Taylor & Francis
Taylor & Francis Group

International Journal of Audiology 2015; Early Online: 1–8

International Journal of Audiology

Original Article

# Development and preliminary evaluation of a new test of ongoing speech comprehension

Virginia Best[*,†], Gitte Keidser[*], Jörg M. Buchholz[*,‡] & Katrina Freeston[*]

[*] *National Acoustic Laboratories and the HEARing Cooperative Research Centre, Australian Hearing Hub, Macquarie University, Australia*
[†] *Department of Speech, Language and Hearing Sciences, Boston University, Boston, USA and* [‡] *Audiology Section, Department of Linguistics, Australian Hearing Hub, Macquarie University, Australia*

The British Society of Audiology

The International Society of Audiology

NAS NORDIC AUDIOLOGICAL SOCIETY

## Abstract

*Objective:* The overall goal of this work is to create new speech perception tests that more closely resemble real world communication and offer an alternative or complement to the commonly used sentence recall test. *Design:* We describe the development of a new ongoing speech comprehension test based on short everyday passages and on-the-go questions. We also describe the results of an experiment conducted to compare the psychometric properties of this test to those of a sentence test. *Study sample:* Both tests were completed by a group of listeners that included normal hearers as well as hearing-impaired listeners who participated with and without their hearing aids. *Results:* Overall, the psychometric properties of the two tests were similar, and thresholds were significantly correlated. However, there was some evidence of age/cognitive effects in the comprehension test that were not revealed by the sentence test. *Conclusions:* This new comprehension test promises to be useful for the larger goal of creating laboratory tests that combine realistic acoustic environments with realistic communication tasks. Further efforts will be required to assess whether the test can ultimately improve predictions of real-world outcomes.

**Key Words:** Speech comprehension; realistic; hearing loss; hearing aids

The speech reception threshold (SRT) is routinely measured in the laboratory to assess speech understanding in noise. Typically, the SRT measures the ability of listeners to repeat short isolated utterances that range from unpredictable nonsense sentences to highly predictable (but out of context) everyday sentences (e.g. Kalikow et al, 1977; Hagerman, 1982). Numerous studies have demonstrated that the characteristics of the background noise influence SRTs. For example, it is generally the case that fluctuating maskers lead to better SRTs than steady-state maskers, as listeners can make use of dips in the masker to 'glimpse' the target (Miller & Licklider, 1950; Cooke, 2006). On the other hand, intelligible speech maskers can lead to poorer SRTs because they are distracting and can cause semantic interference (Kidd et al, 2008). These investigations have greatly advanced our understanding of speech perception in real-world environments, which often fluctuate in level but also contain other competing talkers.

However, sentence repetition does not closely resemble real-world communication, which is ongoing rather than discrete, and involves the extraction of meaning, and generally the formulation of a valid reply. These additional tasks require more higher-level processing than is needed to recognize words in isolated sentences (Pichora-Fuller, 2007; Schneider et al, 2010). According to the Ease of Language Understanding model (Rönnberg, 2003; Rönnberg et al, 2013), working memory plays an important part in speech understanding and comprehension, especially when the input signal is degraded, e.g. by the presence of noise or a hearing impairment. Working memory is considered a limited-capacity system (Baddeley, 1996). Consequently, the more resources that are spent on purely understanding what has been said, the less cognitive resources are available for making sense of what has been heard and for response preparation. It could therefore be expected that in a similar challenging listening environment, a person may show increased listening effort or even poorer performance on a communication style test than on a sentence repetition test.

For laboratory speech tests to be reflective of real-world abilities, there is a need for new approaches that better capture the added higher level processing demands of real communication situations. This need is increasingly being voiced by hearing-aid manufacturers and clinicians wanting to make more powerful assessments of intervention with devices and benefit from different features. Better predictions of real-world outcomes would also allow earlier assessment and fine-tuning of new technologies and thus greatly reduce the number of lengthy and expensive field tests needed during the

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 31 March 2014; accepted 25 May 2015)*

ISSN 1499-2027 print/ISSN 1708-8186 online © 2015 British Society of Audiology, International Society of Audiology, and Nordic Audiological Society
DOI: 10.3109/14992027.2015.1055835

**Abbreviations**

| | |
|---|---|
| 4FAHL | Four-frequency average hearing loss |
| BKB | Bamford-Kowal-Bench (sentences) |
| HI | Hearing-impaired |
| IELTS | International English Language Testing System |
| NAL | National Acoustic Laboratories |
| NH | Normally hearing |
| SNR | Signal-to-noise ratio |
| SRT | Speech reception threshold |

research and development phase. Moreover, it has been suggested that greater sensitivity to cognitive factors will be critical for future efforts to customize hearing-aid technology for the individual listener (e.g. Pichora-Fuller, 2007; Lunner et al, 2009).

The simulation of more natural communication situations in the laboratory is an enormous challenge, but there have been several attempts in research laboratories to move in this direction in recent years. To tap into the *ongoing* nature of real speech communication, MacPherson and Akeroyd recently developed the Glasgow monitoring of uninterrupted speech task (GMUST; MacPherson & Akeroyd, 2013). This task requires participants to listen to an audiobook while simultaneously reading along with a written transcript, which contains occasional errors. The accuracy with which a participant can report these errors is used as a measure of their ability to both listen and keep up. Other approaches have focused on the *comprehension* aspect of communication, which requires not only that the acoustic features of speech are understood but also that meaning is extracted from them (see Humes & Dubno, 2010). The typical format of a comprehension test involves an extended spoken story or lecture (5–15 minutes) that is followed by a set of questions that assess understanding of the content (e.g. Schneider et al, 2000; Murphy et al, 2006; Tye-Murray et al, 2008; Gordon et al, 2009; Sommers et al, 2011). One down side of these tests is that there is a heavy episodic memory requirement, not present in most real conversations, that may dominate performance, especially in older listeners (e.g. Nyborg et al, 1996). A few attempts have been made to overcome this issue by either using much briefer passages (Kei & Smyth, 1997; Kei et al, 2003) or by assessing listeners immediately after the relevant information occurs (Hafter et al, 2013). None of the approaches described above have tapped into a final aspect of real conversations, which is the requirement to further spend cognitive resources on *formulating responses* often while continuing to listen. At the National Acoustic Laboratories (NAL) we decided to investigate a new comprehension test that combines desirable features of the above approaches but incorporates an 'on-the-go' question and answer component. This test requires listeners to follow along continuously, identify relevant pieces of information, and give brief, immediate, written responses.

Here we describe our first efforts to develop a working version of this test for research purposes. We also present the results of an experiment that compared the properties of the test to our standard sentence-based SRT test, to understand the psychophysical consequences of moving to this kind of task. To reduce the number of uncontrolled variables, the speech materials for the comprehension test were spoken by the same talker who spoke the Bamford-Kowal-Bench (BKB) sentences routinely used for SRT testing at NAL (Keidser et al, 2002). The same set of listeners completed speech-in-noise testing using the two different tests under otherwise identical conditions. Finally, because the work was largely motivated by the need for new tests that can predict the real-world benefit of hearing aids, we also examined the ability of the different tests to capture changes in performance due to amplification in hearing-impaired listeners.

## Development of the comprehension test

### *Materials*

Comprehension passages were taken from the listening comprehension component of the International English Language Testing System (IELTS). The IELTS is a widely used test of English language proficiency for education, immigration, and professional purposes in the UK, Australia, and elsewhere. Transcripts of the passages, and their associated comprehension questions, are publicly available in books of past examination papers (e.g. see Jakeman & McDowell, 1995). The passages involve one, two, or three talkers, and are several minutes long when spoken aloud. They cover everyday topics such as public transport information, headline news, information about tourist attractions, etc. Twenty-eight single-talker monologues were selected from the IELTS collection for use in this preliminary evaluation (see Appendix). They were chosen somewhat arbitrarily based on how interesting the experimenters deemed they might be to our Australian participants.

### *Recordings*

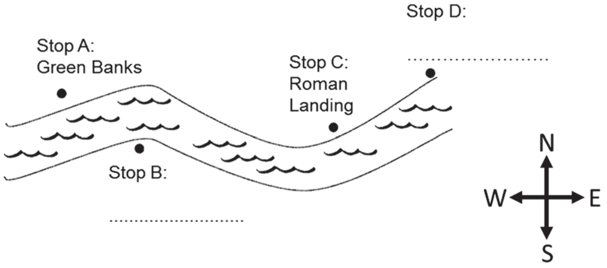
The Australian male talker who had recorded the BKB sentences (Keidser et al, 2002) was recruited for two recording sessions to read aloud the monologues. He was instructed to speak in a natural way, and to stop and restart the sentence if he made an error, stumbled, or needed to cough, etc. Offline editing was done later to remove these errors. The final recordings ranged in duration from 2 minutes 9 seconds to 4 minutes 21 seconds (mean 3 minutes 23 seconds).

The recordings were done in a large anechoic chamber, using a Sennheiser ME 64 microphone (pre-polarized condenser, cardioid) connected to an M-AUDIO MobilePre USB sound card. The gain on the M-AUDIO sound card was set such that the signal-to-noise ratio (SNR) was maximum, preventing clipping at all times. The sampling rate was 44100 Hz. After processing, the monologues were scaled to an equivalent root-mean-square level and saved as .wav files. The recording as well as subsequent processing was done in Adobe Audition 3.0 and CS5.5.

### *Comprehension questions*

Question sheets for each monologue were generated by adapting the question sheets available in the printed IELTS books. Each sheet consists of 10 comprehension questions that pertain to the passage, and the questions are arranged sequentially on the page according to the temporal order in which the information occurs within the passage. The questions come in five broad style categories as described in Table 1. Each question sheet contains a random selection from these categories and does not necessarily have a question from every category. Space is allocated for written answers and the questions are designed to be answered in an 'on-the-go' fashion. Note that the responses expected from subjects are very brief, ranging from ticking a box to labelling an image to writing a few words. It is also worth pointing out that the questions varied in how directly they related to the words spoken in the passages but even where an answer used similar phrasing to the passage the wording around the answer never matched exactly, and a correct response required listeners to

**Table 1.** The five categories of question used to assess comprehension.

| *Style* | *Description* | *Example* |
|---|---|---|
| Multiple choice | A brief question is posed with three alternative answers. | Where does Circus Romano perform?<br>A: in a theatre<br>B: in a tent<br>C: in a stadium |
| Checklist | A list of items is given, a subset of which are true or were mentioned in the passage. | Which TWO of the following can you get advice about from the Union?<br>A: immigration<br>B: grants<br>C: medical problems<br>D: personal problems<br>E: legal matters |
| Fill in the blanks | A sentence from the passage is given, with one item missing. | The government plans to give $........ to assist the farmers. |
| Short-answer | A question requiring a brief answer is posed. | How often do the Top Bus Company tours run? |
| Image | An image corresponding to the passage is given, with some labels missing. |  |

understand what they heard and extract the appropriate keyword(s). In other questions the answer used very different phrasing to the passage and thus a correct answer needed to be inferred. Despite this range, however, continuous comprehension of the passage was required in order for a listener to keep their place and link the questions to the relevant pieces of heard information.

Minor modifications to the question sheets were made to make the text easily readable and the images large and clear. The questions were scored by hand by the experimenters after testing, using the answers provided in the IELTS books.

### *Initial screening*

A short screening was carried out to identify any passages or questions that stood out from the rest as being consistently difficult or confusing even under favourable listening conditions. Twenty-three subjects with normal hearing were recruited (14 female, nine male). Ten were NAL employees and 13 were external participants. They ranged in age from 19 to 39 years (mean 26 years, standard deviation 5 years).

Subjects were seated at a desk in an audiometric booth, and gave written answers on the appropriate question sheets. The monologues were presented diotically over headphones (Sennheiser HD215) either in quiet or in the presence of eight-talker speech babble (taken from the NAL CDs of speech and noise for hearing-aid evaluation, Keidser et al, 2002) that was also diotic. The 28 monologues were divided into two sets. Fourteen subjects listened to the first set of monologues (five subjects listened in quiet, five in babble at 0 dB SNR, and four in babble at − 3 dB). Nine subjects listened to the second set of monologues (three subjects listened in quiet, three in babble at 0 dB, and three in babble at − 3 dB). The order of testing of the monologues within a session was randomized for each subject.

On the basis of this evaluation, several problematic questions were identified (i.e. incorrect answers were given by more than three subjects). These questions were inspected carefully by two of the investigators. In cases where the cause was judged to be ambiguity in the question, the question was reworded. Where the cause was judged to be related to the recording itself (e.g. the talker skipped a keyword or did not enunciate clearly), a replacement question was generated. On average, about one question per passage was modified.

Finally, the 28 monologues were divided into two lists of 14 for use in the experiment. This division was done such that the easier/harder monologues (based on average scores in the screening), as well as the different styles of question, were approximately evenly distributed between the two lists.

## Evaluation study

### *Participants*

Thirty seven listeners participated (11 female, 26 male). Eleven of these had normal hearing (‘normally hearing’, NH). Their age ranged from 18 to 57 years (mean 43 years, standard deviation 12 years) and their four-frequency average hearing loss (4FAHL, mean of left and right ear pure-tone thresholds at 500, 1000, 2000, and 4000 Hz) ranged from 1 to 15 dB (mean 7 dB). The other 26 had bilateral sensorineural hearing losses (‘hearing-impaired’, HI) with air-bone gaps of no more than 10 dB at any frequency. Their age ranged from 29 to 80 years (mean 70 years, standard deviation 11 years) and their 4FAHL ranged from 27 to 78 dB (mean 46 dB). While left- and right-ear asymmetries at a given frequency could be up to 25 dB, mean asymmetries across 500, 1000, 2000, and 4000 Hz were less than 5 dB in all but three listeners (where the differences were 6 dB, 9 dB, and 14 dB). Audiograms for each listener are plotted in Figure 1, along with group averages. Note that 4FAHL was correlated with age in the total pool ($r = 0.63$. $p < 0.001$) but not in the subgroup of HI listeners ($r = -0.28$, $p = 0.16$) as a result of several young listeners with quite severe losses. Reading and vision were not assessed, but participants were instructed to bring their

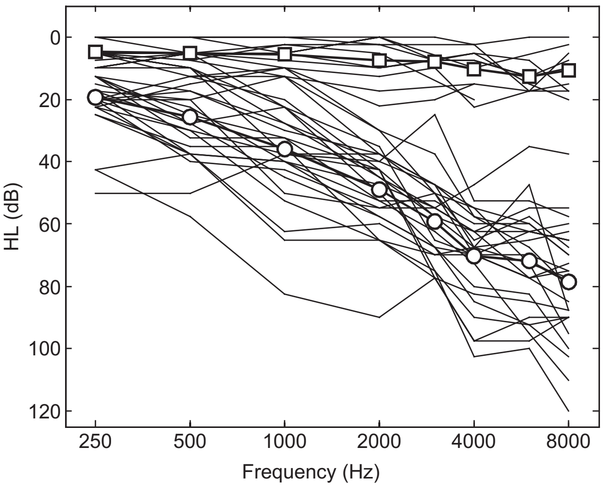


**Figure 1.** Audiograms for each listener (averaged across left and right ears), as well as group means for the NH group (squares) and the HI group (circles).

reading glasses if applicable. No subject reported issues related to seeing/reading the question sheets.

Twenty-three of the 26 HI listeners were regular hearing-aid wearers and participated in the experiment both with and without their own hearing aids. All hearing aids were behind-the-ear styles, seven with the receiver in the canal. They represented a variety of entry-level and high-end devices from Phonak, Resound, Siemens, Oticon, Bernafon, Unitron, and Rexton Day, and were set to the user's most common program for testing. We did not attempt to adjust the gain or compression settings to ensure uniformity across participants but, rather, opted to use the settings each listener was accustomed to using in their daily lives.

All participants were paid a small gratuity for their participation. The 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. Note that these listeners also participated in another study described in a companion paper (Best et al, 2015). The 37 listeners that participated here are a subset of the 47 that participated in the other study, and some of the data presented here (for the sentence test) also appear in that paper (as the 'standard' environment).

### *Environment, stimuli, and tasks*

Testing took place in a large anechoic chamber fitted with a loudspeaker array of radius 1.8 m. The experiment made use of five equalized loudspeakers (Tannoy V8) positioned at 0° elevation (0°, ± 45° and ± 135° azimuth). Stimulus playback was via a PC with a sound card (RME MADI) connected to two D/A converters (RME M-32) and 11 four-channel amplifiers (Yamaha XM4180).

The listener was seated such that the head was in the centre of the array, facing the frontal loudspeaker, and wore a small lapel microphone in order to be heard clearly by the experimenter who was seated outside the chamber wearing headphones. The experimenter monitored participants via webcam to ensure they maintained a relatively fixed head position, and could talk to them via an intercom as required.

In both the sentence and comprehension tests the target speech was presented from the frontal loudspeaker, and four independent samples of the eight-talker speech babble were presented from the other four loudspeakers. The babble was presented continuously throughout a block of trials at a fixed level of 65 dB SPL (measured in the centre of the array).The speech level was set on an individualized basis as described below.

In the sentence test, targets were BKB sentences. Listeners spoke aloud their responses and the experimenter entered the number of correct morphemes (out of a possible 3–8, depending on the sentence) into the software program. In the comprehension test, targets were IELTS monologues. Before each presentation the subject was given the question sheet on a clipboard and had about half a minute to read over it. The monologue was then presented and subjects gave written answers on the question sheet. Answers were scored offline by the experimenters.

### *Procedures*

All 37 listeners completed the sentence test, and then attempted the comprehension test. Two of the HI listeners (both 80 years old) could not complete the comprehension test as they could 'hear' but could not 'keep up' enough to answer the questions. Comprehension results are thus shown only for the remaining 35 listeners. The 23 HI listeners who were tested both unaided and aided completed all sentence testing before any comprehension testing, but with the hearing-aid condition counterbalanced across subjects. Sentence testing and comprehension testing were completed in separate visits, such that all listeners required two visits and hearing-aid wearers required four visits.

For the sentence test, four blocks of trials were completed (per hearing-aid condition for the hearing-aid wearers). In the first block, an adaptive procedure was used to estimate the 50% SRT (for details see Keidser et al, 2013b). A block of 32 sentences was then completed at each of three fixed SNRs: the estimated SRT, the SRT + 2 dB, and the SRT − 2 dB. The order of testing of the three SNRs was randomized, as was the pairing of sentence lists with SNRs, and the order of presentation of sentences within a block. No sentence was presented more than once to any listener. Each block took approximately five minutes to complete, for a total testing time of around 20 minutes.

For the comprehension test, each subject was presented with monologues from one of the two lists described above (and both lists were used for the aided HI listeners). One monologue was presented in quiet and served as a familiarization step. A further 12 monologues were presented in the babble background at three different SNRs (four monologues per SNR). Because adaptive tracking was not possible with the comprehension test, SNRs had to be chosen ahead of time. For the NH group SNRs of − 6, − 3, and 0 dB were chosen, based on informal listening, to cover a large range of the psychometric function whilst not hitting ceiling or floor. Slightly higher SNRs were chosen for the unaided and aided HI groups (− 3, 0, and + 3 dB). If a particular subject showed clear ceiling or floor effects after completing one monologue at the highest or lowest SNR, that passage was discarded (and replaced with the spare 14th monologue in each list), and the SNR range was shifted down or up accordingly by 3 dB to better capture the sloping part of the psychometric function. The pairing of monologues with SNRs and the order of presentation was randomized. The monologue used for familiarization and the spare monologue were randomly different across subjects. Each monologue took approximately five minutes to complete, resulting in a total testing time of around 60 minutes.

For each test and each listener, percent correct scores at the three fixed SNRs were used to generate psychometric functions. Logistic functions were fit to the raw scores using the psignifit toolbox version 2.5.6 for MATLAB (see http://bootstrap-software.org/psignifit/) which implements the maximum-likelihood method described by Wichmann and Hill (2001). SRTs were then estimated as the SNR corresponding to 50% correct according to these fits.

### *Self-report data*

Between visits, all participants were asked to take home and complete a questionnaire addressing their hearing abilities. The general purpose was to provide insight into how speech scores measured in the laboratory relate to real-world experience as measured by self-report. Fifteen questions addressing disability under specific situations were taken from the Speech, Spatial and Qualities of Hearing Scale (SSQ; Gatehouse & Noble, 2004). These questions included the 14 questions in the 'speech' subscale, as well as the question addressing listening effort from the 'qualities of hearing' subscale (question 18). Hearing-aid wearers answered all questions twice, based on listening unaided and aided. For the purposes of this study, a single score was calculated for each subject (separately for unaided and aided listening, where appropriate) by averaging over a subset of eight questions that refer to situations involving selective attention to speech in the presence of noise (speech items 1, 4, 5, 6, 7, 8, 9, 11).

## Results

Correlational analyses showed that unaided SRTs (including both NH and HI listeners) for both tests were correlated with 4FAHL (sentence: Spearman's $\rho = 0.80$; comprehension: $\rho = 0.71$; $p < 0.001$), as well as with age (sentence: $\rho = 0.55$; comprehension: $\rho = 0.56$; $p < 0.001$). Unaided SRTs were also strongly correlated across the two tests (Figure 2; Pearson's $r = 0.83$, $p < 0.001$) with the gradient of the least-squares fit relating the two tests close to one (0.96).

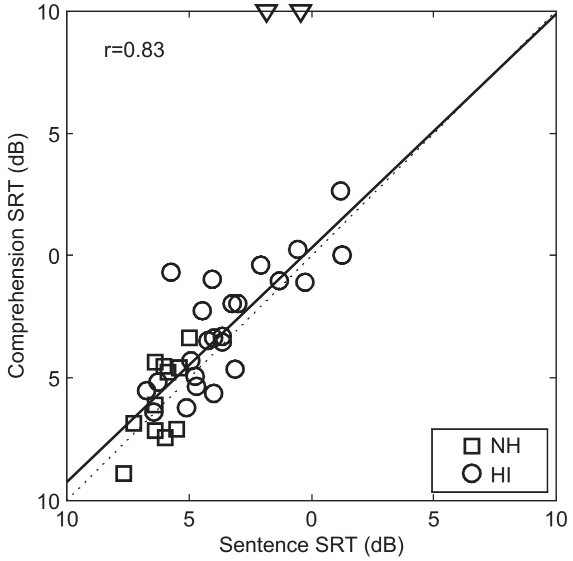


**Figure 2.** Scatterplot showing individual SRTs in the comprehension test against SRTs in the sentence test. Different symbols indicate NH listeners (squares) and unaided HI listeners (circles). The solid line shows the least squares fit. The two HI listeners who could not perform the comprehension test are shown by triangles and are not included in the fit.

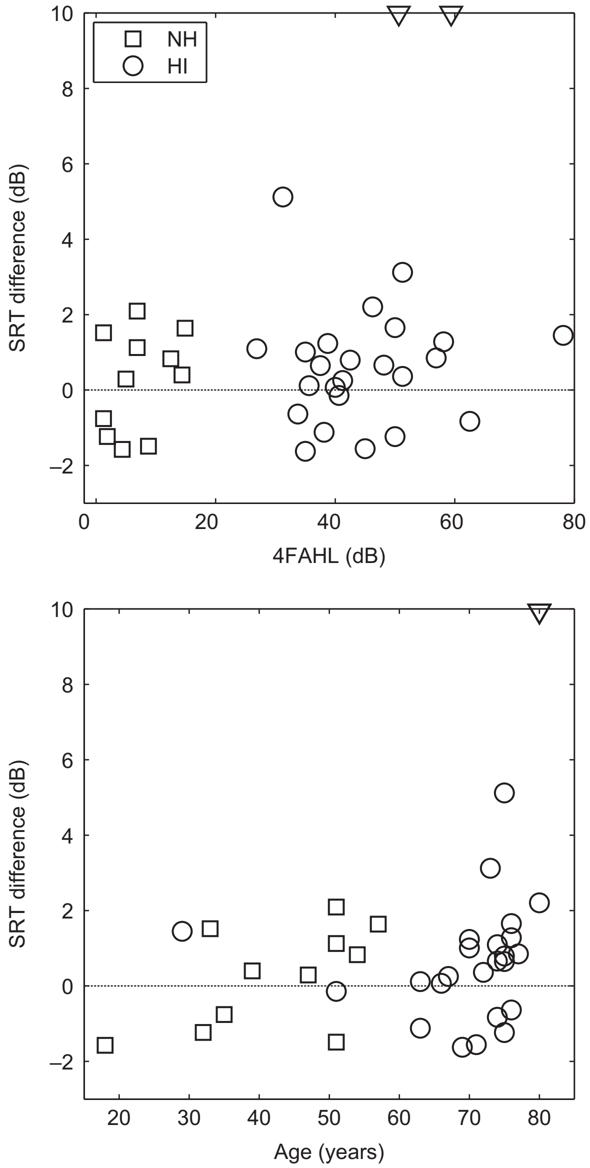


**Figure 3.** Scatterplot showing the difference in unaided SRTs between the two tests (comprehension SRT − sentence SRT) as a function of 4FAHL (upper panel) and age (lower panel). Different symbols indicate NH listeners (squares) and unaided HI listeners (circles). The two HI listeners who could not perform the comprehension test are shown by triangles.

Although correlated, it can be seen from Figure 2 that SRTs could be higher or lower on the comprehension test than on the sentence test, depending on the listener. To further investigate this, the change in SRT when moving from the sentence test to the comprehension test was calculated for each listener (with positive changes indicating an increase in SRT, or poorer performance). This SRT difference is plotted in Figure 3 as a function of 4FAHL (upper panel) and age (lower panel). The effect of test type was not strongly related to hearing loss (Spearman's $\rho = 0.14$, $p = 0.42$) or age ($\rho = 0.25$, $p = 0.15$), but the listeners whose performance dropped the most in the comprehension test were all greater than 70 years of age, including the two 80 year olds who could not perform the task at all.

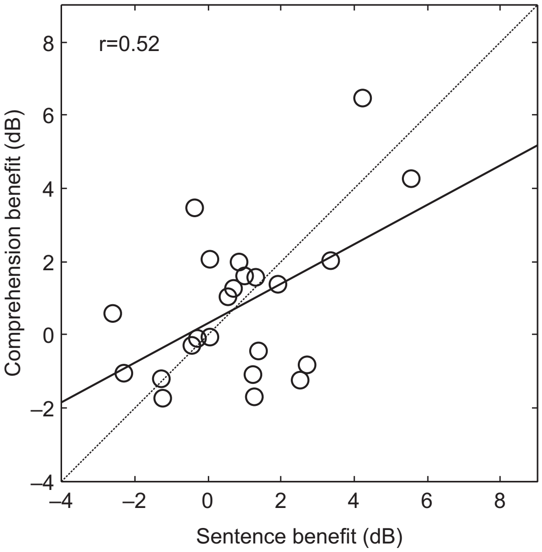


**Figure 4.** Scatterplot showing individual hearing-aid benefits in the comprehension test against hearing-aid benefits in the sentence test (positive benefits indicate a reduction in the SRT, i.e. better aided performance). The solid line shows the least squares fit.

Correlations between unaided self-report scores and unaided SRTs were highly significant in both tests (sentence: Pearson's $r = -0.71$, $p < 0.001$; comprehension: $r = -0.74$, $p < 0.001$). This was expected given the strong correlation between SRTs on the two tests.

Hearing-aid benefits were calculated for each hearing-aid wearer by subtracting unaided SRTs from aided SRTs. To verify that hearing-aid benefits in the comprehension test were not affected by learning effects, we compared hearing-aid benefits for those listeners who did aided testing first versus second and found no significant difference [$t(21) = 1.15$, $p = 0.26$]. A substantial range of benefits was observed across individuals, ranging from − 2.6 dB to 5.6 dB in the sentence test, and from − 1.7 dB to + 6.5 dB in the comprehension test. On average, the benefit of hearing aids was 0.87 dB in the sentence test and 0.78 dB in the comprehension test, and individual benefits were correlated (Figure 4; Pearson's $r = 0.52$, $p = 0.01$).

## Discussion

### *A new speech comprehension test*

A new ongoing speech comprehension test was developed that aims to resemble real communication more closely than current sentence-based laboratory tests. Specifically, the new test incorporates three key aspects of real world listening: it is *ongoing* rather than discrete, it requires *comprehension* of the heard speech, and it uses an *on-the-go* response structure that imitates (to some extent) the requirement for real-time formulation of replies in conversations. In this paper we described the development of the test, and conducted a preliminary evaluation of the effect of moving from sentence recall to comprehension in a group of listeners varying in age and hearing loss.

A novelty of our test is the introduction of the on-the-go response structure. Our intention with this format was to avoid the heavy episodic memory load present in many previous comprehension tests, while engaging some of the operational processes that might be involved in the formation of verbal responses during real communication. However, this format also introduces other cognitive requirements that may not be so relevant for one's ability to cope in real conversations. In particular, the ability of subjects to read and write may influence their performance on our task. In addition, the requirement to read and write while continuing to listen adds a dual-task component to the task. At the moment, the influence of these aspects cannot be untangled from the effects of listening in an ongoing way, or the effects of comprehension abilities per se.

It is worth noting that two of the HI listeners (both 80 years old) could not complete the comprehension test even in quiet, despite reasonable abilities on the sentence test in noise, and in fact the listeners who showed the largest increases in SRT for the comprehension test were over 70 years of age. This is somewhat surprising given that older listeners are better at using semantic context to support speech intelligibility (Pichora-Fuller et al, 1995), which is an inherent part of the comprehension task. On the other hand, both working memory and speed of processing have been found to decline with increasing age (Light & Anderson, 1985; Wingfield et al, 1985; Salthouse et al, 1996). These abilities influence one's ability to communicate effectively (Stephens & Kramer, 2010), and would also affect the ability to manage our more complex test, and in particular the on-the-go response structure. Thus this style of test may have the potential to reveal age-related cognitive limitations in real-life communication that sentence tests may not. Although we did not obtain any measures of cognitive ability as part of this experiment, we did have access to results from the reading span test, used to measure working memory capacity (e.g. Besser et al, 2013), for 19 of the HI subjects who had also participated in other studies in the laboratory at around the same time (Keidser et al, 2013a, 2014). A correlational analysis indicated that these scores were negatively associated with the change in SRT when moving from the sentence test to the comprehension test, meaning that the listeners with smaller working memory spans tended to do *worse* for comprehension than sentences, while those with larger working memory spans were more likely to do *better* at the comprehension test. This association was only weak for unaided SRTs (Spearman's $\rho = -0.25$, $p = 0.31$) but significant for aided SRTs ($\rho = -0.66$, $p = 0.003$). Further measurements on a larger subject pool and the inclusion of a range of cognitive tests will be needed to demonstrate conclusively that the ability to handle the aspects of speech communication introduced by the comprehension test is affected by cognitive ability.

### *Sentence repetition vs. comprehension*

Despite the very different characteristics of the two tests, the two sets of SRTs were highly correlated. As a result, we would not expect the comprehension test in its current form to be able to better predict real-world abilities of individual participants. Indeed, correlations with the self-report data were only marginally better for the comprehension test than for the sentence test. On the other hand, it is encouraging that we have developed a test with increased ecological validity, and which participants found more engaging, that is still able to broadly capture speech understanding as would be measured by a standard sentence test.

The close relationship between scores in the two tests for many subjects may be due to the fact that audibility of the target speech in the noise was the primary limitation, as suggested by the strong association of 4FAHL with unaided SRTs. This is consistent with several previous studies that have found that the performance of listeners with and without hearing loss on speech comprehension tasks (where questions are asked *after* listening to a passage) can be equated by adjusting the level of the materials on an individual basis to equate word intelligibility scores (Schneider et al, 2000; Murphy et al, 2006;

Gordon et al, 2009). We also speculate that under the given test conditions (e.g. with a background of conventional babble noise), the extra cognitive processes introduced in the comprehension test were still managed within the cognitive capacity of most of our listeners. It is also possible that in relation to speech comprehension, cognitive limitations in the capacity of attentional and working memory resources were counterbalanced by utilization of linguistic knowledge and contextual support (Pichora-Fuller et al, 1995; Wingfield & Tun, 2007). This would suggest that the two types of test are only likely to lead to differences in performance under more challenging listening conditions. Under less challenging listening conditions, however, it might be that the increased contextual support available in the comprehension test may actually *improve* performance over a sentence test where every word must be identified exactly.

We are aware of only a few studies that have administered a discrete sentence test and an on going comprehension test to the same population. Percy et al (2013) compared audiovisual speech comprehension to performance on more traditional speech tests in cochlear implantees. One of the traditional speech tests was the Hearing in Noise Test (HINT; equivalent to the BKB sentences used in this study) that was also presented in an audiovisual mode. The authors found significant correlations between percentage correct scores obtained for the two tests presented in quiet and at three different SNRs, although the strength of the correlations did weaken at poorer SNRs. The other tests included a word test and a sentence test with an accented talker, both presented in audio-only mode. Performances on these tests also correlated significantly with performances on the audiovisual speech comprehension test in quiet, but the associations were weaker than for the HINT test and disappeared at the poorest SNR. These data partly support our hypothesis that the two types of test are more likely to provide different information under more challenging listening conditions. Tye-Murray et al (2008) did not directly compare performances on a sentence test and a comprehension test, but compared relative differences measured with each test. Specifically, they observed that older adults with normal hearing were more affected by distorted audiovisual stimuli than young adults when presented with a sentence test than when presented with a speech comprehension test. While it is difficult to compare these results to the current results due to many methodological differences, their results again suggest that the difficulty of the listening condition influences comparisons between sentence and comprehension tests.

### *Future work*

Although we have demonstrated the basic feasibility of this new comprehension test in listeners with hearing loss and provided an initial evaluation, further work is required before it can be put to practical use. Most critically, a much larger set of normative data would allow a thorough evaluation of the test in terms of passage equivalence and test-retest reliability. Furthermore, there are several variations on the test that might be of interest. For example, here we only examined speech comprehension in a relatively simple babble background. Future work will move this test into more realistic and challenging background environments containing reverberation and competing speech, such as one would encounter when conversing in a noisy restaurant (Best et al, 2015). In such a situation, distraction from nearby talkers could interfere more strongly with comprehension than with sentence recall, leading to a larger difference in performance between the two tests. Another nice feature of this test is that it can easily be extended to include multi-person conversations as targets, so that the ability of listeners to follow dynamic variations in voice and location can be examined; we are currently working on extending the test in this way.

One final issue that warrants discussion is that of validating this (or any other) new test in terms of its ability to predict real-world outcomes. As discussed earlier, one of the broad aims of this work is to provide a tool for evaluating new hearing aids or processing schemes. However, how do we tell if a new test provides better predictions than existing tests? How do we quantify real-world performance as a point of reference? The current gold standard is self-report measures, usually obtained via interviews or questionnaires, and it is possible to compare the outcomes of different tests to self-report data (as we did in the present study). But these subjective measures are variable and prone to individual biases (e.g. Kamil et al, 2015), which may make it difficult to observe subtle improvements in predictions. Another problem seems to be that self-report data are often based on a rather broad impression of one's ability in various scenarios, whereas laboratory measures tend to focus on one or a few very specific listening situations. New approaches to this issue will be needed as the field continues to move towards more realistic tests of speech communication.

## Conclusion

In this paper we introduced a new test of ongoing speech comprehension, and described a preliminary evaluation study in listeners with a range of ages and hearing losses. The new test appears to be sensitive to both sensory and cognitive factors, and shows promise for the larger goal of creating more realistic laboratory evaluations.

## Acknowledgements

The authors acknowledge the financial support of the HEARing CRC, established and supported under the Cooperative Research Centres Program, an Australian Government initiative. Virginia Best was also partially supported by NIH/NIDCD grant DC04545. We are grateful to Margot McLelland, Chris Orinos, and Harvey Dillon for their help at various stages of this study. Portions of this work were presented at the International Hearing Aid Research Conference (August 2012), the International Congress on Acoustics (June 2013), the International Conference on Cognitive Hearing Science for Communication (June 2013), and the XXXII World Congress of Audiology (May 2014).

***Declaration of interest:*** The authors declare no conflicts of interest.

## References

Baddeley A.D. 1996. The concept of working memory. In: S. Gathercole (ed.), *Models of Short-term Memory*. Hove, UK: Psychology Press, pp. 1–28.

Besser J., Koelewijn T., Zekveld A.A., Kramer S.E. & Festen J.M. 2013. How linguistic closure and verbal working memory relate to speech recognition in noise: A review. *Trends Amplif*, 17, 75–93.

Best V., Keidser G., Buchholz J.M. & Freeston K. 2015. An examination of speech reception thresholds measured in a simulated reverberant cafeteria environment. *Int J Audiol*, early online, 1–9.

Cooke M. 2006. A glimpsing model of speech perception in noise. *J Acoust Soc Am*, 119, 1562–1573.

Gatehouse S. & Noble W. 2004. The Speech, Spatial and Qualities of Hearing Scale. *Int J Audiol*, 43, 85–99.

Gordon M.S., Daneman M. & Schneider B.A. 2009. Comprehension of speeded discourse by younger and older listeners. *Exp Aging Res*, 35, 277–296.

Hafter E.R, Xia J. & Kalluri S. 2013. A naturalistic approach to the cocktail party problem. In: B.C.J. Moore, R.D. Patterson, I.M. Winter, R.P. Carlyon & H.E. Gockel (eds.) *Basic Aspects of Hearing: Physiology and Perception*. New York: Springer, pp. 527–534.

Hagerman B. 1982. Sentences for testing speech intelligibility in noise. *Scand Audiol*, 11, 79–87.

Humes L.E. & Dubno J.R. 2010. Factors affecting speech understanding in older adults. In: S. Gordon-Salant, R.D. Frisina, A.N. Popper & R.R. Fay (eds.) *The Aging Auditory System*. New York: Springer, pp. 211–257.

Jakeman V. & McDowell C. 1995. *Cambridge English IELTS 1 with Answers: Authentic Examination Papers from Cambridge ESOL*. Cambridge: Cambridge University Press.

Kalikow D.N., Stevens K.N. & Elliott L.L. 1977. Development of a test of speech intelligibility in noise using sentence materials with controlled word predictability. *J Acoust Soc Am*, 61, 1337–1351.

Kamil R.J., Genther D.J. & Lin F.R. 2015. Factors associated with the accuracy of subjective assessments of hearing impairment. *Ear Hear*, 36, 164–167.

Kei J. & Smyth V. 1997. Measuring the ability of hearing impaired children to understand connected discourse: A comparison of two methods. *Br J Audiol*, 31, 283–297.

Kei J., Smyth V., Burge E., Fernando S., Fiteni R. et al. 2003. Measuring the ability of children to understand everyday speech using computer technology: A normative study. *Asia Pacific J Speech, Lang Hear Res*, 8, 235–242.

Keidser G., Ching T., Dillon H., Agung K., Brew C. et al. 2002. The National Acoustic Laboratories' (NAL) CDs of speech and noise for hearing-aid evaluation: Normative data and potential applications. *The Australian and New Zealand Journal of Audiology*, 24, 16–35.

Keidser G., Dillon H., Convery E. & Mejia J. 2013a. Factors influencing inter-individual variation in perceptual directional microphone benefit. *J Am Acad Audiol*, 24, 955–968.

Keidser G., Dillon H., Mejia J. & Nguyen C.V. 2013b. An algorithm that administers adaptive speech-in-noise testing to a specified reliability at selectable points on the psychometric function. *Int J Audiol*, 52, 795–800.

Keidser G., Walravens E. & Mejia J. 2014. Is a non-auditory profile associated with bilateral directional processing benefit in challenging listening situations? *IHCON*. Lake Tahoe, USA.

Kidd G., Jr. Mason C.R., Richards V.M., Gallun F.J. & Durlach N.I. 2008. Informational masking. In: W.A. Yost, A.N. Popper & R.R. Fay (eds.) *Auditory Perception of Sound Sources*. New York: Springer Handbook of Auditory Research, pp. 143–189.

Light L.L. & Anderson P.A. 1985. Working memory capacity, age, and memory for discourse. *J. Gerontology*, 40, 737–747.

Lunner T., Rudner M. & Rönnberg J. 2009. Cognition and hearing aids. *Scandinavian Journal of Psychology*, 50, 395–403.

MacPherson A. & Akeroyd M. 2013. The Glasgow monitoring of uninterrupted speech task (GMUST): A naturalistic measure of speech intelligibility in noise *Proceedings of Meetings on Acoustics (POMA)*. Montreal, Canada.

Miller G.A. & Licklider J.C.R. 1950. The intelligibility of interrupted speech. *J Acoust Soc Am*, 22, 167–173.

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

Nyborg L., Backman L., Erngrund K., Olofsson U. & Nilsson L.-G. 1996. Age differences in episodic memory, semantic memory, and priming: Relationships to demographic, intellectual, and biological factors. *J Gerontology: Psych Sci*, 51B, 234–240.

Percy V., Raymond B., Smith A., Joseph C., Kronk L. et al. 2013. Measuring speech perception abilities in adults with cochlear implants: Comprehension versus speech recognition. *The Australian and New Zealand Journal of Audiology*, 33, 35–47.

Pichora-Fuller M.K. 2007. Audition and cognition: What audiologists need to know about listening. In: C. Palmer & R. Seewald (eds.), *Hearing Care for Adults*. Stäfa, Switzerland: Phonak, pp. 71–85.

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

Rönnberg J. 2003. Cognition in the hearing impaired and deaf as a bridge between signal and dialogue: A framework and a model. *Int J Audiol*, 42, S68–S76.

Rönnberg J., Lunner T., Zekveld A., Sörqvist P., Danielsson H. et al. 2013. The ease of language understanding (ELU) model: Theoretical, empirical, and clinical advances. *Front Syst Neurosci*, 13, 1–17.

Salthouse T.A., Hancock H.E., Meinz E.J. & Hambrick D.Z. 1996. Interrelations of age, visual acuity, and cognitive functioning. *J Gerontology B Psychol Sci Soc Sci*, 51, 317–330.

Schneider B., Daneman M., Murphy D. & See S. 2000. Listening to discourse in distracting settings: The effects of aging. *Psychol Aging*, 15, 110–125.

Schneider B.A., Pichora-Fuller M.K. & Daneman M. 2010. Effects of senescent changes in audition and cognition on spoken language comprehension. In: S. Gordon-Salant, R.D. Frisina, A.N. Popper & R.R. Fay (eds.), *The Aging Auditory System*. New York: Springer, pp. 167–210.

Sommers M.S., Hale S., Myerson J., Rose N., Tye-Murray N. et al. 2011. Listening comprehension across the adult lifespan. *Ear Hear*, 32, 775–781.

Tye-Murray N., Sommers M., Spehar B., Myerson J., Hale S. et al. 2008. Auditory-visual discourse comprehension by older and young adults in favorable and unfavorable conditions. *Int J Audiol*, 47, S31–S37.

Wichmann F.A. & Hill N.J. 2001. The psychometric function: I. Fitting, sampling, and goodness-of-fit. *Perc Psych*, 63, 1293–1313.

Wingfield A., Poon L.W., Lombardi L. & Lowe D. 1985. Speed of processing in normal aging: Effects of speech rate, linguistic structure, and processing time. *J Gerontology*, 40, 579–585.

Wingfield A. & Tun P. 2007. Cognitive supports and cognitive constraints on comprehension of spoken language. *J Am Acad Audiol*, 18, 548–558.

## Appendix

The 28 monologues were taken from Volumes 1–8 of the IELTS books of past examinations (and two books of practice tests) as follows, where TxSx indicates the Test number and Section number within the volume.

Volume 1: T1S2, T2S2, T2S4, T3S2, T4S2
Volume 2: T1S4, T2S2, T3S2, T4S2
Volume 3: T2S2, T3S2, T4S2, T4S4
Volume 4: T1S2, T3S2, T3S4, T4S2
Volume 5: T1S2, T4S2
Volume 6: T3S2, T4S2
Volume 7: T3S2
Volume 8: T1S2, T2S2, T3S2
Practice book 1: T1S2
Practice book 2: T1S2