Automated Screening for High-Frequency Hearing Loss

Marcel S. M. G. Vlaming,[1,2] Robert C. MacKinnon,[1] Marije Jansen,[1] and David R. Moore[1,3,4]

Objective: Hearing loss at high frequencies produces perceptual difficulties and is often an early sign of a more general hearing loss. This study reports the development and validation of two new speech-based hearing screening tests in English that focus on detecting hearing loss at frequencies above 2000 Hz.

Design: The Internet-delivered, speech-in noise tests used closed target-word sets of digit triplets or consonant–vowel–consonant (CVC) words presented against a speech-shaped noise masker. The digit triplet test uses the digits 0 to 9 (excluding the disyllabic 7), grouped in quasi-random triplets. The CVC test uses simple words (e.g., "cat") selected for the high-frequency spectral content of the consonants. During testing, triplets or CVC words were identified in an adaptive procedure to obtain the speech reception threshold (SRT) in noise. For these new, high-frequency (HF) tests, the noise was low-pass filtered to produce greater masking of the low-frequency speech components, increasing the sensitivity of the test for HF hearing loss. Individual test tokens (digits, CVCs) were first homogenized using a group of 10 normal-hearing (NH) listeners by equalizing intelligibility across tokens at several speech-in-noise levels. Both tests were then validated and standardized using groups of 24 NH listeners and 50 listeners with hearing impairment. Performance on the new high frequency digit triplet (HF-triplet) and CVC (HF-CVC) tests was compared with audiometric hearing loss, and with that on the unfiltered, broadband digit triplet test (BB-triplet) test, and the ASL (Adaptive Sentence Lists) speech-in-noise test.

Results: The HF-triplet and HF-CVC test results (SRT) both correlated positively and highly with high-frequency audiometric hearing loss and with the ASL test. SRT for both tests as a function of high-frequency hearing loss increased at nearly three times the rate as that of the BB-triplet test. The intraindividual variability (SD) on the tests was about 2.1 (HF-triplet) and 1.7 (HF-CVC) times less than that for the BB-triplet test. The effect on the HF-triplet test of varying presentation method (professional or cheap headphones and loudspeakers) was small for the NH group and somewhat larger, but nonsignificant for the hearing-impaired group. Test repetition produced a moderate, significant learning effect for the first and second retests, but was small and nonsignificant for further retesting. The learning effect was about two times larger for the HF-CVC test than for the HF-triplet test. The sensitivity of both new tests for high-frequency hearing loss was similar, with an 87% true-positive and 7% false-positive ratio for detecting an average high-frequency hearing loss of 20 dB or more.

Conclusions: The new HF-triplet and HF-CVC tests provide a sensitive and accurate method for detecting high-frequency hearing loss. The tests may signal developing hearing impairment at an early stage. The HF-triplet is preferred over the HF-CVC test because of its smaller learning effect, smaller error rate, greater simplicity, and lower cultural dependency.

Key words: Hearing impairment, Hearing rehabilitation, Hearing screening, High-frequency hearing loss, Speech intelligibility.

(Ear & Hearing 2014;35;667–679)

INTRODUCTION

Many hearing impairments are characterized by an audiometric loss that develops slowly, affecting the higher frequencies first and extending gradually in the lower frequencies. Two of these types of hearing loss are age-related (presbycusis) and noise-induced hearing loss (NIHL). Presbycusis begins to develop relatively early; about 25% of males aged above 55 years have a hearing loss ≥35 dB at frequencies above 4 kHz (Robinson & Sutton 1979). NIHL can affect people at any age and is a result of exposure to damagingly loud sounds, usually from an occupational source (e.g., industry, gunshots or explosions, professional musicians), but also potentially from recreational sources (e.g., motor vehicles, loud music; Morata 2007). NIHL is also characterized by an increase in high-frequency audiometric threshold, but while presbycusis appears as a gradual sloping loss, NIHL manifests as a much sharper hearing loss that may initially be restricted to high frequencies. It is also sometimes accompanied by a specific region of increased threshold, between ~3 and 6 kHz, commonly known as the NIHL "dip" or "notch."

In both cases, the high-frequency hearing loss is unnoticed at first by a listener because overall perception is dominated by low-frequency hearing. Also, the slow progress of hearing impairment makes the gradual change in hearing more difficult to detect. However, high-frequency hearing loss is associated with and may underpin everyday hearing problems, including difficulty in speech understanding in noisy environments. This can impair communication in social contacts and the ability to work. Because of the gradual buildup of hearing impairment, people may delay or fail to seek professional help (Trumble & Pitterman 1992). This is one of the reasons why many people who could benefit from hearing aids do not use them (Davis et al. 2007; Dawes et al. 2014).

Remotely deliverable screening tests have been developed that can detect hearing loss reliably, quickly, and easily and that are accessible by telephone (Smits et al. 2004; Smits & Houtgast 2005; Watson et al. 2012) or Internet (Smits et al. 2006; Vlaming et al. 2011; Jansen et al. 2013). These tests present short words in the presence of background noise (speech-in-noise testing). Users are tasked to correctly identify the words, which are most commonly a digit triplet sequence (in English, from the 9 monosyllabic numbers 0–9, e.g., 5-2-8). Using adaptive procedures, the signal (speech) to noise ratio (SNR) can be found where 50% of the words are correctly identified. This is termed the speech reception threshold (SRT) in noise. The SRT is largely independent of the absolute presentation level over a wide dynamic range (e.g., Plomp 1986; Wagener & Brand 2005) so that testing can be performed without level calibration. Tests use closed sets of tokens (e.g., digits,

[1]National Institute for Health Research, Nottingham Hearing Biomedical Research Unit, Nottingham, UK; [2]ENT-Audiology, VU University Medical Center, Amsterdam, The Netherlands; [3]MRC Institute of Hearing Research, Nottingham, UK; and [4]Communication Sciences Research Center, Cincinnati Children's Hospital, Cincinnati, USA.

consonant–vowel–consonant (CVC) words) and are simple enough to be self-administered by users with a wide range of ages (4 to >95 years) and intellectual capabilities (Smits et al. 2013; C. Smits, personal communication). Digits are particularly convenient because responses can be made on a telephone or computer keypad (Smits et al. 2004; Smits & Houtgast 2005). The background noise used in the test is typically stationary speech-shaped noise, derived from the target speech material to have the same average spectral shape. Digit triplet-based screening tests have been developed for several languages (Ozimek et al. 2009; Jansen et al. 2010, 2013; Vlaming et al. 2011; Zokoll et al. 2012) and have been publically available as website- and telephone-based screening services for several years. They have high levels of sensitivity and specificity, and the results correlate strongly (*r* = 0.74–0.86) with audiogram pure-tone average (PTA) measures (Smits et al. 2004; Watson et al. 2012; Jansen et al. 2013).

Tests of speech perception should ideally mimic the everyday situation in which the speech must be understood in a background of noise generated from other speakers. However, for screening, it may be preferable to optimize the sensitivity of the hearing test rather than to mimic everyday listening. Leensen et al. (2011) developed a number of screening tests in the Dutch language based on the understanding of nine CVCs. They used different variants of the masking noise, including low-pass filtering and temporal modulation, in an attempt to optimize detection of high-frequency hearing loss. They found that low-pass filtered noise made a CVC-based test more sensitive. Modulated masking noise also improved sensitivity, but not as much as the low-pass filtered noise. Combining low-pass filtering and temporal modulation was less effective than low-pass filtering alone. The increased sensitivity of the test is likely due to the increased requirement to understand high-frequency speech components when the low speech frequencies are selectively masked.

This study reports on the development and validation of new 'high frequency' digit triplet (HF-triplet) and CVC (HF-CVC) tests in British English with increased sensitivity for detecting high-frequency hearing loss.

MATERIALS AND METHODS

Development of Test Materials

The design for the new hearing tests is based on the method used for developing the original Dutch digit triplet test (DTT) (Smits et al. 2004).

Speech Targets

The digits were initially recorded from two female speakers selected to have a typical, southern English accent. The recorded speech patterns were assessed, and one speaker was dismissed on the bases of accent and some hesitant speech (Kent & Read 2002). The digit triplets were spoken at normal effort and speed. Recordings were made in a sound-attenuating chamber using a Sennheiser MKH 40 P48 microphone and an Audiofire digital sound card at 44.1 kHz sample rate and stored as 16 bit.wav files. The digit 'seven' was omitted as this digit is bisyllabic, whereas the other digits 0–9 are all monosyllabic. The digit 'zero' was pronounced /o/ (as in 'soap'). The remaining nine digits were recorded as triplets. The first digit was preceded (200 ms pause) by a trailer, "The numbers …." Recordings of each digit were obtained in three variants: as first (a), middle (b) and last (c) position in the triplet. A list of 18 triplets was made where each digit occurred twice at each position. This list was recorded six times. From these repeated recordings, the most natural sounding nine digits for the three positions (a,b,c) were selected to give 27 digits and one trailer. The 27 individual digits and the one trailer were isolated from the recorded triplets using Adobe Audition 3.0 by cutting at the zero crossings on visual inspection of the speech pattern. The test triplets were constructed by connecting three digits keeping the normal timing between the digits. The trailer was appended in front of each triplet. Total presentation length was constant, as determined by the noise burst (below).

The English CVC words were recorded from the same female speaker used for the triplets. The CVC words (initially *n* = 130) were selected from different CVC lists used for children (Roeser 1996; Lanternfish 2010). Each CVC was recorded four times, and the clearest, most natural sounding examples were selected. From these 130 CVC selected recordings, the consonants and vowels were isolated, and their spectral content was analyzed using a 1024-point Fourier transform. The CVC words were then ranked according to the highest relative spectral power of the consonants for frequencies in a band between 3 and 5 kHz. CVC words having the two vowels /i/ and /a/ had the most consonant energy in that band, so the list was restricted to those words to reduce recognition based more on vowels and less on consonants. The highest ranked CVC words were used to construct two lists of 12 CVC words that, through the homogenization tests (see below), were reduced to one set of the best 12 CVC words. No trailer was used for CVCs.

Masking Noise

A quasi-stationary masking noise with the same spectrum as the speech test materials was constructed for each of the HF-triplet and HF-CVC tests. For the HF-triplet test, the noise was made by repeatedly superimposing the triplets according to the procedure described by Wagener et al. (2003) and also used for the HearCom BB-triplet tests (Vlaming et al. 2011). For the HF-CVC test, the 24 CVCs were repeatedly superimposed by the same procedure. The low-pass (LP) noise version (for HF tests) was constructed by filtering (LP cutoff frequency 1500 Hz) using a tenth order Butterworth filter and summing with the original noise attenuated by 15 dB. Age-related hearing loss typically has a 1–2 kHz lower limit (Dubno et al. 2013). For the LP-triplet noise, a fragment of 4170 msec was selected, and for the LP-CVC noise a fragment of 1300 msec was chosen. These noises started 500 ms before the speech (with trailer, for the triplets) and finished at the same time as the speech. They were faded in-out by a 5 ms ramp. The spectra of unfiltered and filtered masking noise for the HF-triplets and the HF-CVCs are shown in Figure 1.

Homogenization

Background and Preliminary Procedure. Speech intelligibility as a function of SNR can be described by a psychometric function (Brand and Kollmeier 2002; Jansen et al. 2010):

$$\mathrm{SI}\left(\mathrm{SNR}\right) = \gamma + (1-\gamma)\frac{1}{1+\mathrm{e}^{4s(\mathrm{SRT}-\mathrm{SNR})}} \qquad (1)$$

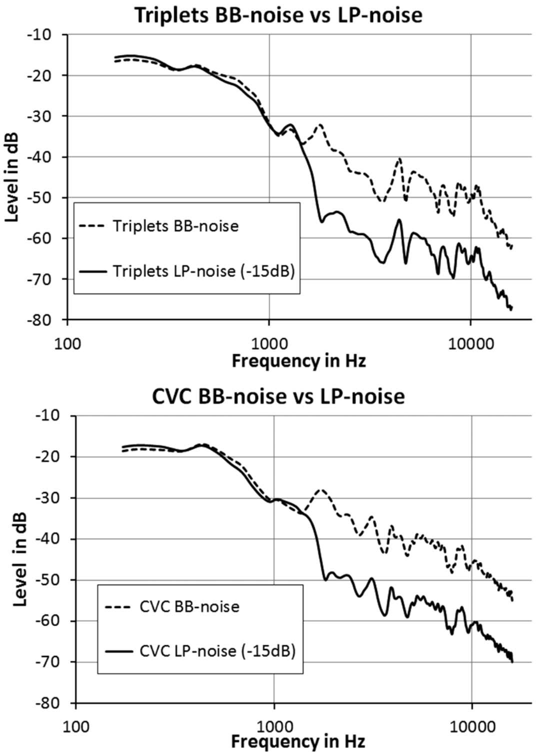

Fig. 1. Spectra of the broadband (BB) and low-pass (LP) noise spectra for the triplets (A) and consonant–vowel–consonant (B) tests shown at relative sound levels at a fixed 86 Hz bandwidth.

Where SI, speech intelligibility; γ, guess level; SRT, speech reception threshold; SNR, signal to noise ratio; and *s*, the slope at the SRT.

The SRT is the SNR at 50% speech intelligibility. For a single digit having 10 alternatives (the user does not know that "7" is not presented), the guess level is 1/10 and the above SRT will correspond to 55% speech intelligibility. For the CVC test with 12 CVC words, the guess level is 1/12 and the above SRT corresponds to 54.16% speech intelligibility. For the triplets test consisting of three digits, the guess level will be 1/1000 and the SRT corresponds to 50.05%.

The individual digit and CVC tokens were "homogenized" with respect to SRT. This helped ensure that each digit or CVC had an equal chance of being selected without bias generated by speech intelligibility. To achieve this, the speech intelligibility (psychometric) function of each digit and each CVC was averaged across 10 NH participants in a round of homogenization tests. From that function, the mean SRT was determined for each digit and each CVC. Next, the level of each digit or CVC was shifted to the mean SRT such that the psychometric functions of each digit or CVC coincided at the mean SRT point.

A set of 36 triplets was developed in which all 27 digit tokens were presented four times. These 36 triplets were then mixed with the triplets LP masking noise at 10 SNR levels varying by 2 dB SNR, making a total 360 triplets in noise. Two sets of 12 CVCs were developed at 15 levels of SNR varying by 2 dB SNR, a total of 360 CVCs in noise.

Participants. All participants in the study signed an informed consent for participation. All homogenization testing was performed using a single group of 10 NH participants, each having UK-English as their mother tongue and with normal hearing (maximum audiogram thresholds in each ear of 20 dB HL at frequencies 0.25, 0.5, 1, 2, 3, 4, and 6 kHz and 25 dB at 8 kHz, with asymmetry at each frequency of not more than 15 dB).

Procedure. An Internet-delivered test program was developed to present the test stimuli at the different SNRs in a random order. The test stimuli were presented diotically in a sound attenuating chamber using Sennheiser HD 25-1 II headphones via an Audiofire digital sound card. For presenting through the Internet, the sound files were MP3 coded at a bit rate of 192 kb/s to reduce the download time for each triplet. The test program ran on a server of the University of Nottingham. After presentation of the full stimulus, the listener selected by mouse clicking the three digits or the one CVC, confirmed by clicking the 'submit' field, after which the next stimulus was presented. If the participant failed to recognize the stimulus, they were advised to make a best guess. A selection was a requirement for presentation of the next stimulus.

Homogenization testing of the triplets was completed in one session for each participant. A fixed noise presentation level of 65 dBA was used, as measured at the headphones by an artificial ear (B&K 4153). Each participant was presented with the 360 triplets in noise in random order. After 180 triplets, a short break was taken. The total test of 360 triplets took about 45–60 min to complete.

Homogenization testing of the CVCs was completed in two sessions for each participant. In the first session, two tests (1A and 1B) were completed, each involving presentation of 180 CVCs in noise in random order (i.e., a total of 2 sets of 12 CVCs at 15 SNR levels). Noise level was fixed at 65 dBA. Each test took about 30 min including a short break between tests. From the results of the first session on all 24 CVCs, a final set of 12 CVCs was chosen as those having the steepest slope of the psychometric function. In the second session (test 2), 3–6 weeks after session 1, the procedure was repeated for the final set of 12 CVCs at 13 SNR levels (156 CVCs).

HF-Triplet and HF-CVC Tests

The new HF-triplet and HF-CVC tests were constructed from the homogenized triplets and CVCs. For each test, 25 triplets or CVCs are presented, and the SNR level is varied adaptively to determine the SRT. The speech level is varied while the background noise is kept at a constant level, as for the UK BB-triplet test (Phipps 2007; Vlaming et al. 2011). Following a correct response, the SNR is reduced by 2 dB (making the next trial more difficult), and following an incorrect response, it is increased by 2 dB. The initial SNR is −14 dB, about 8–10 dB above the expected SRT for NH listeners. The SRT is calculated as the mean SNR of the last 19 stimulus presentations. It is not practical to prescribe a fixed sound level because of the variation in audio hardware. A demo triplet or CVC is thus presented prior to each test. The demo stimuli have a fixed SNR of −4 dB so that they can be understood by all normal hearing and most hearing-impaired listeners. The participant is asked to adjust the overall presentation sound level (i.e., noise + signal) to a "comfortable volume" by moving a slider and confirm by clicking an enter field. That level was increased during testing by an extra 3 dB to provide some extra margin for detecting speech at levels that may be close to the threshold of hearing.

Validation

Background and Overview. The new tests were validated by (i) assessing their relation to the audiogram PTA, (ii) comparing them with the conventional, broadband triplet test (BB-triplet), and (iii) comparing them with the clinically used ASL sentence speech-in-noise test (Macleod & Summerfield 1990). Performance on the new tests was also characterized by the measured SRT and psychometric function slope for comparison with other DTT speech-in-noise tests (Plomp 1986; Leensen et al. 2011; Vlaming et al. 2011). The slope of the psychometric function is indicative of the ability of the adaptive procedure to reach a precise SRT: the steeper the psychometric function, the faster and more accurately the SRT will be obtained.

The new tests are designed for use through the Internet for which various setups for audio presentation are used (e.g., headphones or loudspeakers, usually of unknown quality and characteristics). We assessed how these different audio setups affected test outcome measures. Outcome measures will also be affected by intraindividual variability that, because of repeated test use over time, may include training. This "test–retest reliability" was assessed by comparing measures within and between sessions.

Participants. Normal-hearing participants (NH group; *n* = 24; audiometrically verified) were recruited by advertisement in a local newspaper. Participants with hearing impairment (HI group; *n* = 50) were recruited from people previously diagnosed at Nottingham Audiology Services. Candidates in the NH group with hearing loss interaural asymmetry of more than 20 dB (0.125–4 kHz) were excluded. Candidates with conductive loss >10 dB (PTA; 0.125–4 kHz) were excluded from both groups. Eight participants in the HI group had interaural asymmetry >20 dB. Figure 2 shows the mean (and SD) hearing level for both groups. The age of the NH group ranged from 18 to 47 years (mean of 29.8 years) and that of the HI group ranged from 31 to 75 years (mean of 63.4).

Procedure. All tests were performed in a quiet lab room, except for two tests completed at home. The new HF-triplet and HF-CVC tests and the BB-triplet test (Phipps 2007; Vlaming et al. 2011) were Internet-delivered. A login procedure with individual codes was used for participant identification and storage of results. Sound stimuli were presented diotically using, initially, Sennheiser HD 25-1 II headphones via a digital sound card (Echo Audiofire 4). In the separate "audio setup" tests, stimuli were presented via three transducers considered representative of setups used at home. The first was a set of cheap headphones, for example, as offered in economy class of planes. The second was the two built-in, small loudspeakers of an IIyama PC monitor (30 cm L/R separation at about 80 cm distance from the participant). The third was two good quality loudspeakers (Genelec 8030A) at 80 cm L/R separation and about 140 cm from the participant. The responses of the headphones were measured using a B&K 4153 artificial ear. For free-field (loudspeaker) measurements, the sound was recorded at the center position of the head (substitution method). Recordings were made using third octave analysis and a B&K 2250 sound level meter connected to the digital sound card. In Figure 3, the frequency-response of all four transducers is presented for a constant input level white noise (10 Hz–20 kHz) stimulus. The differences between the response functions of the four transducer types used were less than 20 dB for a constant input signal. The ASL speech-in-noise sentence test was presented through a PC program, as used clinically at Nottingham Audiology Services. Audiograms were measured using a Unity audiometer.

At a preliminary session, the ears of all candidate participants were inspected using an otoscope, and a full audiogram (0.25–8 kHz) was recorded. After acceptance for testing, the participant was instructed on the tests. Validation testing was performed in five sessions for the NH group and in two sessions for the HI group (Table 1). At session 1, the NH group performed both HF-triplet and HF-CVC tests four times (1.1–1.4) and the HI group three times (1.1–1.3). Each test takes about 3–5 min. Both groups then performed the BB-triplet test (1.1). At session 2 (10 ± 5 days later), all participants performed both HF tests using the four transducer types, as outlined above. Next, the BB-triplet test was performed, followed by the ASL sentences-in-noise test. The ASL test used 30 sentences presented adaptively against a 60 dBA steady-state noise mixed from the sentences. Further details of the ASL implementation are in Ferguson et al. (2014).

At the end of session 2, the participants were asked to perform both new HF tests at home using their own home computer. For this, each participant received a URL link with personal identifier code. If no home computer was available or when other objections existed, this home test was skipped for that participant. At sessions 3 and 4, the NH participants performed both HF tests

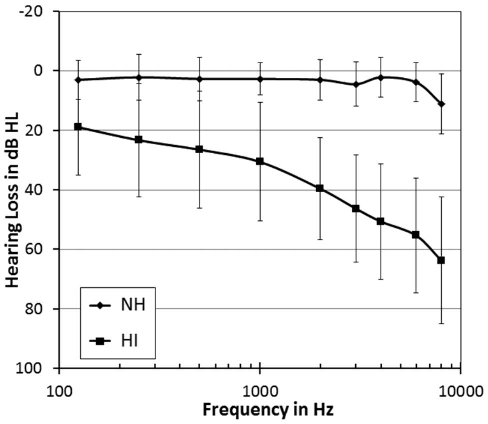

Fig. 2. Mean (±SD) hearing loss for the normal-hearing and hearing-impaired groups.

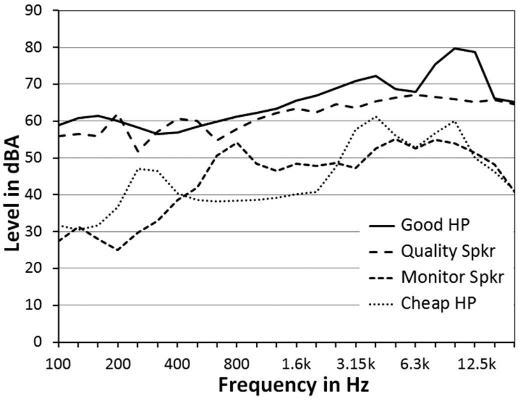

Fig. 3. Frequency response functions of the transducers (good quality and cheap headphones, HP; good quality and small monitor loudspeakers, Spkr). Measured are third-octave response sound levels in decibels, A-weighted [dBA].

TABLE 1. Overview of validation tests and testing order

	Normal Hearing (*n* = 24)	Hearing Impaired (*n* = 50)
Session 1	4 × HF-triplet (1.1–1.4)	3 × HF-triplet (1.1–1.3)
	4 × HF-CVC (1.1–1.4)	3 × HF-CVC (1.1–1.3)
	BB-triplet (1.1)	BB-triplet (1.1)
Session 2	HF-triplet (2/Good HP)	HF-triplet (2/Good HP)
	HF-triplet (Cheap HP)	HF-triplet (Cheap HP)
	HF-triplet (Small Spkr)	HF-triplet (Small Spkr)
	HF-triplet (Good Spkr)	HF-triplet (Good Spkr)
	HF-CVC (2/Good HP)	HF-CVC (2/Good HP)
	HF-CVC (Cheap HP)	HF-CVC (Cheap HP)
	HF-CVC (Small Spkr)	HF-CVC (Small Spkr)
	HF-CVC (Good Spkr)	HF-CVC (Good Spkr)
	BB-triplet (2)	BB-triplet (2)
	ASL test	ASL test
Home	HF-triplet (Home Spkr, *n* = 13)	HF-triplet (Home Spkr, *n* = 30)
	HF-CVC (Home Spkr, *n* = 15)	HF-CVC (Home Spkr, *n* = 31)
Session 3	HF-triplet (3)	
	HF-CVC (3)	
	BB-triplet (3)	
Session 4	HF-triplet (4)	
	HF-CVC (4)	
	BB-triplet (4)	
Session 5	HF-triplet (PM-function; *n* = 17)	
	HF-CVC (PM-function; *n* = 17)	

See text for details.
HF-triplet, high frequency digit triplet; HF-CVC, high frequency Consonant–Vowel–Consonant; BB-triplet, broadband triplet; HP, headphone; PM, psychometric.

and the BB-triplet test. Finally, in session 5, 17 NH participants performed one HF-triplet and two HF-CVC psychometric tests. The remaining 7 NH participants were unable to return for this session. Session 5 psychometric tests followed a similar procedure as for the homogenization tests described above. Each of the 36 HF-triplets and 12 HF-CVCs was presented at eight SNR levels. The resulting set of 288 triplets in noise was presented once and the set of 96 CVCs in noise was presented twice at a fixed noise level of 80 dBA. This level was based on those chosen in session 4 by the first 14 participants (HF-triplets, 81 dBA; HF-CVCs, 78.5 dBA). These levels appear to be high for the HF tests, where the average level used chosen by the same listeners of the BB-triplet test was 71.5 dBA.

RESULTS

Results of Homogenization Testing of the Triplets

The scores for each token as a function of SNR were averaged across the 10 NH listeners of the homogenization tests. These scores were fitted with the logistic function [Equation 1] for slope and SRT using a maximum-likelihood estimation method. Results for each of the nine digit tokens in a, b, and c positions are shown in Table 2.

The digit “4” had the lowest abc average slope of 6.9%/dB. The mean of the remaining eight digits was 13.1%/dB. This digit “4” also had the largest mean absolute deviant SRT (dSRT) of 9.28 dB, compared to the mean dSRT for the remaining eight digits of only 2.71 dB. To improve the homogeneity of the final tests, it was decided to remove the digit “4” from the list of tokens. The shift (the dSRT excluding the digit “4”) was the shift in SNR applied to the tokens for the final triplets test. The deviant SRT and slope values for the token “4” were most likely caused by the relative lack of high-frequency power compared to the other digits. Note that the removal of a single digit has only a small effect on guess characteristics. Even if a participant was to know or suspect that the tokens “7” and “4” were missing, it would reduce the guess factor of the whole triplet from 1/729 (one missing digit) to 1/512 (two missing digits), which still is close to zero. Also digit triplets tests in some other languages have eight or less digits, for instance eight digits in Dutch and seven digits in Greek (Vlaming et al. 2011).

An estimation of the final triplet psychometric function was modelled by first fitting psychometric functions (at 2 dB SNR resolution) for each of the eight homogenized digits at each of the three digit positions (abc). From that an average function for each position was calculated. The triplet function was then

TABLE 2. Individual and averaged fitted slopes and SRTs for the nine digits in each of the three positions (abc) for the triplet

	a-Position				b-Position				c-Position				Average abc		
Digit	Slope (%)	SRT	dSRT	Shift	Slope (%)	SRT	dSRT	Shift	Slope (%)	SRT	dSRT	Shift	Slope (%)	SRT	dSRT
0	9.9	−30.67	−5.20	−4.04	7.7	−24.32	1.15	2.31	8.4	−24.88	0.59	1.75	8.7	−26.62	−1.15
1	12.1	−24.60	0.87	2.03	8.3	−22.77	2.70	3.86	6.5	−21.17	4.30	5.46	9.0	−22.85	2.62
2	18.4	−32.60	−7.13	−5.97	11.3	−29.96	−4.49	−3.33	11.5	−29.92	−4.45	−3.29	13.7	−30.83	−5.36
3	24.2	−24.55	0.92	2.08	16.1	−24.81	0.66	1.82	7.6	−25.45	0.02	1.18	16.0	−24.93	0.54
4	6.9	−20.89	4.59		6.2	−12.89	12.58		7.6	−14.80	10.67		6.9	−16.19	9.28
5	8.1	−22.74	2.73	3.89	10.6	−24.47	1.00	2.16	13.3	−20.06	5.41	6.57	10.7	−22.42	3.05
6	20.3	−31.57	−6.10	−4.94	8.1	−31.71	−6.24	−5.08	6.7	−29.01	−3.54	−2.38	11.7	−30.76	−5.29
8	29.2	−29.03	−3.56	−2.40	9.7	−29.97	−4.50	−3.34	14.3	−28.08	−2.61	−1.45	17.8	−29.03	−3.56
9	16.1	−27.74	−2.27	−1.11	21.0	−28.74	−3.27	−2.11	14.0	−20.33	5.14	6.30	17.0	−25.60	-0.13
Average	16.1	−27.15	−1.68		11.0	−25.52	−0.05		10.0	−23.74	1.73		12.4	−25.47	0.00
Average −“4”	17.3	−27.94		−1.31	11.6	−27.09		−0.46	10.3	−24.86		1.77	13.1	−26.63	

dSRT denotes the difference of each SRT from the (abc) average SRT. Shift denotes the difference from the average abc minus the digit /4/ and equates to the level shift applied for each digit. SRT, speech reception threshold.

TABLE 3. Fitted slopes and SRTs for the 24 CVC words, their averages and reduction to the final 12 CVC words

	Test 1A and 1B					Test 2			
	CVC	Slope (%)	SRT	dSRT	Shift	CVC	Slope (%)	SRT	Shift
1A	Cat	10.8	–23.83	–1.28	2.11	Cat	13.3	–22.91	0.10
	Dip	12.3	–21.22	1.33	–0.50	Dip	11.6	–20.32	–2.48
	Dish	6.8	–28.52	–5.97					
	Fat	15.5	–24.85	–2.29	3.13	Fat	13.9	–26.26	3.46
	Fin	8.0	–24.29	–1.74					
	Hat	11.5	–17.11	5.44	–4.61	Hat	7.6	–17.70	–5.10
	His	No fit possible							
	Hit	8.0	–30.29	–7.74					
	Kiss	16.0	–31.98	–9.42	10.26	Kiss	25.3	–32.14	9.34
	Lid	6.0	–17.22	5.33					
	Sin	10.4	–13.59	8.96	–8.12	Sin	7.9	–19.69	–3.12
	Sit	20.0	–25.59	–3.04	3.87	Sit	11.0	–24.32	1.51
1B	Dad	10.7	–18.87	3.68	–2.85	Dad	10.6	–21.26	–1.55
	Did	24.0	–16.51	6.05	–5.21	Did	15.7	–18.17	–4.64
	Fish	7.3	–36.17	–13.62					
	Fit	No fit possible							
	Kid	13.0	–21.95	0.60	0.23	Kid	9.4	–24.65	1.84
	Kin	6.5	–16.81	5.75					
	Pig	37.0	–17.69	4.86	–4.03	Pig	14.9	–20.84	–1.96
	Pit	8.2	–21.18	1.37					
	Sad	36.0	–27.42	–4.87	5.71	Sad	18.9	–25.40	2.59
	Tab	6.4	–19.92	2.64					
	Tap	8.3	–18.61	3.95					
	Zip	No fit possible							
Average		13.5	–22.55	0	0		13.3	–22.80	0

Tests 1A and 1B show the results of the first homogenization test round, where dSRT is the deviance from the average SRT score. 'Shift' is the difference from the average for the 12 final CVC words and the level shift applied to each token. In Test 2, the fitted slopes and SRTs with dSRT/shifts of the second round for the final 12 CVC words are shown.
SRT, speech reception threshold; CVC, consonant–vowel–consonant; dSRT, deviant SRT.

constructed by multiplying the three position functions at each SNR, giving the triplet curve for which a function fit was made to give slope and SRT. This resulted in a modelled SRT of −23.8 dB and a slope of 13.0%/dB.

Results of Homogenization Testing of CVCs

Session 1 speech intelligibility scores for each of the 24 CVC tokens are shown in Table 3 (Test 1A and 1B). Results for "his," "fit," and "zip" were skipped as no good fit to the psychometric function was found. The 12 remaining tokens having the highest slope were selected for test 2. The level of these tokens was shifted by values as given in Table 3 before they were retested in test 2. The results of test 2 are shown in Table 3. An estimation of the final CVC test psychometric function was modeled by averaging the SRT and slopes of the 12 individual CVC tokens. This modeled SI function gave an SRT of −22.8 dB and a slope of 13.3%/dB. The SRT corresponds to the 54.2% point of the logistic functions with a guess level of 1/12. When correcting to the 50% level, an SRT of −23.1 dB and slope of 12.1%/dB was found.

Results of Validation Testing

SRTs refer to the first test for each participant for each condition as though this was a clinical test. For reference of test performance of the new HF tests, the SRTs of the NH group were analyzed. Mean SRTs of −21.3 dB (HF-triplets; SD = 2.4 dB) and −21.1 dB (HF-CVC; SD = 2.1 dB) were found. Both compare well with the modeled SRTs from homogenization testing. For the BB-triplet test, the mean SRT of −10.3 dB (SD = 1.1 dB) compared well to that reported (−11.1 dB) by Phipps (2007).

Relation to Audiometry

SRTs for the HF-triplet and HF-CVC tests were analyzed in relation to PTA_{LF} (0.5, 1, 2, and 4 kHz) and PTA_{HF} (3, 4, 6, and 8 kHz) hearing loss (Fig. 4). Note that the "lower frequency" PTA covers the frequency range of typical PTA audiometry. Hearing loss was referred to the better ear (whether left or right). Regression lines are presented only for the HI group as those for the NH group were not significant. Three outliers from the HI group in the BB-triplet test were removed (here and for subsequent analyses) as their standard residual of regression is >3 of the remaining participants. The slopes, intercepts, and correlation of the regression lines are given in Table 4. For the HI group, PTA and SRT correlated significantly ($p < 0.05$) after Bonferroni correction ($n = 6$) for all tests. The correlations of the HF-triplet (0.79) and HF-CVC (0.82) tests with PTA_{HF} were significantly better (Hotelling's t-test and Steigler's Z-test; both $p < 0.01$) than for the BB-triplet test (0.62).

To gain further insight into the association between the HF tests and audiometry, the correlation between the SRT of each of the tests with single audiometric frequencies (best ear) was calculated as shown in Figure 5 (NH and HI groups combined). The BB-triplet test had a relatively constant correlation across all audiometric frequencies, while both HF tests showed an increased correlation for frequencies ≥1 kHz.

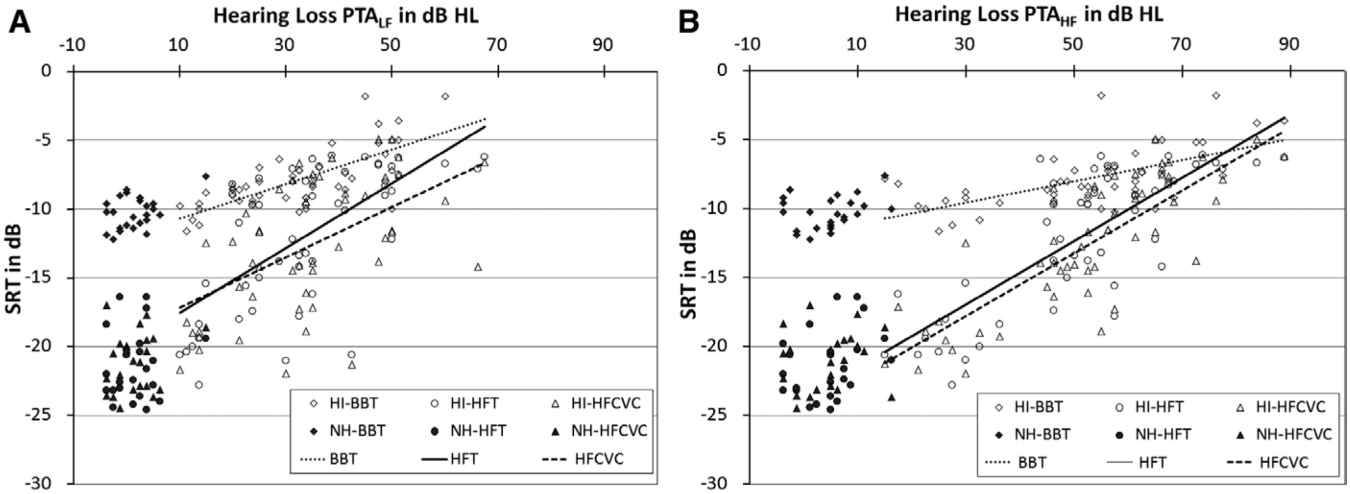

Fig. 4. Speech reception threshold (SRT) levels as a function of lower-frequency (PTA$_{LF}$, A) and high-frequency (PTA$_{HF}$, B) hearing loss. Shown are the HF-triplet (HFT), HF-CVC (HFCVC), and BB-triplet (BBT) tests for the NH and HI groups. PTA measures were for the better ear. A more negative SRT means better hearing.

Relation to ASL Sentence Test

The correlations of the HF-triplet, HF-CVC, and BB-triplet tests with the ASL sentence test and with each other are shown in Table 5 for the HI group. All correlations were significant at $p < 0.05$ after Bonferroni correction ($n = 6$). The correlations of the three tests with the ASL tests were moderate to high and did not differ much between tests. The correlations of the new HF tests with the BB-triplets were also moderate to high. The correlation between the HF-triplet and the HF-CVC tests was better (but not significantly), as expected from the similarity of the maskers used in the HF tests. For the NH group, correlations with ASL measurements could not be derived as the ASL test step size (1.2 dB) was too large in relation to the test variance (SD = 1.5 dB) of this group.

Test–Retest and Training Effects

For a reliable screening or diagnostic test, the results should reproduce when repeated over time. Reproducibility is influenced by learning; test results will improve as the participant is better able to memorize the closed set of tokens, separate the auditory characteristics of speech and noise, master the method of entering the responses, and so on. Together with these systematic effects of training, there is also an overall test variability that should remain relatively constant. The effects of repeating the test over time were analyzed separately for the NH and HI group (Fig. 6). For the NH group, the test–retest effects were tested over four sessions and, for the HI group, over two sessions. In the first session, the tests were repeated two and three times for the HI and NH groups, respectively.

TABLE 4. Intercepts (dB SNR), slopes, standard errors of slope, and correlation coefficients (Pearson *r*) of the regression lines of Figure 4

	HI			
PTA$_{LF}$	Intercept	Slope	Slope se	*R*
HF-triplet	−19.85	0.23	0.04	0.66
HF-CVC	−18.94	0.18	0.04	0.54
BB-triplet	−11.93	0.12	0.02	0.72
	HI			
PTA$_{HF}$	Intercept	Slope	Slope se	*r*
HF-triplet	−23.87	0.23	0.03	0.79
HF-CVC	−24.60	0.23	0.02	0.82
BB-triplet	−11.87	0.08	0.01	0.62

PTA, pure-tone average; HI, hearing impaired; HF-triplet, high frequency digit triplet; HF-CVC, high frequency consonant–vowel–consonant; BB-triplet, broadband triplet.

There was a general improvement on the retests relative to the first test. Analysis of the HF-triplet data (paired sample *t*-test) for the NH group showed significant improvements compared to the first test (1.1) for all subsequent tests in session 1 (1.2, 1.3, and 1.4) and in sessions 2–4. However, none of the differences between the second test in session 1 (1.2) and the later tests (1.3, 1.4, 2, and 3, 4) were significant, showing that the small amount of learning (0.5 dB) occurred predominantly in the first test. None of the differences between the repeated BB-triplet tests were significant. Note that the BB-triplet test was repeated at the end of the second session and, therefore, will have benefited from learning as a result of repeated, interleaved testing using the HF-triplet tests, which use the same test tokens and test procedure. For the HI group (Fig. 6), HF-triplet performance showed a gradual, small but significant (at test 1.3 and 2) improvement across tests. Analysis of the HF-CVC data also showed, for both groups,

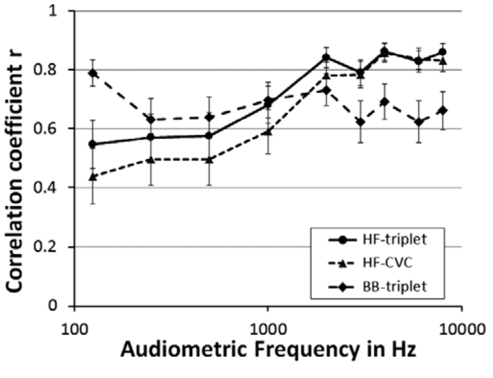

Fig. 5. Correlation coefficients (Pearson's '*r*') of the HF-triplet, HF-CVC, and BB-triplet SRT test results with hearing level (best ear) at audiometric frequencies. Error bars here and in Figures 6 and 7 are standard errors.

TABLE 5. Correlation coefficients (Pearson *r*) of the new HF tests with the BB-triplet and ASL tests (HI group)

HI	HF-CVC	BB-Triplet	ASL
HF-triplet	0.79	0.67	0.67
HF-CVC	-	0.65	0.68
BB-triplet	-	-	0.64

HI, hearing impaired; HF-triplet, high frequency digit triplet; HF-CVC, high frequency consonant–vowel–consonant; BB-triplet, broadband triplet.

a gradual but larger (1–1.5 dB) and significant improvement for all repeats of the test (1.2–4) compared to the first test (1.1). In summary, the main training effect for both groups was about −0.6 dB for the HF-triplet and about −1.2 dB for the HF-CVC test, obtained after the first or second retest. In subsequent testing, a small trend of continuing improvement could be seen in the HI group for both HF tests and for the NH group for HF-CVC test, but this trend was smaller than the standard error and not significant.

The test–retest accuracy of the tests was also examined using the mean intraindividual SD (intra-SD) of the repeated tests (Table 6). The intra-SD was calculated by computing the SD for each subject across retests and take the root mean square across participants. For the HF-triplet test, the training effect during the first two sessions of about −0.6 dB (Fig. 6) was less than half of the intra-SD. For the HF-CVC test, the training effect (−1.2 dB) was about the same size as the intra-SD. These SRTs (Table 6, left) cannot be directly compared with each other because of the different types of noise used. However, by reference to Table 4, the speech test data for the HI listeners were converted to hearing loss (Table 6, right), thus enabling comparisons between all three tests. Intra-SDs (in dB HL) for the HF-triplet and HF-CVC tests were quite similar to each other, but the Intra-SD (in dB HL) for the BB-triplet test is about twice as large.

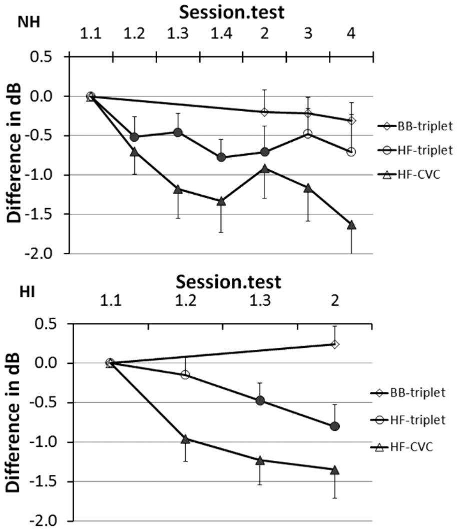

Fig. 6. Retest speech reception thresholds, relative to Test 1.1, for the triplet and consonant–vowel–consonant tests across subsequent sessions for normal-hearing and hearing-impaired listeners (panels NH and HI). Filled symbols indicate significant differences from Test 1.1. A more negative difference represents a better threshold.

TABLE 6. Intraindividual standard deviations (Intra-SD) of the repeated tests (Fig. 6) for NH and HI groups expressed in SRT dB

	SRT dB		HL dB
Intra-SD	NH	HI	HI
HF-triplet	1.3	1.3	5.6
HF-CVC	1.3	1.6	6.9
BB-triplet	0.7	1.0	12.3

For the HI group also expressed in PTA_{HF} hearing loss (dB HL).
Intra-SD, intraindividual standard deviation; SRT, speech reception threshold; NH, normal hearing; HI, hearing impaired; HF-triplet, high frequency digit triplet; HF-CVC, high frequency consonant–vowel–consonant; BB-triplet, broadband triplet.

Audio Presentation Mode

The HF-triplet and HF-CVC tests are designed for use through the Internet at home and in other environments in which the acoustic conditions of presentation cannot be controlled. Major contributors to the loss of control are likely to be through the use of different sound delivery systems (i.e., audio transducers) and different testing environments. The effect of different audio transducers, used in the lab or at home, is shown in Figure 7 relative to the first test in session 2 that used good quality headphones (Good HP). For the NH group (*n* = 24), 13 participants performed the HF triplet tests and 15 the HF-CVC test at home using loudspeakers (Home Spkr); another 3 used headphones. For the HI group, 30 performed the HF-triplet tests and 31 the HF-CVC test at home using loudspeakers; 5 used headphones. The results of home headphones are not shown in the figure as the number of participants was too low.

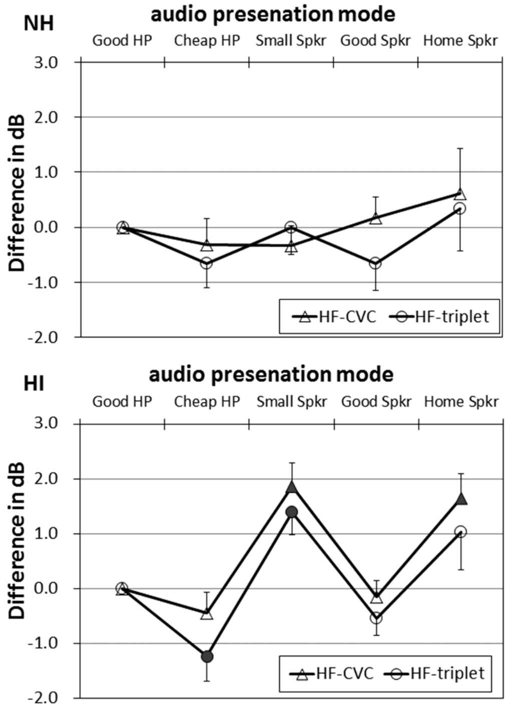

Fig. 7. Effect of audio presentation mode for normal-hearing and hearing-impaired listeners relative to the good-HP mode (panels NH and HI). A more negative difference represents a better threshold. The effect for testing through Internet at home using loudspeakers refers to a subgroup of 15 participants. Filled symbols show significant differences from the Good-HP mode. Further details and abbreviations in text.

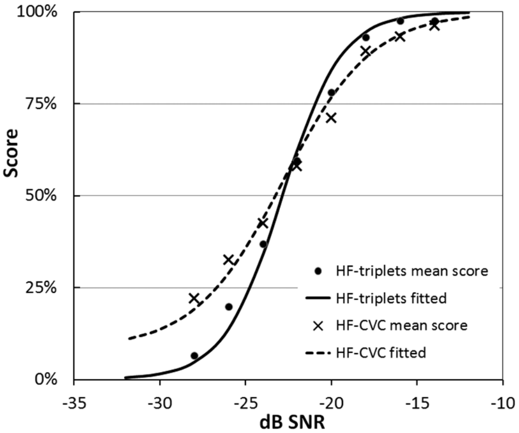

Fig. 8. Mean scores of the triplets and consonant–vowel–consonants as a function of SNR, with their mean psychometric functions based on the fitted slopes and speech reception thresholds.

For the NH group, the differences between using good headphones in the lab and the other transducers/environments were all nonsignificant (t-test $p < 0.05$). For the HI group, several of the results using small loudspeakers (Small Spkr), cheap headphones (Cheap HP), or home testing did produce significantly (t-test, $p < 0.05$) different thresholds, both higher and lower, on both the HF-triplet and the HF-CVC tests.

Psychometric Functions of the New Tests

Psychometric functions for the HF-triplet and HF-CVC material were fitted according to formula (1) to the session 5 test data for each participant ($n = 17$) to give an individual slope and SRT. These were averaged, and the mean data with fitted function are shown in Figure 8. SRT and slopes are given in Table 7, including the correction for the 1/12 guess chance of the CVC test. Table 7 also shows the corresponding modeled values from the homogenization tests and from the adaptive HF tests. The SRT and slope values of the BB-triplet test (Phipps 2007) have been added for comparison. It is expected that the modeled slopes and SRTs of the homogenization testing would correspond to those measured from the psychometric tests, despite the data being from two different groups of NH participants. The measured SRTs of the HF-triplets and HF-CVCs matched well with each other. The measured slope of the HF-triplets was a little better than expected from the model. For the HF-CVC, the measured slope was markedly flatter than the modeled slope.

DISCUSSION

The new HF-triplet and HF-CVC speech intelligibility tests described here were developed to detect hearing impairment above 2.0 kHz. SRTs on the new tests for the NH group were about 11 dB lower than those measured with the BB-triplet test. This difference is attributable to the LP filtering of the noise, unmasking the speech by 15 dB at frequencies above 1.5 kHz. Regression lines between SRT and high-frequency PTA_{HF} for the new tests had gradients that were, for the HI group, almost three times steeper than those of the BB-triplet test. For lower-frequency PTA_{LF}, the regression slopes of the new tests were about 1.5 (HF-CVC) to 2 (HF-triplet) times those of the BB-triplet test. High correlation coefficients were found between SRTs and PTA_{HF} and medium correlations were found with PTA_{LF}, demonstrating the increased sensitivity of the new HF tests for high-frequency (3–8 kHz) hearing loss. Expressed as test–retest repeatability, the HF-triplets and HF-CVC tests had a higher consistency and reliability compared to the original BB-triplet test. The intra-SD (in dB HL) of the BB-triplet test for the HI group was about 1.9 larger than that of the HF-triplet and about 1.7 larger than that of the HF-CVC test (Table 6).

Relation to Sentence Tests

The relation between the new HF tests and a clinical sentence-in-noise test (the ASL) was found to be moderate to high ($r = 0.67$), about the same as the relation between the BB-triplet test and the ASL. The original triplet test for telephone (in Dutch) was previously found among HI listeners to correlate highly ($r = 0.87$) with SRT measured by the Plomp and Mimpen (1979) sentence test (Smits & Houtgast 2006). Recently, Smits et al. (2013) have reported an even higher correlation ($r = 0.96$) for NH listeners with simulated hearing loss. The differences between these estimates appear to depend largely on differences in age, hearing impairment, and language skills between participants, rather than on any intrinsic properties of the tests. Smits et al. (2013) conclude that the BB-triplet test can be used across virtually the full range of hearing abilities. We suggest that the HF-tests described here retain these properties and provide greater sensitivity for the detection of high frequency HI.

Training Effects

One of the benefits of using a speech intelligibility test based on a closed set of small and simple speech materials is that the effects of repeat testing (a form of training) are expected to be small. A smaller training effect allows for more accurate and

TABLE 7. Mean (and SD) SRTs and slopes from the psychometric functions as fitted from the scores as a function of the signal to noise ratios

	Measured				Modeled		Adaptive	
Test	SRT (dB)	SD SRT (dB)	Slope (%/dB)	SD slope (%/dB)	SRT (dB)	Slope (%/dB)	SRT (dB)	SD SRT (dB)
High-frequency-triplet	–22.9	1.8	14.7%	1.1%	–23.8	13.0%	–21.3	2.4
High-frequency-consonant–vowel–consonant	–23.2	1.3	8.8%	2.1%	–23.1	12.1%	–21.1	2.1
Broadband-triplet	–11.1	n.a.	17.0%	n.a.	n.a.	n.a.	–10.3	1.1

The broadband-triplet data are from Phipps 2007. The modeled SRTs and slopes are from the homogenization test results. The last column presents the mean (and SD) SRTs as measured by the adaptive tests.
SRT, speech reception threshold.

efficient assessment of hearing loss development or hearing rehabilitation over time. Speech tests based on sentences (e.g., ASL, Plomp-sentence-test) cannot be used repeatedly as the sentence or some of the key words can be remembered when the sentence is reused. This can be overcome by having a large corpus of sentences that have been "homogenized" so that different sentence lists will give the same results. Nevertheless, these tests still have a limited set of sentences that may "run-out" after some time. As an alternative, the Matrix or BKB sentence tests (Hagerman 1982; Jansen et al. 2012) can be used. These tests have a closed set of key words that are used to randomly generate the sentences. Jansen et al. (2012) observed a relatively large training effect during the first two tests, but later repetitions produced more stable results.

For the new HF-triplet test, the training effect was relatively small and was largest in the first retest for the NH group. For the HI group, the largest retest effect was found after the second retest. This training effect (mean of −0.6 dB) was only about half of the intra-SD during the initial sessions. For the NH group, no significant training effect was observed during later repeats, but no longer-term observations were obtained. Note that additional testing of the NH group occurred (using alternative transducers) during session 2, after the "training" data were obtained, and (at home) between sessions 2 and 3 (Table 1). It is likely that these additional tests would have contributed to training effects observed in sessions 3 and 4 (Fig. 6; NH).

Leensen et al. (2011) also found a significant learning effect in the first retest, but did not perform further retests. They found a mean test–retest effect of the same size (−0.7 dB; NH group) for the Dutch HF-CVC test, using LP noise similar to that used here. Smits et al. (2013) also found a small training effect (−1.0 dB) on the first BB-triplet retest, but performance was then stable over 24 subsequent measures. The larger training effect for the HF-CVC test found here during later sessions (−1.2 dB, close to the intra-SD) was presumably caused by the low familiarity of the 12 CVC words compared to the 10 digits. Recall of the CVC words would not be expected to play a large role, as the CVC response matrix was visible all the time. However, the recognition and reproduction of the CVCs may need additional time, relative to the digits, to be stored and processed in working memory. It may also be noted that pure-tone thresholds from conventional audiometry show a training improvement in the order of 5–10 dB across three test sessions (Zwislocki et al. 1958).

Psychometric Functions

It was expected that the psychometric function mean SRTs (Table 7) would be similar to those obtained by the adaptive tests, as the same materials were used and only the procedure for obtaining the SRT was different. In fact, the SRT measured from the psychometric function testing was lower (better) than the adaptive tests by 1.6 dB for the HF-triplets and 2.1 dB for the HF-CVCs. However, the psychometric SRTs were not corrected for the training effect that would have reduced the differences by 0.6 and 1.2 dB, respectively. When restricting the results of the adaptive test to the same 17 NH participants for whom psychometric data were available, a negligible difference of 0.1 dB was found for both tests.

When comparing the measured with the modeled psychometric function of the homogenization tests, the modeled SRT for the HF-triplets was about 0.9 dB lower (better). This difference is not significant and may in part be attributable to the two different groups of participants. Differences in training were expected to be negligible as in both procedures large numbers of triplets were experienced. For the HF-CVC, the measured and modeled SRTs compared very well. The measured slope of 8.8% was worse than the modeled 12.1%, with a difference of about 2 SD. For the HF-triplet test, the slope seems to correspond a little better with difference <1.5 SD.

During the homogenization testing, we had no good indication what the best test presentation level would be. It was assumed that the commonly used noise level of 65 dBA would be comfortable and applicable to speech understanding. The new tests used a level at which the speech noise was most comfortable for the listener. That seemed a practical approach when level prescription was impossible in home situations and difficult for HI participants. However, the actual levels set in session 4 were around 80 dBA. This level, for the HF-triplet LP-noise (NH participants), was some 9.5 dB higher than that chosen for the triplet broadband noise (BB-noise), possibly because the energy of the filtered noise is concentrated in the low frequencies.

The results of the psychometric function test may be compared to those of Leensen et al. (2011). In their tests, using Dutch CVC words with similar LP noise as used here, an SRT of −20.8 dB and a slope of 11.2% were found. This compares reasonably well with the HF-CVC SRT and slope found here. For the completely novel HF-triplet test, no direct comparison with other tests was possible. However, the slope of the mean HF-triplet function (14.7%) was comparable to that of the English BB-triplet test (17.0%, Phipps 2007) and the telephone Dutch triplet test (16.0%, Smits et al. 2004). All these triplet tests had significantly steeper slopes than the HF-CVC tests.

Sensitivity and Specificity of the Tests

The newly developed tests were designed to have a high sensitivity to discriminate between people with and without high-frequency hearing loss. For high sensitivity, the percentage of true positives (people correctly classified as having a hearing loss) relative to all real positives (all true positives plus all false negatives) should be maximized. Also, the percentage of false positives (people wrongly classified as having hearing loss) relative to all real negatives (all false positives plus all true negatives) should be minimized. The specificity is 100% minus the percentage of false positives. The proportion of true and false positives is dependent on the criterion for distinguishing between normal and impaired hearing. $PTA_{HF} < 20$ dB corresponds to no or very mild HF hearing loss, between 20 and 60 dB to mild/moderate HF loss and >60 dB to severe/profound HF loss. We compared the criteria of $PTA_{HF} > 20$ dB and $PTA_{HF} > 60$ dB. Receiver operating characteristics (ROCs) were calculated (Fawcett 2004; Eng 2006) for each test (Fig. 9), based on results of the first validation test in session 1.

The specific criterion for hearing loss is a "cutoff" value on the ROC curve that is a compromise between sensitivity and specificity. We took a best compromise by taking the point where a further increase of the true positive rate is equal to an increase of the false-positive rate, as indicated in Figure 9. From these ROC curves, the corresponding SRT cutoff values were derived (Table 8).

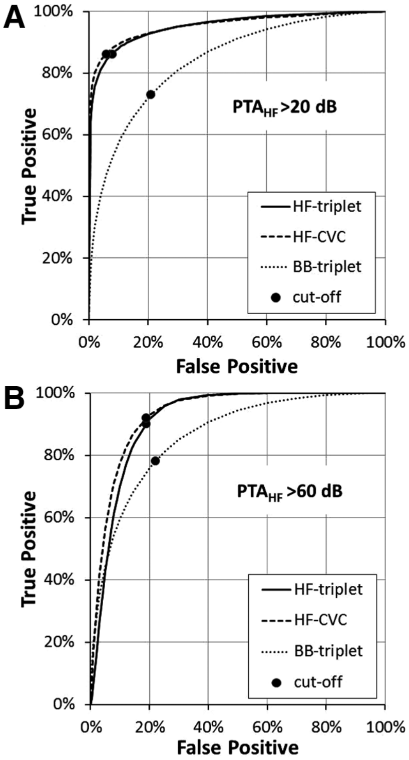

Fig. 9. ROC curves showing the fraction of true positives (sensitivity) and false positives (1-specificity) for the HF-triplet, HF-CVC, and BB-triplet tests based on PTA_{HF} (3, 4, 6, and 8 kHz) hearing loss of 20 dB (panel A) and 60 dB (panel B). The filled dots correspond to the chosen cutoff SRT of each test (Table 8).

The HF-triplet and HF-CVC tests had nearly the same ratios for false and true positives, for both hearing loss scenarios, and thus had similar sensitivity and specificity. The HF-CVC SRT (the ROC optimal "sweetspot") was 0.4 dB (cutoff 1) and 2.6 dB (cutoff 2) lower than the HF-triplet test sweetspot. This difference is partly explained by the difference between the HI group regression curves (Fig. 4), which was about 0.7 dB. For both HF tests, the sensitivity was considerably greater than that of the BB-triplet test. Nevertheless, a recent study by Jansen et al. (2013) has also shown very high sensitivity of a BB-triplet test for detecting HF hearing loss in noise-exposed workers, possibly because of the relative homogeneity of that group of HI participants. For the BB-triplet test in our study, the SRT cutoffs were quite similar (−9.2 and −8.2 dB SRT) and within the SD for the NH group. This is unsurprising as the BB-triplets had a relative poor correlation with PTA_{HF}. Additional analysis was conducted at the low-frequency loss PTA_{LF} at <20 dB and <45 dB HL. The corresponding ROC areas were 0.91 and 0.90 with cutoff 1 = −7.2 dB SRT and cutoff 2 = −9.1 dB SRT. These cutoffs compare closely to those reported by Phipps (2007; −6.9 and −9.0 dB SRT).

The PTA_{HF} > 20 dB criterion was chosen to detect early signs of hearing loss that may be due to NIHL or presbycusis. Such a criterion could warn youngsters of potential NIHL due to loud noise or music exposure. For older people, it may warn of the first signs of presbycusis, which may motivate them to consult a healthcare professional and, perhaps, initiate regular check-ups for hearing problems. The PTA_{HF} > 60 dB criterion would indicate a substantial high-frequency hearing loss. This criterion could help to identify those who would potentially benefit from a hearing aid. This is especially important given that the insidious, creeping nature of presbycusis means hearing problems often go unnoticed for some time. The actual need for a hearing aid may, of course, further depend on any low-frequency hearing loss.

Effect of Presentation Mode

The new HF tests have been developed for Internet use with simple PC equipment and without the need of supervision. The tests were introduced with written or spoken instructions, and the procedure and sounds were demonstrated by the Internet application. Ideally the tests should be done using headphones of good quality. However, it is impractical to prescribe a preferred presentation mode, as most users at home will not be willing to buy new equipment or to connect infrequently used equipment they already have. In many cases, the small loudspeakers built into a PC monitor or laptop will be used. Cheap headphones, when used by the HI group, seemed to have a small advantage compared to the lab-quality headphones. This advantage may be attributed to an emphasized sound level in the 3 to 10 kHz region, which benefited the HI group more than the NH group.

Small monitor speakers showed a significant disadvantage in the lab of about 1.5 dB in the HI group for both HF tests. This was consistent with the results of the participants at home, where speakers of low quality were likely to have been used. This disadvantage, which occurred only for the HI group, may indicate an extra handicap when using cheap loudspeakers. The frequency response curves of the low-quality speakers suggested a higher response at the mid-frequencies (700–1300 Hz) compared to the good headphones and the good loudspeakers. At these mid-frequencies, an upward spread of masking (Oxenham & Plack 1998) may affect hearing at higher frequencies, thus increasing the measured SRT, suggesting worse hearing. A similar disadvantage of using cheap speakers, compared to

TABLE 8. ROC area and true positive, false-positive percentages of tests at cutoff SRT test values obtained from ROC curves at PTA_{HF} >20 dB and >60 dB

	PTA_{HF} > 20 dB				PTA_{HF} > 60 dB			
		Cutoff 1				Cutoff 2		
Test	ROC area	True Pos	False Pos	SRT [dB]	ROC area	True Pos	False Pos	SRT [dB]
High-frequency-triplet	0.95	87%	8%	–17.1	0.92	90%	19%	–9.8
High-frequency-consonant–vowel–consonant	0.96	87%	6%	–17.5	0.93	92%	19%	–12.4
Broadband-triplet	0.84	73%	21%	–9.2	0.86	78%	22%	–8.2

SRT, speech reception threshold; ROC, receiver operating characteristic; PTA, pure-tone average.

cheap headphones, was also found by Smits et al. (2006). They evaluated the public Internet version of the standard telephone (low frequency) Dutch DTT for more than 30,000 participants with all levels of hearing loss. The measured SRT was about 1.1 dB higher (poorer hearing) for participants who used speakers rather than (cheap) headphones.

Overall, for both NH and HI groups, it may be concluded that the presentation mode has only a small influence and that an overall systematic correction for any of the presentation modes is not possible for these general modes. However, it may be concluded that the use of small loudspeakers, home speakers or cheap headphones does have some effect on the results for the HI group. This can be avoided by using good quality headphones for professionally administered tests. For testing at home, headphones (either cheap or quality) should be recommended as they will generally produce less variable results, will reduce the effects of environmental noises and reverberation (if of the "insert" or "closed circumaural" type), and will enable separate testing of each ear. Those with moderate–severe hearing loss should avoid testing using cheap speakers, especially those built into some laptop computers.

Noise-Induced Hearing Loss

The new HF tests were more sensitive and correlated better with high-frequency hearing loss than did the BB-triplet test. NIHL, as with presbycusis, is characterized by a high-frequency loss. While noise damage typically causes a more 'notched' loss, often affecting frequencies between 3 and 6 kHz more than higher frequencies, presbycusis tends to start at higher frequencies and gradually affects lower frequencies. The new tests are therefore likely to detect both types of hearing loss much earlier than the BB-triplet test; by the time lower frequencies are affected, more damage is likely to have taken place. To augment this "early warning" of hearing loss, a validated criterion measure (e.g., "cutoff 1" as described here) could be used to increase sensitive detection of a mild loss as it develops. Early detection of a high-frequency loss is of particular importance, as it could influence the use of (any or better) hearing protection or cessation/avoidance of the damaging sound exposure.

Using a similar low-pass background noise masker, Leensen et al. (2011) compared the sensitivity and specificity of their test to distinguish between participants with a "narrow" high-frequency loss (likely to be NIHL) and a "broad" high-frequency loss (likely to be more severe NIHL). They defined a SRT cutoff that gave the test a sensitivity of around 91% and a specificity of around 85% for distinguishing between narrow and broad hearing loss. The test could thus be effective for distinguishing between the two conditions, making it even more valuable as a screening tool. In their study, however, the mean thresholds of the "broad" and "narrow" high frequency hearing loss groups do not appear to differ significantly.

It could be argued, finally, that the nature of the background noise and target speech makes the test more effective at detecting NIHL than presbycusis. The low-pass filtering of the noise at 1.5 kHz effectively defines the lower limit of increased test sensitivity, while the use of speech as a target means there is little informational content at frequencies above 6 kHz. It could therefore be theorized that optimal test sensitivity would be found in the 2–6 kHz region, where the low-pass filtering has an effect and where significant speech information lies. This is the region we would expect to reveal NIHL rather than presbycusis in its early phase. For presbycusis, however, the same region will be affected in a later phase, so making any differentiation impossible.

Application of the HF Hearing Tests

The standard, BB-triplet tests (Smits et al. 2004; Phipps 2007; Vlaming et al. 2011) have been developed for general hearing screening. They use speech with a noise masker of the same long-term spectrum to simulate the multi-talker babble noise of speech maskers in everyday life. The BB-triplet test correlated well with the ASL sentence-based speech-in-noise test, which is even more representative of everyday listening situations. The newly developed tests will detect high-frequency loss that may not be immediately noticed, as hearing may still be good at low frequencies while high-frequency mediated speech intelligibility decreases. Correlations between the ASL and the HF- and BB-triplet tests were similar. As predicted, the HF tests as well as the BB-triplet test correlated with PTA_{LF} and so are equally sensitive for the detection of low frequency loss. The new HF tests should therefore also be good predictors of general hearing problems. The main advantage of the HF tests is that they have much better accuracy and sensitivity, making them preferable to the BB test for large group screening programs. For individuals, the HF tests will be sensitive enough to provide advice on the use of hearing aids.

The new HF tests can also be used to evaluate and discriminate between different hearing rehabilitation strategies. Care should be taken to relate the results to the intra-SD which is, for the HF-triplet tests, a factor of 2.2 lower (better) than that of the BB-triplet test when converted to predicted audiogram hearing level at high frequencies. The training effect for repeated testing can either be corrected for, or an extra training trial can be given at the beginning. The new tests are expected to improve rehabilitation testing of hearing aids that give most amplification to the higher frequencies. Amplification between 1500 and 6000 Hz is the most important region for NIHL and presbycusis and for the use of open fittings where there is hardly any amplification below 1000 Hz. The HF tests correlate best with high-frequency hearing loss. As most hearing aids are fitted to such losses, the HF tests are expected to assess their rehabilitative benefit with greater sensitivity than the BB-triplet or speech sentence tests. Further research is required, however, to address this specific prediction.

ACKNOWLEDGMENTS

Thanks are expressed to Sandra N. Smith for collecting data for the project.

David R. Moore is currently at Communication Sciences Research Center, Cincinnati Children's Hospital, Cincinnati, Ohio, USA.

This work was supported by the National Institute for Health Research (M.S.M.G.V., R.C.M.), Action on Hearing Loss (M.J.), and the Medical Research Council (D.R.M.). This work has partly been funded by Action on Hearing Loss (www.actiononhearingloss.org.uk).

The authors declare no other conflicts of interest.

Address for correspondence: Marcel S. M. G. Vlaming, Nottingham Hearing Biomedical Research Unit, 113 The Ropewalk, Nottingham, NG1 5DU, UK. E-mail: vlamingm@yahoo.co.uk

Received June 2, 2013; accepted May 11, 2014.

REFERENCES

Brand, T., & Kollmeier, B. (2002). Efficient adaptive procedures for threshold and concurrent slope estimates for psychophysics and speech intelligibility tests. *J Acoust Soc Am*, *111*, 2801–2810.

Davis, A., Smith, P., Ferguson, M., et al. (2007). Acceptability, benefit and costs of early screening for hearing disability: a study of potential screening tests and models. *Health Technol Assess*, *11*, 1–294.

Dawes, P., Fortnum, H., Moore, D. R., et al. (2014). Hearing in middle age: A population snapshot of 40–69 year olds in the UK. *Ear Hear*, *35*, e44–51.

Dubno, J. R., Eckert, M. A., Lee, F. S., et al. (2013). Classifying human audiometric phenotypes of age-related hearing loss from animal models. *J Assoc Res Otolaryngol*, *14*, 687–701.

Eng, J. (2006). ROC analysis: Web-based calculator for ROC curves. Baltimore: Johns Hopkins University [updated 2006 May 17th, cited May 4th 2012]. www.jrocfit.org.

Fawcett, T. (2004). ROC Graphs. Notes and Practical Considerations for Researchers. ReCALL 31(HPL-2003–4), 1–38.

Ferguson, M. A., Henshaw, H., Clark, D. P. A., et al. (2014). Benefits of phoneme discrimination training in a randomised controlled trial of 50–74 year olds with mild hearing loss. *Ear Hear*, 35, e110–21.

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

Jansen, S., Luts, H., Wagener, K. C., et al. (2010). The French digit triplet test: a hearing screening tool for speech intelligibility in noise. *Int J Audiol*, *49*, 378–387.

Jansen, S., Luts, H., Wagener, K. C., et al. (2012). Comparison of three types of French speech-in-noise tests: a multi-center study. *Int J Audiol*, *51*, 164–173.

Jansen, S., Luts, H., Dejonckere, P., et al. (2013). Efficient hearing screening in noise-exposed listeners using the digit triplet test. *Ear Hear*, *34*, 773–778.

Kent, R. D., & Read, C. (2002). *The Acoustic Analysis of Speech*. Albany: Delmar, Thomson Learning.

Lanternfish ESL website. (2010). First CVC words [cited Nov 15th, 2010], www.bogglesworldesl.com.

Leensen, M. C., de Laat, J. A., Snik, A. F., et al. (2011). Speech-in-noise screening tests by internet. Part 2: Improving test sensitivity for noise-induced hearing loss. *Int J Audiol*, *50*, 835–848.

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audio-visual speech-reception thresholds for sentences in noise: rationale, evaluation, and recommendations for use. *Br J Audiol*, *24*, 29–43.

Morata, T. C. (2007). Young people: Their noise and music exposures and the risk of hearing loss. *Int J Audiol*, *46*, 111–112.

Oxenham, A. J., & Plack, C. J. (1998). Suppression and the upward spread of masking. *J Acoust Soc Am*, *104*, 3500–3510.

Ozimek, E., Kutzner, D., Sęk, A., et al. (2009). Development and evaluation of Polish digit triplet test for auditory screening. *Speech Commun*, *51*, 307–316.

Phipps, H. L. (2007). *Assessment of Telephone Bandwidth on the English Number Recognition in Noise Test*. Institute of Sound and Vibration Research. MSc thesis, University of Southampton.

Plomp, R., & Mimpen, A. M. (1979). Improving the reliability of testing the speech reception threshold for sentences. *Audiology*, *18*, 43–52.

Plomp, R. (1986). A signal-to-noise ratio model for the speech-reception threshold of the hearing impaired. *J Speech Hear Res*, *29*, 146–154.

Robinson, D. W., & Sutton, G. J. (1979). Age effect in hearing: A comparative analysis of published threshold data. *Audiology*, *18*, 320–334.

Roeser, R. J. (1996). *Roeser's Audiology Desk Reference*. NY: Thieme.

Smits, C., Kapteyn, T. S., Houtgast, T. (2004). Development and validation of an automatic speech-in-noise screening test by telephone. *Int J Audiol*, *43*, 15–28.

Smits, C., & Houtgast, T. (2005). Results from the Dutch speech-in-noise screening test by telephone. *Ear Hear*, *26*, 89–95.

Smits, C., & Houtgast, T. (2006). Measurements and calculations on the simple up-down adaptive procedure for speech-in-noise tests. *J Acoust Soc Am*, *120*, 1608–1621.

Smits, C., Merkus, P., Houtgast, T. (2006). How we do it: The Dutch functional hearing-screening tests by telephone and internet. *Clin Otolaryngol*, *31*, 436–440.

Smits, C., Theo Goverts, S., Festen, J. M. (2013). The digits-in-noise test: Assessing auditory speech recognition abilities in noise. *J Acoust Soc Am*, *133*, 1693–1706.

Trumble, S. C., & Piterman, L. (1992). Hearing loss in the elderly. A survey in general practice. *Med J Aust*, *157*, 400–404.

Vlaming, M. S. M. G., Kollmeier, B., Dreschler, W. A., et al. (2011). HearCom: Hearing in the Communication Society. *Acta Acust United Ac*, *97*, 175–192.

Wagener, K., Josvassen, J. L., Ardenkjaer, R. (2003). Design, optimization and evaluation of a Danish sentence test in noise. *Int J Audiol*, *42*, 10–17.

Wagener, K. C., & Brand, T. (2005). Sentence intelligibility in noise for listeners with normal hearing and hearing impairment: Influence of measurement procedure and masking parameters. *Int J Audiol*, *443*, 144–156.

Watson, C. S., Kidd, G. R., Miller, J. D., et al. (2012). Telephone screening tests for functionally impaired hearing: Current use in seven countries and development of a US version. *J Am Acad Audiol*, *23*, 757–767.

Zokoll, M. A., Wagener, K. C., Brand, T., et al. (2012). Internationally comparable screening tests for listening in noise in several European languages: The German digit triplet test as an optimization prototype. *Int J Audiol*, *51*, 697–707.

Zwislocki, J. J., Maire, J., Feldman, A. S., et al. (1958) On the effect of practice and motivation on the threshold of audibility. *J Acoust Soc Am*, *30*, 254–262.