-world communication. These measures focussed on working memory, which is associated with speech-in-noise performance in both young and older listeners (Akeroyd 2008; Besser et al. 2013; Souza and Arehart 2015; Gordon-Salant and Cole 2016) and was shown to contribute to performance on the NAL-DCT in our previous study (Best et al. 2016), as well as the speed of language processing, which we presumed would be critical when there is a requirement to follow and comprehend speech in an ongoing way.

In addition to its potential for tapping into cognitive aspects of communication, the NAL-DCT is uniquely suited for examining dynamic aspects of real-world communication settings. In particular, it allows an assessment of whether natural variations in talker voice and location, inherent to group conversations, influence how well listeners comprehend ongoing speech. Our original prediction, based on previous studies that incorporated spatial dynamics into word- and sentence-recall tests (Best et al. 2008; Jensen et al. 2012) was that increasing the number of talkers would lead to attention-switching costs and poorer performance. However, in our previous study with NH listeners (Best et al. 2016), we found only a modest effect of number of talkers, and in fact scores for the two- and three-talker conditions were slightly *better* on average than for the single-talker condition. We interpreted that discrepancy in terms of a number of other factors in realistic conversations (simplified language, repetition, etc.) that can offset any attention-switching costs. Here we were interested in examining this question again for older listeners with HI, as there is some evidence that older listeners are less able to make use of spatial location to perceptually segregate talkers in conversations (Murphy et al. 2006).

Finally, in parallel with the NAL-DCT, we also collected subjective ratings of ''listening effort'', a concept that is hard to define but most certainly encapsulates the combined difficulties associated with hearing, listening and comprehending (e.g. see discussion in McGarrigle et al. 2014). While subjective ratings of listening effort, like speech recognition measures, are sensitive to the presence and type of noise and signal-to-noise ratio (SNR), there is some debate about whether rated listening effort simply reflects perceived speech recognition accuracy. On the one hand, there are examples demonstrating perceived differences in listening effort across conditions that did not show relative performance differences (e.g. Rudner et al. 2012; Johnson et al. 2015; van den Tillaart-Haverkate et al. 2017), and Humes (1999) found the two variables to be separate hearing aid outcome measures. On the other hand, there are examples in which ratings of effort do follow recognition scores (e.g. Feuerstein 1992; Fraser et al. 2010). The rating of listening effort was introduced in this study to examine how it relates to our measure of speech *comprehension* as the SNR and the complexity of the talker condition vary.

Methods

Participants and hearing aids

Forty-one participants (11 female, 30 male) with hearing loss were recruited from the National Acoustic Laboratories' database. They ranged in age from 60 to 80 years (mean 72 years). Four-frequency average hearing loss (4FAHL; average threshold across the two ears at 0.5, 1, 2 and 4 kHz) ranged from 12 to 65 dB HL. Hearing losses were sensorineural (air-bone gap no more than 10 dB at any of the four frequencies) and symmetric (no more than 20 dB difference between the ears at any of the four frequencies and a maximum of 13 dB difference between the ears on average). Age and hearing loss were weakly but significantly correlated ($r = 0.38$; $p = 0.02$).

The 32 participants with hearing levels suitable for hearing aid fitting (4FAHL >25 dB) were fitted with a real-time master hearing aid (custom-designed at NAL), connected to transducers mounted in behind-the-ear cases (Siemens Centra S). Custom earmoulds were used for each participant with venting appropriate to their degree of hearing loss given that the hearing aids did not have feedback cancellation. Amplification and wide-dynamic range compression thresholds were set in 16 channels according to the NAL-NL2 prescription. Compression was fast-acting with attack and release times of 10 ms and 100 ms, respectively. For verification, insertion gain was measured at 65 dB SPL using the International Speech Test Signal with the participant facing a frontally positioned loudspeaker. The match to targets was generally very good (within 1.7 dB on average, from 250 Hz to 4 kHz). During the experiment, the hearing aids were used in omnidirectional mode.

Each participant attended 3 or 4 appointments of about 2 hours each. Across these appointments, participants completed adaptive sentence-in-noise testing, comprehension and sentence-in-noise testing at fixed SNRs, and cognitive testing. Participants received a small gratuity for their participation. Treatment of participants was approved by the Australian Hearing Ethics Committee and conformed in all respects to the Australian government's National Statement on Ethical Conduct in Human Research.

Comprehension testing

The speech comprehension materials upon which the NAL-DCT is based, and details of the simulated cafeteria environment used to present the test, are described in detail in our previous paper (Best et al. 2016).

Briefly, the NAL-DCT is based around 20 single-talker monologues, 20 two-talker conversations and 20 three-talker conversations that were recorded from a group of six talkers. The talkers read from transcripts but delivered the lines in a natural way. The test

requires listeners to follow and comprehend the speech in each monologue or conversation (3–4 min on average). Before each presentation, listeners are provided with a single page containing 10 questions related to the content with space provided for answers, and have half a minute to look it over. The listener's ability to comprehend the speech is assessed via their written answers, given ''on-the-go'' at the relevant point during the passage. Written answers consist of anything from a tick in a box, to a single number or letter answer, to a short answer of a few words. The 20 passages within each talker condition are grouped into four sets of approximately equivalent difficulty (as described in Best et al. 2016), allowing up to four experimental conditions to be tested.

Testing took place in a large anechoic chamber fitted with a spherical loudspeaker array. The loudspeaker array was used to simulate a virtual room which incorporated the target speech materials (monologues and conversations) as well as a background of competing conversations. The target talkers (one, two or three) were assigned randomly to three target locations (−67.5°, 0°, or +67.5° azimuth), and LEDs on top of the relevant loudspeakers indicated which one was currently active. The background scene contained 14 masker talkers, distributed around the listener at different locations in the virtual room. They were arranged in seven two-talker pairs, with the two members of each pair taking turns in a conversation, such that there were seven active maskers at any moment in time. The background noise was presented at a fixed level of 65 dB SPL (measured in the centre of the array), and the target speech level was varied to adjust the SNR.

For each participant, three consecutive SNRs were selected from a predetermined set of 11 (−12.5 to 12.5 dB in 2.5 dB steps) with the goal of spanning the sloping portion of the psychometric function and allowing a 50% speech comprehension threshold (SCT) to be extracted. To choose appropriate SNRs for each listener, we were guided by the speech reception threshold (SRT) obtained via adaptive testing with sentence materials (see next section). If the SRT was within 1 dB of a member of the SNR set, it was chosen as the middle SNR for that listener (e.g. for an SRT of −7.3 dB the three SNRs would be −10, −7.5 and −5 dB). If not, the three were chosen so that two were above and one was below the SRT (e.g. for an SRT of 1.2 dB the three SNRs would be 0, 2.5 and 5 dB). Before formal testing began, two spare passages were presented, one at the highest test SNR to familiarise participants with the task, and one at the lowest test SNR to ensure that the speech was at least partially audible at this level. If the participant was unable to answer at least two out of 10 questions correctly at the lowest SNR then the SNR range was shifted up (this was necessary for 13 participants).

Comprehension of single-, two- and three-talker passages was measured (unaided) at these three SNRs for each participant and talker condition. Aided performance was then tested at one of these three SNRs (because only four sets of NAL-DCT passages are available per talker condition). The SNR used for aided testing was chosen for each participant such that the final set of aided SNRs used was as small as possible (and thus the number of participants at each SNR was maximised). The final set of aided SNRs was 5, 2.5 and 10 dB (with 19, 14 and 8 participants tested, respectively). The assignment of passages to conditions, and the order of testing of conditions, was randomised and balanced across participants. Performance was captured as percentage correct scores based on 50 questions per condition. At the completion of each individual passage, participants were asked to rate their perceived listening effort on a scale from 0 to 12 corresponding to ''no effort'' through ''extreme effort'' (Luts et al. 2010; Johnson et al. 2015).

Sentence-in-noise testing

Sentence-in-noise testing was done in two stages. First, adaptive sentence-in-noise testing was conducted in the multitalker background to obtain an unaided SRT for each participant. This SRT was used to inform the selection of SNRs for testing with the NAL-DCT (see previous section), and was also used as a reference for the SCTs estimated from the NAL-DCT. Speech stimuli were Bamford-Kowal-Bench (BKB) sentences spoken by a male talker, which had been modified slightly to increase the slope of the psychometric function (Best et al. 2014). The background noise was fixed at 65 dB SPL, and the target level was adapted using custom software that tracks 50% correct sentence recall using morpheme-level scoring with a maximum of 32 trials (see Keidser et al. 2013 for details). Three adaptive tracks were completed, and the values were averaged to obtain a single SRT.

In the second stage of sentence-in-noise testing, which took place after completion of the NAL-DCT, the same software and procedures were used but the target level was fixed rather than adaptive. The level was chosen for each listener to set the SNR to that used for aided NAL-DCT testing (i.e., from the set −5, 2.5 and 10 dB). This test contained 32 trials which were used to calculate percentage correct scores, and was completed once unaided and once aided (where applicable). Scores on the fixed-level test were used for direct comparisons with the equivalent unaided/aided scores on the NAL-DCT obtained at the same SNR. Different lists of sentences were used for the adaptive and fixed-level tests within a participant, and the choice of lists was random and counterbalanced across participants.

Cognitive measures

To characterise the relevant cognitive abilities of the group, measures of working memory (percent correct on the Reading Span Test; Daneman and Carpenter 1980; Rönnberg et al. 1989), verbal information-processing speed (average response time on Letter Matching and Lexical Decision tasks; Hällgren et al. 2001) and reading comprehension speed (correct answers per second on the Stanford Speed Reading Test; Karlsen and Gardener 1984) were collected. These tests were conducted in an audiometric booth fitted with a computer monitor, keyboard and mouse.

Reading Span Test: A string of sentences appeared on the monitor, one at a time. After each sentence the participant was required to say ''yes'' or ''no'' to indicate whether it was a meaningful sentence or not. Then, at the end of the string of sentences, they were asked to recall either the first word or last word of each sentence in the string. After one practice string of three sentences, participants heard three strings each of three sentences, four sentences, five sentences and six sentences. This amounted in a total of 54 sentences (and thus 54 words for recall). Scores were based on the percentage of words correctly remembered in any order.

Letter Matching and Lexical Decision tasks: In the Letter Matching task, the participant saw two letters appear on the monitor at the same time and they had to indicate using a ''yes'' and ''no'' key whether the two letters were identical or not. For the Lexical Decision task, the same keys were used to indicate whether a three-letter word presented on the monitor was a real English word or a

nonsense word. For both tasks, participants were instructed to respond as quickly as possible. Performance was calculated as the average response time for each correct answer. As the average response times for the two measures were highly correlated ($r = 0.76$; $p < 0.001$), the two measures were further averaged to produce a single measure of lexical processing speed. Furthermore, to facilitate comparisons to the other cognitive measures, the reaction times were multiplied by -1 such that higher values represent better cognitive abilities.

Stanford Speed Reading Test: Participants received a written passage about cable cars in paper form. The passage was missing a word in each of 30 sentences, and the participant's task was to choose the best of three options to fill in the gap. The task was terminated at three minutes regardless of whether the participant was finished, but if they finished earlier then the time taken was noted by the experimenter. Scores on this test were highly skewed for our participant group, with many obtaining a perfect score. However, there was a large variation in the time taken to complete the task. Thus, we divided the number of correct responses by the time taken to give a measure of reading comprehension speed (''correct answers per second''), which followed a more normal distribution.

The cognitive abilities were all significantly inter-correlated, although not strongly so (r varying from 0.34 to 0.37; and p varying from 0.02 to 0.03). A principal components factor analysis was performed, with the three cognitive measures as independent variables. This analysis identified one factor with an eigenvalue greater than one, which explained 57% of the variation in the cognitive data. Factor loadings for the three cognitive measures were all significant, and of similar magnitude (-0.76, -0.75, -0.75). For further analysis concerned with cognitive abilities, we used the resulting factor score for each participant as a more robust composite measure of ''cognitive deficit''. The factor score was significantly associated with age ($r = 0.45$, $p = 0.003$) but not 4FAHL ($r = 0.17$, $p = 0.29$).

Results

Effect of number of talkers

Figure 1 shows mean unaided performance on the NAL-DCT as a function of SNR (low, mid and high) for the three talker conditions. The figure shows the expected effect of SNR, but only minor differences between the three talker conditions. To determine which factors influenced performance, a repeated-measures ANOVA was conducted with factors of SNR (low, mid and high) and talker condition (one-, two- and three-talker) as repeated measures, and 4FAHL as a covariate. This analysis confirmed that there was a significant effect of SNR [$F(2,78) = 39.9$; $p < 0.001$], but no effect of talker condition [$F(2,78) = 1.0$; $p = 0.37$], and no interaction [$F(4,156) = 0.4$; $p = 0.83$]. The effect of 4FAHL was not significant [$F(1,39) = 3.9$; $p = 0.05$], and did not interact with SNR [$F(2,78) = 2.1$; $p = 0.14$] or talker condition [$F(2,78) = 0.3$; $p = 0.75$], suggesting that the effects of hearing loss were mitigated by the use of individualised SNR ranges. The three-way interaction was not significant [$F(4,156) = 1.5$; $p = 0.20$].

To further confirm that there was no effect of talker condition, logistic functions were fitted to the data for each participant in each talker condition. Where possible, 50% SCTs were extracted from the fits. This was not possible in a number of cases with extremely flat or nonmonotonic functions, and the seven participants affected were excluded from this analysis. Average SCTs for the one-, two- and three-talker conditions were -3.1, -2.4 and -2.7 dB, respectively. A repeated-measures ANOVA indicated that these small differences across talker condition were not significant [$F(2,66) = 5.0$; $p = 0.43$].

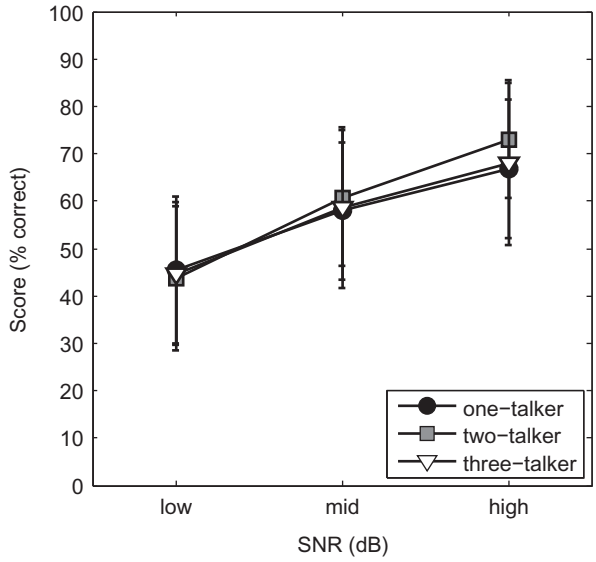

Figure 1. Mean unaided performance on the NAL-DCT as a function of SNR (low, mid, high) for the three talker conditions. Error bars depict across-subject standard deviations.

Because we found no effect of talker condition on the percent correct data or SCTs, the raw data were collapsed across talker condition for the analyses that follow. Percent correct scores therefore represent composite scores averaged over 150 questions. In the unaided condition, a single new logistic function was estimated for each participant and the SCTs extracted from these functions are used below.

Predictors of unaided performance on the NAL-DCT and the sentence test

To understand the factors driving performance on each of the two tests, regression analyses were conducted for each test using the SRT/SCT as the dependent variable, and 4FAHL, age, and the cognitive deficit factor as predictors. The analyses showed that the 4FAHL provided the only significant contribution to the SRTs (standardised coefficient = 0.89; $p < 0.001$), explaining 76% of the variance. Degree of hearing loss also made the strongest significant contribution to the SCTs (standardised coefficient = 0.75; $p < 0.001$), followed by the cognitive deficit factor (standardised coefficient = 0.34; $p < 0.001$). Combined, the two factors explained 74% of the variance; i.e. a similar proportion to that explained by degree of hearing loss alone for the sentence test.

Of primary interest in this work is to understand what listener-related factors affect performance on the NAL-DCT above and beyond what can be measured on a sentence test. To focus on this question, the difference in threshold between the two tests was calculated (SCT–SRT; with positive values meaning the comprehension task increased thresholds, i.e. produced worse performance). A regression analysis was conducted on these differences with the same predictors as above. Consistent with the observations

made above, the only significant predictor that emerged was the cognitive deficit factor (standardised coefficient =0.53; $p < 0.001$), explaining 37% of variance in the difference in performance on the two tests. This suggests that those with poorer cognitive abilities tended to perform more poorly on NAL-DCT than on the sentence test (see Figure 2). Conversely, participants with relatively good cognitive abilities tended to perform relatively better on the comprehension-style test.

Effect of amplification

Each of the 32 hearing aid candidates completed unaided and aided testing for both the NAL-DCT and the BKB sentence test at a fixed SNR. Scores are shown in Figure 3 as a function of test SNR.

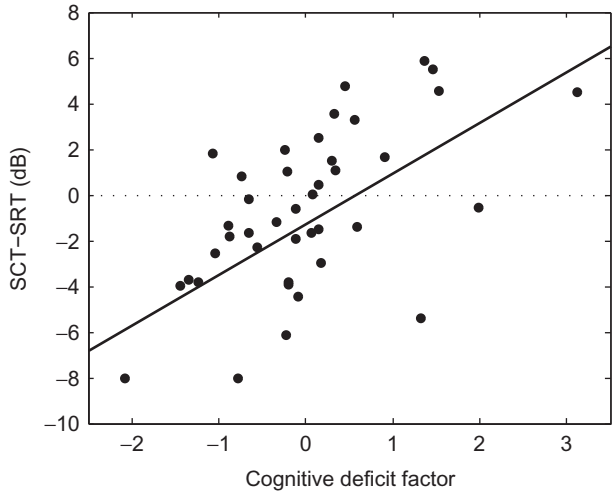

Figure 2. Scatterplots showing the difference in threshold on the two tests (SCT–SRT) as a function of the composite cognitive deficit factor.

To understand which factors influenced performance, a repeated-measures ANOVA was conducted with within-subjects factors of test (NAL-DCT and BKB) and listening mode (unaided and aided), and SNR as a between-subjects factor. This analysis revealed significant main effects of test [$F(1,29) = 4.5$; $p = 0.04$], listening mode [$F(1,29) = 6.0$; $p = 0.02$], and SNR [$F(2,29) = 34.1$; $p < 0.001$], as well as significant two-way interactions between test and SNR [$F(2,29) = 20.7$; $p < 0.001$], and between listening mode and SNR [$F(2,29) = 7.4$; $p = 0.003$]. The interaction between test and listening mode was not significant [$F(1,29) = 0.8$; $p = 0.39$], nor was the three-way interaction [$F(2,29) = 2.5$; $p = 0.10$]. The interaction between test and SNR is explained by the more compressed range of performance for the NAL-DCT compared to the sentence test. Post hoc comparisons (paired *t*-tests, using scores averaged over listening mode for each listener) indicated that NAL-DCT scores were superior to BKB scores for listeners tested at −5 dB ($p < 0.001$) but inferior for listeners tested at positive SNRs ($p = 0.005$ for 2.5 dB; $p = 0.004$ for 10 dB). The interaction between listening mode and SNR is consistent with the idea that the benefit of amplification depends on SNR. Post-hoc comparisons (paired *t*-tests, using scores averaged over test for each listener) indicated that aiding offered an advantage for listeners tested at positive SNRs that was significant for 2.5 dB ($p = 0.03$) but just shy of significance for 10 dB ($p = 0.05$). However, for listeners tested at −5 dB, aided scores were significantly *poorer* than unaided scores ($p = 0.02$).

Predictors of aided performance on the NAL-DCT and the sentence test

To understand the factors driving aided performance on each of the two tests, regression analyses were conducted for each test using the test SNR, 4FAHL, age, and the cognitive deficit factor as predictors. The analyses showed that the SNR provided the only significant contribution to the scores obtained with the sentence test (standardised coefficient = 1.15; $p < 0.001$); with 71% of the

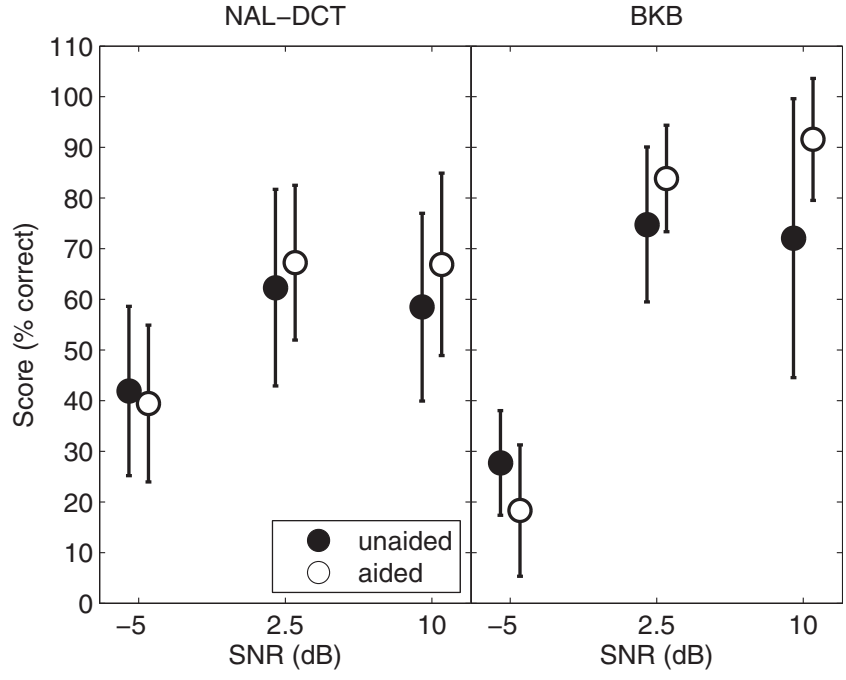

Figure 3. Unaided scores (filled symbols) and aided scores (open symbols) as a function of SNR for the NAL-DCT (left) and BKB sentence test (right). Error bars depict across-subject standard deviations.

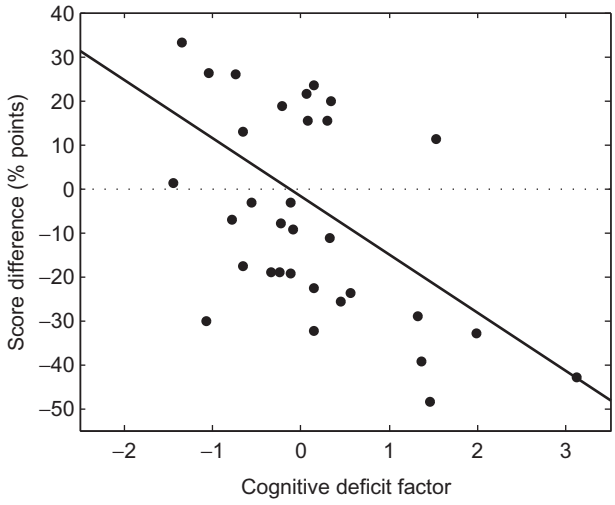

Figure 4. Scatterplots showing the difference in aided scores on the two tests (NAL-DCT–BKB) as a function of the composite cognitive deficit factor.

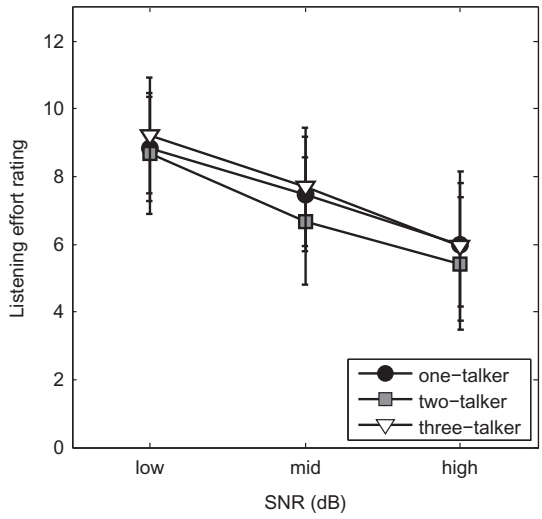

Figure 5. Mean unaided listening effort ratings as a function of SNR (low, mid and high) for the three talker conditions. Error bars depict across-subject standard deviations.

variance explained, and better performance being associated with the higher SNRs. Note that this indicates that the use of individualised test SNRs eliminated the contribution of 4FAHL to performance. SNR also made the strongest significant contribution to the scores obtained with NAL-DCT (standardised coefficient $=0.96$; $p<0.001$), followed by the cognitive deficit factor (standardised coefficient $=-0.54$; $p<0.001$). As for the sentence test, better performance was associated with the higher SNRs, and similar to what was seen for the unaided data, better performance was associated with better cognitive abilities. Combined, the two factors explained 53% of the variance.

To parallel the analyses conducted on unaided thresholds, the differences in aided scores between the two tests were calculated (with positive values meaning the comprehension task reduced percent correct). A regression analysis was conducted on these differences with the same predictors as above. SNR made the strongest contribution to this difference (standardised coefficient $=-0.86$; $p<0.001$) followed by the cognitive deficit factor (standardised coefficient $=-0.41$; $p<0.001$), explaining 79% of variance in the difference in performance on the two tests. This provides further support for the idea that those with lower cognitive abilities performed more poorly on NAL-DCT than on the sentence test (Figure 4).

Subjective ratings of listening effort

Figure 5 shows mean ratings of unaided listening effort as a function of SNR for the three talker conditions (c.f. Figure 1). The data show the expected effect of SNR, and only small differences between the three talker conditions. To confirm these observations, a repeated-measures ANOVA was conducted with factors of SNR (low, mid and high) and talker condition (one-, two- and three-talker) as repeated measures, and 4FAHL as a covariate. This analysis revealed a significant effect of SNR [$F(2,78)=33.1$; $p<0.001$] and talker condition [$F(2,78)=6.5$; $p=0.003$], but no interaction [$F(4,156)=2.1$; $p=0.09$]. Post-hoc comparisons (paired *t*-tests, with ratings collapsed across SNR for each listener) suggested that ratings were lower for the two-talker condition than for the one- and three-talker conditions ($p<0.05$). There was no main effect of 4FAHL [$F(1,39)=0.2$; $p=0.66$], but it did interact significantly with talker condition [$F(2,78)=3.5$, $p=0.04$]. Closer inspection of the data suggests that it was particularly those with milder hearing losses who found it less effortful to listen to the two-talker conversations. The interaction between 4FAHL and SNR was not significant [$F(2,78)=2.1$; $p=0.13$], nor was the three-way interaction [$F(4,156)=1.5$; $p=0.20$]. Overall, these results mirror those found for the NAL-DCT scores, although the listening effort ratings were sensitive to the talker condition whereas the NAL-DCT scores were not. When examined for each of the 41 participants, correlations between unaided NAL-DCT scores and the corresponding listening effort ratings across listening condition (3 SNRs $\times 3$ talker conditions) ranged from -0.41 to -0.96, and were significant at $p<0.05$ in 35 of the 41 participants.

Unaided and aided listening effort ratings are shown in Figure 6 as a function of test SNR. Qualitatively, the general trends in the listening effort ratings mirror the observations made above for the objective scores (c.f. Figure 3). To determine which factors significantly influence listening effort, a mixed ANOVA was conducted with within-subjects factors of talker condition (one-, two- and three-talker) and listening mode (unaided and aided), and SNR as a between subjects factor. This analysis revealed significant main effects of talker condition [$F(2,58)=7.7$; $p=0.001$] and SNR [$F(2,29)=23.2$; $p<0.001$], but no main effect of listening mode [$F(1,29)=0.0$; $p=0.94$]. Post-hoc comparisons (paired *t*-tests, with unaided/aided ratings collapsed for each listener) suggested that significantly lower effort ratings were given in the two-talker condition than in the three-talker condition. The interaction between listening mode and SNR was not significant [$F(2,29)=2.0$; $p=0.16$], nor were the interactions between listening mode and talker condition [$F(2,58)=0.2$; $p=0.86$] or talker condition and SNR [$F(4,58)=0.8$, $p=0.55$]. The three-way interaction was not significant [$F(4,58)=2.0$; $p=0.11$].

Discussion

A test that is sensitive to cognitive factors

Our primary goal in designing the NAL-DCT was to provide a complement to traditional sentence-in-noise tests that assesses

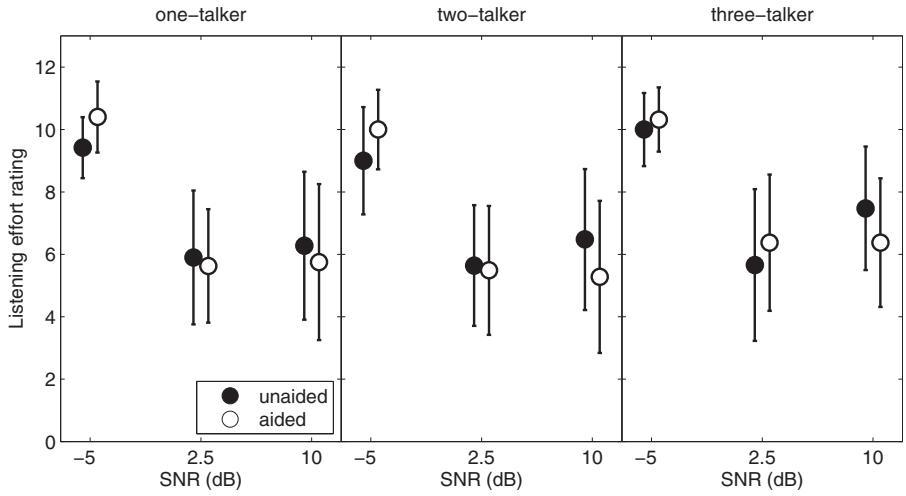

Figure 6. Unaided listening effort ratings (filled symbols) and aided listening effort ratings (open symbols) as a function of SNR for the one-, two- and three-talker conditions (left, middle and right). Error bars depict across-subject standard deviations.

comprehending in addition to *hearing* and *listening*. To increase the realism of the test, we incorporated multi-person conversations as targets, and presented the materials in a simulated noisy environment containing competing conversations. The overall intention was to tap into some of the higher-level processes involved in everyday communication.

The strong contribution of hearing loss to performance on the NAL-DCT suggests that the known ''SNR loss'' accompanying hearing loss propagates from lower-level speech processing to higher-level speech processing. Moreover, evidence that the NAL-DCT achieves the goal of tapping into higher-level processing comes from the fact that we also found a significant contribution of cognitive abilities to performance. In our previous study in young NH listeners, we found an association between NAL-DCT performance and working memory. Here, we made use of a composite cognitive factor that encompasses working memory and the speeded processing of language, and showed that it was predictive of the *difference* in performance on the NAL-DCT relative to a BKB sentence test. This suggests that sentence recognition tests that are commonly used in hearing clinics may overpredict speech communication performance in people who are less cognitively able, presumably because they do not incorporate the requirement to ''keep up'' with speech in an ongoing way while formulating responses to the received information. Conversely, it appears that sentence recognition tests may underpredict performance in those with strong cognitive skills, possibly because they do not offer contextual information that these individuals can use to support performance.

These tentative conclusions are consistent with recent work looking at speech recognition and learning in children, whose cognitive abilities are less well developed than in adults. It was reported, for example, that children performed more poorly than adults on a listening comprehension test conducted in a simulated classroom, whereas the groups were both near ceiling for a BKB sentence test under the same conditions (Valente et al. 2012). The comprehension task was also better able to differentiate children with and without hearing loss (Lewis et al. 2015). There is much scope for future studies that aim to more carefully unpack the specific cognitive abilities that give rise to individual differences in real-world speech communication across different populations.

The NAL-DCT and subjective listening effort

Another finding of the current study was that scores on the NAL-DCT tended to mirror self-reported ratings of listening effort. The two sets of scores showed similar patterns with respect to SNR and talker condition, and were generally correlated within subjects. Although the listening effort ratings appeared more sensitive than the performance scores to the dynamic changes in speech, we see no strong evidence in our data that the two measures tap into different constructs as previously reported (e.g. Humes 1999; Rudner et al. 2012; Johnson et al. 2015; van den Tillaart-Haverkate et al. 2017). That is, we cannot rule out a hypothesis that the subjective ratings of listening effort reflected how well participants felt they performed on the task just completed, rather than the mental exertion they experienced during the task. Conversely, it is also possible that the observed association between the two sets of scores suggests that NAL-DCT performance is in fact influenced by listening effort. This would not be surprising when one considers that the NAL-DCT requires the listener to comprehend conversational speech in a continuous way for a sustained period of time, with the added requirement of keeping track of and answering the associated questions. In fact, the main goal of the NAL-DCT was to capture factors that are important to the listener's experience in a given situation, and in that sense the relationship is encouraging. However, to test this idea it is necessary to show that NAL-DCT scores are more strongly associated with listening effort than sentence recognition scores, which is unfortunately not possible with the current data set. In general, further investigations are warranted into the relationship between speech comprehension scores and listening effort, perhaps incorporating objective measures of listening effort (e.g. Ayasse et al. 2017; Gagné et al. 2017).

The dynamic aspect of the NAL-DCT

A unique feature of the NAL-DCT is that it includes conversations between pairs and triplets of talkers, and provides the opportunity to

include unpredictable changes in voice and location within a single comprehension passage. In our previous study in NH listeners (Best et al. 2016), we found only a modest effect of the number of talkers, and in fact scores for the one-talker condition were slightly lower on average than for the two- and three-talker conditions. In the current study, we did not find any significant effects of talker condition on NAL-DCT scores, and the listening effort data indicated that the two-talker conversations were slightly easier to follow. Overall, there is no evidence that our older participants had particular difficulties making use of spatial information to follow a conversation between more than one person as has been reported previously (Murphy et al. 2006).

It is worth noting, however, that the simplified language that is characteristic of multiple-talker conversations might outweigh any difficulties associated with having to switch attention between talkers in the NAL-DCT. It is possible that we would have observed stronger effects of talker condition had we used the same set of passages and varied only the number of locations and/or talkers (as per Murphy et al. 2006; Valente et al. 2012; Lewis et al. 2015).

Hearing aid benefit and disbenefit

Another goal of this work was to examine the potential utility of the NAL-DCT for measuring the real-world benefit of hearing aids. Scores on both the BKB and the NAL-DCT were sensitive to the listening mode (unaided vs. aided) but there was evidence of an interaction between hearing aid benefit and the test SNR. This result is in line with recent reports showing that input SNR can affect objective measures of hearing aid benefit (e.g. Naylor and Johannesson 2009; Naylor 2016). A more comprehensive study is needed, in which the SNR is varied within rather than across subjects, to examine this issue further.

Further work is also needed to determine if the NAL-DCT can provide new information (above that provided by traditional speech-in-noise tests) about the benefits of different hearing aid processing schemes. In general, the NAL-DCT might complement traditional speech tests by revealing whether improved recognition translates into improved comprehension. And more specifically, we believe the NAL-DCT could be particularly well-suited for evaluating highly directional devices such as binaural beamformers. While these devices can offer large advantages for fixed, frontal targets, there are indications that spatial dynamics reduce this benefit (Best et al. 2015). The NAL-DCT provides a tool for measuring the benefit of such devices under conditions containing realistic, conversational dynamics. In a clinical context, the output of such a test would be useful for counselling hearing aid candidates and setting realistic expectations about how hearing aids will impact their experience in social settings.

Conclusions

The NAL-DCT appears to be suitable for testing older listeners with hearing loss under realistic communication conditions. When compared to a more traditional sentence-based test, while performance for both tests was strongly driven by hearing loss, performance on the NAL-DCT was additionally related to a measure of cognitive deficit. Benefits of amplification were measurable with the NAL-DCT but depended on the test SNR. Performance on the NAL-DCT was generally aligned with subjective measures of listening effort.

Acknowledgements

Preliminary data were presented at the International Hearing Aid Conference (Lake Tahoe, USA, August 2016). Work supported by a grant from the Hearing Industry Research Consortium and the Australian Government Department of Health. Virginia Best was also partially supported by NIH-NIDCD grant DC04545. The authors would like to thank Adam Westermann and James Galloway for their technical help at various stages.

Declaration of interest: No potential conflict of interest was reported by the authors.

References

Akeroyd, M. A. 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.'' *International Journal of Audiology* 47: S53–S71.

Ayasse, N. D., A. Lash, and A. Wingfield. 2017. ''Effort Not Speed Characterizes Comprehension of Spoken Sentences by Older Adults with Mild Hearing Impairment.'' *Frontiers in Aging Neuroscience* 8: 329.

Besser, J., T. Koelewijn, A. A. Zekveld, S. E. Kramer, and J. M. Festen. 2013. How Linguistic Closure and Verbal Working Memory Relate to Speech Recognition in Noise – A Review.'' *Trends in Amplification* 17: 75–93.

Best, V., G. Keidser, K. Freeston, and J. M Buchholz. 2016. ''A Dynamic Speech Comprehension Test for Assessing Real-world Listening Ability.'' *Journal of the American Academy of Audiology* 27: 515–526.

Best, V., M. McLelland, and H. Dillon. 2014. The BEST (Beautifully Efficient Speech Test) for Evaluating Speech Intelligibility in Noise World Congress of Audiology. Brisbane, Australia.

Best, V., J. Mejia, K. Freeston, R. J. van Hoesel, and H. Dillon. 2015. ''An Evaluation of the Performance of Two Binaural Beamformers in Complex and Dynamic Multitalker Environments.'' *International Journal of Audiology* 54: 727–735.

Best, V., E. J. Ozmeral, N. Kopčo, and B. G. Shinn-Cunningham. 2008. ''Object Continuity Enhances Selective Auditory Attention.'' *Proceedings of the National Academy of Science* 105: 13173–13177.

Daneman, M., and P. Carpenter. 1980. ''Individual Differences in Working Memory and Reading.'' *Journal of Verbal Learning and Verbal Behavior* 19: 450–466.

Feuerstein, J. 1992. ''Monaural versus Binaural Hearing: Ease of Listening, Word Recognition, and Attentional Effort.'' *Ear and Hearing* 13: 80–86.

Fraser, S., J. P. Gagné, M. Alepins, and P. Dubois. 2010. ''Evaluating the Effort Expended to Understand Speech in Noise using a Dual-task Paradigm: The Effects of Providing Visual Speech Cues.'' *Journal of Speech, Language, and Hearing Research* 53: 18–33.

Gagné, J. P., J. Besser, and U. Lemke. 2017. ''Behavioral Assessment of Listening Effort using a Dual-task Paradigm: A Review.'' *Trends in Hearing* 21. doi:10.1177/2331216516687287.

Gordon-Salant, S., and S. S. Cole. 2016. ''Effects of Age and Working Memory Capacity on Speech Recognition Performance in Noise Among Listeners with Normal Hearing.'' *Ear and Hearing* 37: 593–602.

Hällgren, M., B. Larsby, B. Lyxell, and S. Arlinger. 2001. ''Evaluation of a Cognitive Test Battery in Young and Elderly Normal-hearing and Hearing-impaired Persons.'' *Journal of the American Academy of Audiology* 12: 357–370.

Humes, L. E. 1999. ''Dimensions of Hearing Aid Outcome.'' *Journal of the American Academy of Audiology* 10: 26–39.

Jensen, N. S., R. B. Johannesson, S. Laugesen, and R. K. Hietkamp. 2012. ''Measuring Speech-in-speech Intelligibility with Target Location Uncertainty.'' In *Speech perception and auditory disorders. Proceedings of the International Symposium on Audiological and*

Auditory Research (ISAAR), edited by Dau, T., Jepsen, M.L., Christensen-Dalsgaard, J., and Poulsen, T., 135–142. Denmark.

Johnson, J., J. Xu, R. Cox, and P. A. Pendergraft. 2015. ''A Comparison of Two Methods for Measuring Listening Effort as Part of An Audiologic Test Battery.'' *American Journal of Audiology* 24: 419–431.

Karlsen, B., and E. F. Gardener. 1984. *Stanford Diagnostic Reading Test*. NY, USA: Psychological Corporation.

Keidser, G., H. Dillon, J. Mejia, and C. V. Nguyen. 2013. ''An Algorithm that Administers Adaptive Speech-in-noise Testing to a Specified Reliability at Selectable Points on the Psychometric Function.'' *International Journal of Audiology* 52: 795–800.

Kiessling, J., M. K. Pichora-Fuller, S. Gatehouse, D. Stephens, S. Arlinger, et al. 2003. ''Candidature for and Delivery of Audiological Services: Special Needs of Older People.'' *International Journal of Audiology* 42: 92–101.

Lewis, D. E., D. L. Valente, and J. L. Spalding. 2015. ''Effect of Minimal/Mild Hearing Loss on Children's Speech Understanding in a Simulated Classroom.'' *Ear and Hearing* 36: 136–144.

Luts, H., K. Eneman, J. Wouters, M. Schulte, M. Vormann, et al. 2010. ''Multicenter Evaluation of Signal Enhancement Algorithms for Hearing Aids.'' *Journal of Acoustical Society of America* 127: 1491–1505.

McGarrigle, R., K. J. Munro, P. Dawes, A. J. Stewart, D. R. Moore, et al. 2014. ''Listening Effort and Fatigue: What Exactly are We Measuring? A British Society of Audiology Cognition in Hearing Special Interest Group 'white paper'.'' *International Journal of Audiology* 53: 433–445.

Murphy, D. R., M. Daneman, and B. A. Schneider. 2006. ''Why do older adults have difficulty following conversations?'' *Psychology Aging* 21: 49–61.

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: 504–514.

Naylor, G., and R. B. Johannesson. 2009. ''Long-term Signal-to-noise Ratio at the Input and Output of Amplitude-compression Systems.'' *Journal of the American Academy of Audiology* 20: 161–171.

Rönnberg, J., S. Arlinger, B. Lyxell, and C. Kinnefors. 1989. ''Visual Evoked Potentials: Relation to Adult Speechreading and Cognitive Function.'' *Journal of Speech Hearing and Research* 32: 725–735.

Rudner, M., T. Lunner, T. Behrens, E. S. Thorén, and J. Rönnberg. 2012. ''Working Memory Capacity may Influence Perceived Effort During Aided Speech Recognition in Noise.'' *Journal of the American Academy of Audiology* 23: 577–589.

Souza, P. and K. Arehart. 2015. ''Robust Relationship Between Reading Span and Speech Recognition in Noise.'' *International Journal of Audiology* 54: 705–713.

Valente, D. L., H. M. Plevinsky, J. M. Franco, E. C. Heinrichs-Graham, and D. E. Lewis. 2012. ''Experimental Investigation of the Effects of the Acoustical Conditions in a Simulated Classroom on Speech Recognition and Learning in Children.'' *Journal of the Acoustical Society of America* 131: 232–246.

van den Tillaart-Haverkate, M., I. de Ronde-Brons, W. A. Dreschler, and R. Houben. 2017. ''The Influence of Noise Reduction on Speech Intelligibility, Response Times to Speech, and Perceived Listening Effort in Normal-hearing Listeners.'' *Trends in Hearing* 21. doi:10.1177/2331216517716844.