Closed Set Speech Discrimination Tests for Assessing Young Children

Deborah A. Vickers[1], Brian C.J. Moore[2], Arooj Majeed[3], Natalie Stephenson[3], Hala Alferaih[3], Thomas Baer[2], and Josephine E. Marriage[4]

[1]University College London, Department of Speech Hearing and Phonetic Sciences, 2 Wakefield Street, London. WC1N 1PF, UK

[2]Dept of Experimental Psychology, University of Cambridge, Downing St, Cambridge CB2 3EB, UK

[3]University College London, Ear Institute, 332 Grays Inn Road, London. WC1X 8EE, UK

[4,2]Chear, 30 Fowlmere Road, Shepreth, Royston, Herts SG8 6QS, UK

Send correspondence to:

Deborah A. Vickers
email: d.vickers@ucl.ac.uk

Objective: To obtain data assessing normative scores, test-retest reliability, critical differences and the effect of age for two closed-set consonant discrimination tests.

Design: The two tests are intended for use with children aged 2-8 years. The tests were evaluated using normal-hearing children within the appropriate age range. The tests were: (1) The closed-set consonant confusion test (CCT) and (2) The consonant-discrimination sub-test of the closed-set Chear Auditory Perception Test (CAPT). Both were word-identification tests using stimuli presented at a low fixed level, chosen to avoid ceiling effects while avoiding the use of background noise. Each test was administered twice.

Results: All children in the age range 3 years 2 months to 8 years 11 months gave meaningful scores, and were able to respond reliably using a computer mouse or a touch screen to select one of four response options displayed on a screen for each trial. Assessment of test-retest reliability showed strong agreement between the two test runs (inter-class correlation $\geq$ 0.8 for both tests). The critical differences were similar to those for other monosyllabic speech tests. Tables of these differences for the CCT and CAPT are provided for clinical use of the measures. Performance tended to improve with increasing age, especially for the CCT. Regression equations relating mean performance to age are given.

Conclusions: The CCT is appropriate for children with developmental age in the range 2 to 4.5 years and the CAPT is appropriate as a follow on test from the CCT. If a child scores 80% or more on the CCT they can be further tested using the CAPT, which contains more advanced vocabulary and more difficult contrasts. This allows the assessment of consonant perception ability and of changes over time or following an intervention.

INTRODUCTION

Good auditory discrimination is particularly important for the development of speech and language skills in the early years of life when brain plasticity is greatest (Sharma et al. 2005; Kuhl et al. 2014). Therefore, there is great value in having appropriate speech tests with known reliability for use with younger age groups to assess interventions such as the provision of hearing aids or a program of training. Such tests need to be sensitive to the perceptual changes that are likely to arise from the intervention (Kirk 2012). Discrimination or recognition speech perception tests for infants should be designed to be age appropriate and to avoid floor and ceiling effects (Govaerts et al. 2006). There is a trade off between what is feasible for younger children and the sensitivity and reliability of the measures derived. Generally, only relatively imprecise measures can be obtained from young children.

There are several problems in conducting speech testing for hearing-impaired children aged less than 6 years. The use of open-set speech tests with verbal responses requires clear articulation by the child to allow scoring, especially when phoneme scores are required (Stiles et al. 2012). If written responses are obtained, then basic phonological reading and writing skills are required (Scollie 2008). The requirement for clear articulation or written responses limits the minimum age at which valid and repeatable testing is possible using open-set tests. Even when closed-set tests are used, there are difficulties in testing young children, who typically have a short attention span and limited skills in speech understanding and language use. It is therefore understandable that few clinical measures are available for assessing speech perception in very young children. However, it is crucial that suitable tests are available to assess whether clinical interventions are effective.

Tomblin et al. (2015) recognised the importance of a flexible strategy for acquiring speech perception data across the developmental age range. They tested children with hearing aids starting at six months of age up to nine years. For the children under four years of age they were restricted to parental questionnaires and live voice tests, due to the difficulties of testing young children. For older children, monosyllabic word tests were used, while for children above seven years of age it was possible to assess the perception of speech in noise. Such an adaptable strategy is important to ensure that children of different ages can be appropriately assessed, but makes it difficult to assess long-term developmental trends for individual children or to compare results for different age groups.

Speech tests may be used both for assessing trends over time for individual children or groups of children and for comparing different groups of children in research studies. Regardless of the purpose, it is useful to know the inherent variability of the outcome measure. This can be important in assessing whether a given child is showing improved or poorer performance over time, and when choosing group sizes in research studies. To be appropriate for assessing changes in the effectiveness of hearing aid provision, or of changes in the frequency-gain characteristic of a hearing aid, a speech test should assess the use of acoustic cues across a wide frequency range. Ideally the test should be reliable, have little redundancy, be easy for young children to complete and not be reliant on speech production. Another important aspect of speech tests for young children is the availability of normative data. Such data are important for allowing comparisons of speech scores for groups and individuals with scores that would be typical for their age.

In what follows, we briefly review existing tests that can be used to assess speech perception for children aged six years or less and we assess their merits and limitations. Then we give the rationale for the development and evaluation of the speech tests that are presented in this paper. These tests are intended to be applicable to the evaluation of children aged between two and eight years.

Parental-response questionnaires are typically used with children under four years of age. The subjectivity of these can make them insensitive to small changes. However, a validated questionnaire can be useful for monitoring relatively large changes in auditory perception over time. The Infant Toddler Meaningful Auditory Integration Scale (IT-MAIS, McConkey Robbins et al. 2004) is a validated measure, with known normative ranges, that has been shown to be sensitive to changes in perception with age. The Categories of Auditory Performance test (CAP; Archbold et al. 1995) uses a hierarchical rating scale with eight levels of auditory perception from “no awareness of environmental sounds” to “uses the telephone”. Although this appears to be a fairly gross measure, it has been shown to be sensitive to differences in performance over time, as exemplified over the first 12 months of hearing experience for children receiving cochlear implants at an early age (Zhou et al. 2013). However, for both the IT-MAIS and the CAP, the variability of the outcomes, the limited number of discrete scores, and the subjective nature of the responses, prevent these measures from being viable in assessing the impact of small changes in sound delivery, for example, changes in the frequency-gain characteristic of a hearing aid. They do, however, have a role in detecting gross changes in perception.

Tests of word recognition are typically used with children aged four years or older. A useful measure of the reliability of such tests when comparing performance on two conditions, for example listening with and without hearing aids, is the critical difference. This is the smallest difference between scores obtained from an individual required to be 95% confident of a “true” difference across conditions, for example to be 95% confident that the use of hearing aids is beneficial. The critical difference is a conservative measure and the values are often large relative to the differences across conditions that are likely to occur. Unfortunately, critical differences are rarely provided for the speech tests that are used with children.

Thornton and Raffin (1978) calculated theoretical critical differences for the CID W-22 word test and compared them with obtained critical differences. The CID W-22 test is an open-set monosyllabic speech perception test. They showed that both the theoretical and obtained critical difference values were greatly affected by the number of items used to evaluate each condition. This leads to a dilemma when using speech tests with children: for example, presenting ten words per condition would not provide a sufficiently reliable measure of any change in performance across conditions, but presentation of many more items to increase reliability could make the test too time consuming for clinical practice or could lead to loss of attention of the child. Probably because of the limited reliability of the speech scores obtained with young children, many studies on early intervention for hearing-impaired children do not report speech recognition scores for children younger than about six years (Davidson & Skinner 2006; Strauss & van Dijk 2008).

For British English there are very few validated measures of speech perception with high sensitivity and reliability that can be used with young children. The McCormick Toy

Test is the main speech perception test in the UK that is used with very young children (Cullington et al. 2013). It is an adaptive discrimination test using words presented in either speech-shaped noise or two-talker babble. Lovett et al. (2013) demonstrated that the McCormick Toy Test had a large critical difference when tested with young children. The average critical difference for the speech reception threshold in noise was 7.5 dB for one run. This makes it difficult to monitor performance on an individual basis, because the differences that might be expected over time or across conditions are usually smaller than the critical difference for the test; with multiple runs the performance estimates are more robust but there is always the possibility of fatigue and loss of attention with young children.

Other closed-set tests for young children, using American English, are the pediatric speech intelligibility (PSI) test (Jerger et al. 1980) and the online imitative test of speech pattern contrast perception (OLIMSPAC) (Boothroyd et al. 2005). The critical differences for these tests have not been reported, making the results difficult to interpret on a case-by-case basis. However, the tests have been demonstrated to be effective measures for group level data; see, for example, Sininger et al. (2010). Holt and Lalonde (2012) described a test assessing toddler speech sound discrimination, for two different contrasts, using a change/no change paradigm. They measured test-retest reliability for normal-hearing 2- and 3-year old children and found a strong correlation between scores for two successive runs ($r = 0.886$, $p = 0.037$). However, critical difference values were not presented.

This paper describes the design and evaluation of two tests that have potential for the functional speech assessment of young children. The tests have already been used in hearing aid and cochlear implant research. They have been shown to be sensitive to hearing aid gain settings (Marriage & Moore 2003) and have been used to derive cochlear implant candidacy criteria (Lovett et al. 2015). These tests are the consonant confusion test (CCT) and the Chear Auditory Perception Test (CAPT). Both tests use four response alternatives on each trial, based on the observation that children as young as two years old are able to make a choice among four alternatives. The pattern of phoneme confusions made by a child in the tests can give some frequency-specific information about the audibility and discrimination of speech cues. All the items in the tests are real words that should be familiar to children in the target age range.

In a companion paper (Marriage et al. 2017), we describe the use of the tests to compare speech scores for children using hearing aids fitted with the DSL i/o (Cornelisse et al. 1995), DSL V (Scollie et al. 2005), and NAL-NL1 (Byrne et al. 2001) procedures, and we show that the tests are capable of revealing differences between the procedures.

The goals of the present paper were: to determine the appropriate age ranges for the use of the CCT and the CAPT; to provide normative data for the tests; to evaluate changes in performance with age; to determine the test-retest reliability of each test; and to provide critical differences for each test.

METHOD

Ethical Approval

Ethical approval was obtained from the University College London ethics committee (4059/001).

Participants

Thirty one children aged between 38 and 107 months (mean age = 74 months; 19 females and 12 males) were assessed with the CCT and 55 children aged between 48 and 107 months (mean age = 81 months; 31 females and 24 males) were assessed with the CAPT. All children were screened with pure-tone audiometry at the beginning of the session to have hearing thresholds at 0.5, 1.0, 2.0 and 4.0 kHz that were less than 20 dB HL. The following demographic characteristics were collected: chronological age, whether English was their only language or one of two or more languages spoken, and results of the Renfrew word-finding vocabulary scale (Renfrew 1995). The latter provides a quick assessment of expressive vocabulary based on a picture-naming task, giving gender-appropriate age-equivalent scores. These scores are used as a proxy for English language development. To be included children were required to have sufficient speech and language skills to be able to understand and participate in the tests. This was evaluated by showing them cards of the pictures used in the tests to ensure that they understood what word was associated with each picture. Of the children tested with the CCT, 13 had English as their only language (E1L group) and 18 spoke more than one language (English as additional language, EAL, group). Of the children tested with the CAPT, 22 fell in the E1L group and 33 in the EAL group. All children were attending English speaking nurseries or schools and had intelligible spoken English.

Consonant Confusion Test (CCT)

The CCT is intended for use with children aged two years or older. On each trial, one of four monosyllabic words is presented. All words are intended to be familiar to children with a vocabulary age of two years or more. The response alternatives are represented by pictures. The requirement to use familiar vocabulary items that can be represented through pictures constrains the acoustic features that can be used and means that the response alternatives differ in more than one speech sound. Each word group has (phonemically) the same vowel, and different contrastive consonants are used in both word-initial and word-final positions, thus giving multiple cues for identification of each item. The CCT was developed from the Michael Reed picture test screening cards (Reed 1959). The test items for the CCT are available as a CD recording and responses are available in a printed booklet. The CCT is also incorporated into the ParrotPlus speech test system (www.soundbytesolutions.co.uk). The materials for the computer-based version of the test are also freely available by contacting the corresponding author.

For the present study, there were 40 words in total (10 word groups, each containing 4 words; see table 1 for the words in the test). The test was conducted twice within the same test session but with a break in between and a different order of presentation of items for each run.

TABLE 1. Word groups for the consonant confusion test (CCT). Note that the vowel sound remains approximately the same within each group but word-initial and word-final consonants can change

	Word 1	**Word 2**	**Word 3**	**Word 4**
Group 1	Cow	Owl	House	Mouse
Group 2	Bed	Hen	Peg	Egg
Group 3	Fan	Man	Cat	Hat
Group 4	Key	Three	Feet	Sheep
Group 5	Pig	Chick	Fish	Ship
Group 6	Horse	Ball	Fork	Door
Group 7	Shoe	Moon	Spoon	Food
Group 8	Pipe	Pie	Kite	Five
Group 9	Sock	Cot	Doll	Dog
Group 10	Jug	Duck	Bus	Cup

Chear Auditory Perception Test (CAPT)

The CAPT uses the same format as the CCT but is intended for slightly older children with a more advanced vocabulary, who are beginning to recognize written words. This allows monosyllabic words to be used that differ in only one speech sound. Children can be trained to recognize the words in a play situation, Younger children or those with motor constraints can use a touch screen to select their choices, while older children can use a mouse or keypad. The test can be delivered in a short form, intended to be appropriate for children from three years upwards or the standard form that incorporates the words in the short form plus additional words that are appropriate for children with developmental ages of five years and above.

The CAPT contains different sections to assess: (1) discrimination of consonants, where the four words differ in just one consonant, for example fat, bat, cat, mat; (2) vowel discrimination, where the four words have the same consonants and differ only in the vowel, for example two, tar, tea, tie or cat, cot, cut, cart; and (3) detection of consonants, where performance depended on the detection of one or more consonants, for example: eye, ice, lice, slice, or why, wine, eye, wise. For the short form of the test there are 28 words for the discrimination of consonants, 12 words for vowel discrimination, and 12 words for consonant detection. For the long form there are 48 words for discrimination of consonants, 20 words for vowel discrimination, and 20 words for consonant discrimination. The test can be separated into the component parts, depending on the perceptual aspect being studied. The most commonly used section is the consonant discrimination section. A point is given for each word scored correctly.

The test-retest reliability of the shortened form of the consonant discrimination section was evaluated with normal-hearing school-aged children and the average critical difference across the performance range was found to be 17.6% (the critical difference varies across the performance range). This means that scores for two conditions would need to differ by 3 or 4 items for the difference across conditions to be considered as significant (Vickers et al. 2013).

Only the long form of the consonant discrimination section was used here because that is the most critical section for assessing the effect of spectral changes (e.g. changes in the frequency-gain characteristic of a hearing aid) and it is the section of the CAPT that is most similar in nature to the CCT for the purpose of the comparison between measures. There were 48 words in total (12 word groups, each containing 4 words; see table 2 for the words in the test). The test was conducted twice within the same test session but with a break in between and a different order of presentation of items for each run.

TABLE 2. Word groups for the CHEAR auditory perception test (CAPT). Note that the vowel sound remains approximately the same within each group and only the word-initial or word-final consonant changes.

	Word 1	Word 2	Word 3	Word 4
Group 1	Mat	Bat	Cat	Fat
Group 2	Wine	Wise	White	Wipe
Group 3	Fin	Tin	Shin	Chin
Group 4	Stork	Talk	Chalk	Fork
Group 5	Bun	Bug	Bud	Buzz
Group 6	Kick	Tick	Thick	Pick
Group 7	White	Right	Light	Night
Group 8	Law	Raw	War	Your
Group 9	What	Wash	Want	Watch
Group 10	Jug	Drug	Bug	Mug
Group 11	Cheap	Cheat	Cheek	Cheese
Group 12	Caught	Call	Corn	Core

Speech Test Delivery

On each trial, the four word options were shown on the screen of the PC. Each word was depicted by a picture with the target word written underneath. The child used a mouse (or touch screen) to select the word they thought that they had heard. Responses were recorded via the PC.

Within a session, each child was tested with the CCT or CAPT alone or with both tests. This was decided at the beginning of the session, based on the child's developmental language age and time commitments. Five children were tested with the CCT alone, 29 with the CAPT alone, and 26 with both tests. The test and the re-test for a given test (CCT or CAPT) were run consecutively, with a ten-minute break in between. If both tests were administered, there was a 15-minute break between administration of the two tests. The CCT was always administered first. If a child had performed poorly on the CCT, they would not have been further tested using the CAPT. However, this did not happen for any child.

Stimuli for both tests were generated via the built-in sound card of a laptop PC (sampling rate = 44100 Hz, 16-bit precision) and presented via Sennheiser HD600 headphones. These headphones have a diffuse-field response and stimuli were presented diotically at an equivalent diffuse-field level of 30 dB SPL (the actual level at the eardrum was higher, especially for frequencies around 3 kHz, because of the diffuse-field response of the headphones). This low level was selected to avoid ceiling effects. In theory, ceiling effects can also be avoided by using background noise, but the speech reception threshold in noise is hardly altered by substantial variations in frequency-gain response (van Buuren et al. 1995), and this test was intended to be sensitive to such variations. For speech with a diffuse-field level of 30 dB SPL, the mean level at the eardrum in a 1/3 octave band around 3 kHz would be about 25 dB SPL, with speech peaks reaching levels of about 37 dB SPL (Moore et al. 2008). Hence, the 30 dB SPL level would have led to a sensation level (SL) of about 25-30 dB. Stimuli with similar SLs are often used in studies with hearing-impaired people, since loudness recruitment precludes the use of high SLs. However, we acknowledge that a child would only rarely have to try to understand speech with a level as low as 30 dB SPL. The implications of the use of this low level are discussed later.

The sensitivity and frequency response of the headphones were checked by mounting them on a KEMAR Type 45DA head assembly, fitted with G.R.A.S. RA0045 ear simulator, 40AG microphone, and 26 AC preamplifier. The input signals were 1 volt 0.125-, 0.25-, 1.0, 2.0-, 4.0-, and 8.0-kHz pure tones. The output of the preamplifier was analyzed using a Hewlett-Packard HP3561A dynamic signal analyzer. Since the headphones have a sensitivity at low frequencies (where the diffuse-field-to-eardrum transfer function has a value close to 0 dB) of 102 dB SPL/V, the sound level of 30 dB SPL was obtained by setting the root-mean-square voltage of the speech at the input to the headphones to $10^{(30-102)/20}$, i.e., 0.25 mV. Calibration procedures for sound field delivery can be found in the companion paper (Marriage et al., 2017).

Conversion of Scores to d′ Values

Percent correct scores were converted to discriminability index (d') scores (Macmillan & Creelman 2005). The value of d' increases monotonically with percent correct for a given number of response alternatives, and it increases monotonically with number of alternatives for a fixed percent correct. The value of d' can be readily obtained from standard tables (Hacker & Ratcliff 1979), although for our data this value is only approximate since it is based on the assumption that all response alternatives are equally confusable with the target, which was probably not the case. There are two advantages of using d' rather than percent correct: d' scores are less affected than percent correct scores by floor and ceiling effects; and d' scores allow approximate comparison across tests with different numbers of response alternatives.

RESULTS

Consonant Confusion Test (CCT)

The mean percent correct score for the CCT was 2.1% higher for the first than for the second run, but a paired t-test showed that the difference was not significant (t = 0.39, df = 30, p = 0.70).

To determine the test-retest reliability of the CCT, an inter-class correlation (ICC) using a two-way random-effects model, with type absolute agreement, with averaged measures was calculated based on the d′ scores for each run of the CCT (Bland & Altman 1986). The ICC showed a very strong agreement of 0.80 between the two runs. A within-subject s_ω (Bland & Altman, 1996) was calculated to derive the 95% confidence interval of the score for an individual. The quantity s_ω is the square root of the mean group variance (mean across individuals of the variance calculated for each individual). An individual's observed score is expected to lie within $\pm 1.96 s_\omega$ of their "true" score (for 95% of observations; the confidence interval). The critical difference is calculated as $\sqrt{2}*1.96s_{\omega}$. If scores obtained on two different occasions differ by $\sqrt{2}*1.96s_{\omega}$ or more, then they differ significantly at $p < 0.05$. The mean values obtained in this way were $s_{\omega} = 0.35d'$, $1.96s_{\omega} = 0.69d'$, and $\sqrt{2}*1.96s_{\omega} = 0.97d'$. When calculated as a percentage, the mean critical difference was 14.2% (the exact value varies across the performance range).

Table 3 shows how the critical difference varies across the performance range and what the critical difference is for an individual score out of 40. The critical differences were calculated in terms of d′, but have been converted back to scores out of 40 for ease of interpretation. As an example, assume that a child scored 30 on the CCT on one occasion. If the child scored between 18 and 37 on the next occasion this would not be viewed as a significant change in performance. However, a score of 38 on the second occasion would be taken as a significant improvement. Figure 1 shows the lower and upper bounds of the critical difference plotted as a function of the initial score.

TABLE 3. Critical differences for the CCT expressed as d′ values and converted back to scores out of 40. The upper and lower values for the critical difference are indicated. Equivalent percentage values are shown in parentheses

Initial score (out of 40)(**%**)	initial score (d')	lower boundary of critical difference (d')	upper boundary of critical difference (d')	lower boundary (score out of 40 (**%**
	x	**x-0.96**	**x+0.96**	
38 (**95**)	2.92	1.95	3.89 (max=3.80)	32 (**80**)
36 (**90**)	2.45	1.48	3.42	28 (**70**)
34 (**85**)	2.14	1.17	3.11	24 (**60**)
32 (**80**)	1.89	0.92	2.86	21 (**53**)
30 (**75**)	1.68	0.71	2.65	18 (**45**)
28 (**70**)	1.49	0.52	2.46	16 (**40**)
26 (**65**)	1.22	0.25	2.19	13 (**33**)
24 (**60**)	1.15	0.18	2.12	12 (**30**)
22 (**55**)	0.99	0.02	1.96	10 (**25**)
20 (**50**)	0.84	-0.13	1.81	9 (**23**)
18 (**45**)	0.68	-0.29	1.65	7 (**18**)
16 (**40**)	0.52	-0.45	1.49	6 (**15**)
14 (**35**)	0.36	-0.61	1.33	5 (**13**)
12 (**30**)	0.19	-0.78	1.16	4 (**10**)
10 (**25**)	0	chance level		

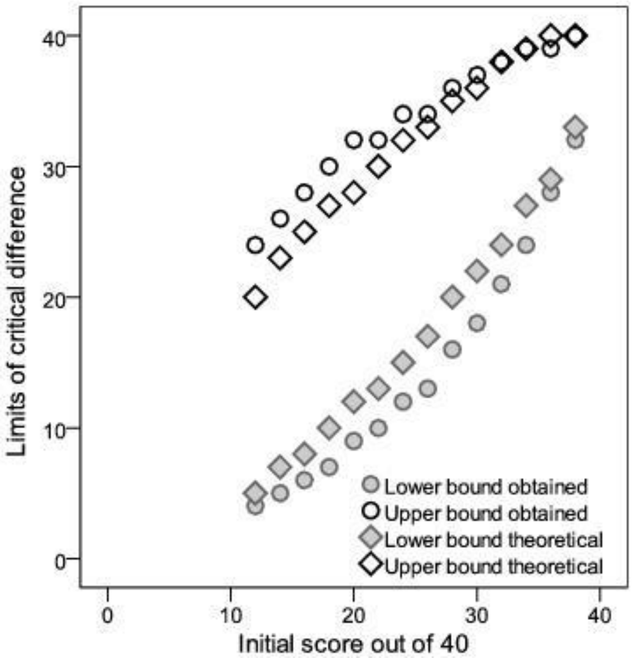

Fig. 1. Upper and lower bounds of the critical difference for the CCT when 40 items are presented. The x-axis shows the score obtained on the first test session. The dark circles show the upper bound and the light circles the lower bound within which a score for a second test would not be considered to be significantly different from that for the first test.

It is of interest to compare the critical difference values in Figure 1 with those that would be expected for a 40-item test, based on the binomial distribution. This was done using the following steps: (1) The initial proportion correct, *P*, was arcsine transformed (= $2\arcsin(\sqrt{P})$); (2) The expected standard deviation, SD_e of the test scores on the same transformed frequency scale (= $1/\sqrt{(N+1)}$) (Thornton & Raffin 1978) was calculated, where N is the number of test items (40 in this case); (3) The value of SD_e was multiplied by $1.96*\sqrt{2}$ to calculate the critical difference in the transformed variable; (4) This critical difference was added to and subtracted from the transformed initial score; (5) The upper and lower bounds of the transformed score were converted back to proportions $(\sin(value/2))^2$, and from that to the corresponding number of items, rounded to the nearest whole number. The outcomes are shown as diamonds in Figure 1. The theoretical critical differences are consistently slightly smaller than the obtained critical differences, by a factor of about 1.3, indicating that the children were not entirely consistent across the two tests, probably reflecting fatigue or boredom, or an increase in proficiency due to practice.

Chear Auditory Perception Test (CAPT)

The mean percent correct score for the CAPT was 1.8% lower for the second than for the first run, but a paired t-test showed that the difference was not significant ($t = -1.69$, df = 54, $p = 0.10$).

A similar test-retest reliability analysis as described above was conducted for the CAPT. The ICC showed very strong agreement between the two test runs of 0.84. The value of s_ω was 0.29d′, so the boundaries of the 95% confidence intervals around a specific obtained score fell at $1.96s_\omega = 0.58d'$, and the critical difference was $\sqrt{2}*1.96s_\omega = 0.82d'$. When calculated as a percentage the mean critical difference was 13.7% (the exact value varies across the performance range). Table 4 shows the critical difference values across the performance range for the CAPT. Figure 2 shows the lower and upper bounds of the critical difference plotted as a function of the initial score.

TABLE 4. As table 3 but for the CAPT.

Initial score (out of 48)(**%**)	initial score (d')	lower boundary of critical difference (d')	upper boundary of critical difference (d')	lower boundary (out of 48)	u
	x	**x-0.82**	**x+0.82**		
46 (**96**)	3.05	2.23	3.87 (max=3.80)	42 (**88**)	
43 (**90**)	2.45	1.63	3.27	36 (**75**)	
41 (**85**)	2.14	1.32	2.96	31 (**65**)	
38 (**79**)	1.89	1.07	2.71	27 (**56**)	
36 (**75**)	1.68	0.86	2.5	24 (**50**)	
34 (**71**)	1.49	0.67	2.31	22 (**46**)	
31 (**65**)	1.22	0.4	2.04	18 (**38**)	
29 (**60**)	1.15	0.33	1.97	17 (**35**)	
26 (**54**)	0.99	0.17	1.81	14 (**29**)	
24 (**50**)	0.84	0.02	1.66	12 (**25**)	
22 (**46**)	0.68	-0.14	1.5	11 (**23**)	
19 (**40**)	0.52	-0.3	1.34	9 (**19**)	
17 (**35**)	0.36	-0.46	1.18	7 (**15**)	
14 (**29**)	0.19	-0.63	1.01	6 (**13**)	
12 (**25**)	0	chance level			

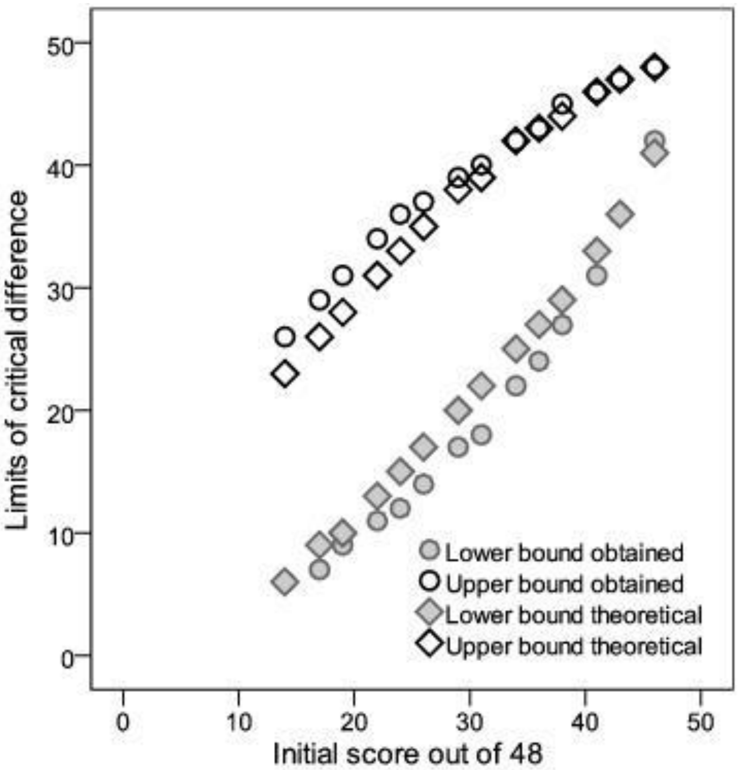

Fig. 2. As figure 1 but for the CAPT when 48 items are presented.

We also calculated the theoretical critical differences based on the binomial distribution, as described for the CCT but with $N = 48$. Again, the theoretical critical differences were consistently slightly smaller than the obtained critical differences, by a factor of about 1.2, indicating that the children were not entirely consistent across the two tests.

Relationship of Scores with Age and Language Group for the CCT and CAPT

An analysis of the linear relationship between age and scores for the CCT and CAPT was conducted using the Pearson product-moment correlation for children for whom both chronological and expressive vocabulary age were available. This was done because a high proportion of the participants fell in the EAL group, so it could not be assumed that their chronological age was a valid indicator of their developmental language age. Both chronological and expressive vocabulary age were used in the analyses.

For the CCT, both chronological age and expressive vocabulary age were available for all 31 children. The mean chronological and expressive vocabulary ages were 63 and 76 months, respectively, for the E1L group (13 children) and 82 and 79 months for the EAL group (18 children). For the CAPT there were 43 children for whom both vocabulary age and chronological age were available. For these, the mean chronological and expressive vocabulary ages were 79 and 91 months, respectively, for the E1L group (19 children) and 83 and 82 months, respectively, for the EAL group (24 children). Figures 3 and 4 show the distributions of the chronological and expressive vocabulary ages for the children assessed with each test.

To avoid ceiling effects, individuals with average scores on the tests that were close to ceiling (above a d′ value of 3.30, corresponding to about 97.5%) were excluded from the correlation analyses. This avoided outliers having a strong influence on the correlations.

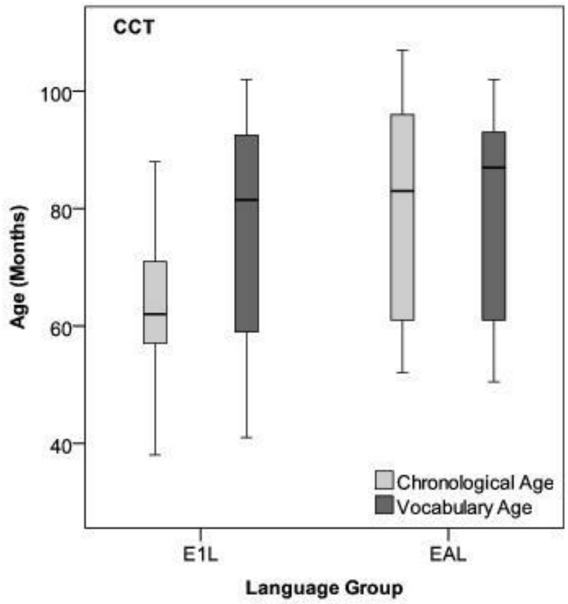

Fig. 3. Boxplots of scores for the children tested with the CCT to show the distribution of vocabulary age and chronological age (in months), separated into those with English as first language (E1L) and those with English as an additional language (EAL). The light and dark boxes indicate chronological and vocabulary age, respectively. The line in the boxes shows the median and the whiskers indicate the range of values.

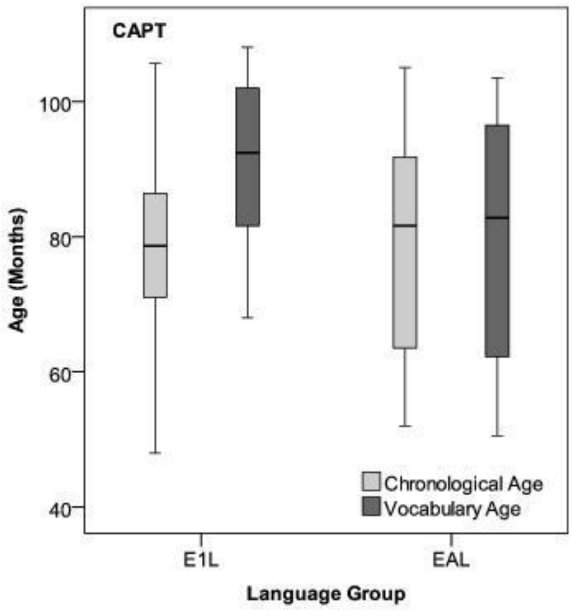

Fig. 4. As for figure 3, but for the children tested with the CAPT.

Figure 5 shows a scatter plot of performance on the CCT versus age for the 17 children whose scores were not excluded. There were significant correlations between d′ scores and both chronological age (r = 0.68, n = 17, p < 0.002; the equation for the relationship was d′ = 0.028age+0.50, where age is in months) and vocabulary age (r = 0.64, n = 17, p = 0.006; the equation for the relationship was d′ = 0.026age+0.64).

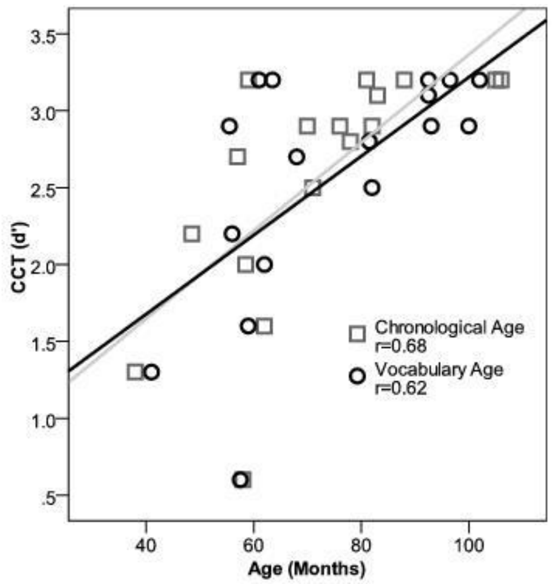

Fig. 5. Scatter plots showing the relationship between age (in months) and d′ score for the CCT. Grey squares indicate chronological age and dark circles indicate vocabulary age. Pearson correlation coefficients are shown in brackets.

For the CAPT, the scores for 37 children were included in the correlation analysis. A scatter plot for these is shown in figure 6. There were significant correlations between d′ scores and both chronological age (r = 0.36, n = 37, p = 0.03; the equation for the relationship was d′ = 0.010age+1.33) and vocabulary age (r = 0.43, n = 37, p = 0.008; the equation for the relationship was d′ = 0.011age+1.23).

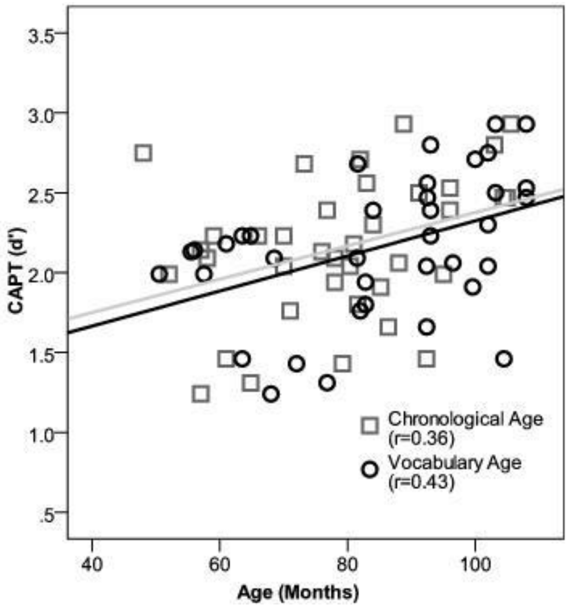

Fig. 6. As for figure 5 but for the CAPT scores.

For both the CCT and the CAPT the correlation was similar for vocabulary and for chronological age. For the CCT the relationship with age accounted for approximately 43% of the variance in the scores (chronological = 46%; vocabulary age = 41%) whereas for the CAPT the relationship with age only accounted for approximately 15% of the variance (chronological age = 13%; vocabulary age = 19%). This finding of a lower strength of the relationship between age and score for the CAPT than for the CCT occurred partly because the vocabulary level requirement for the CAPT prevented very young children from taking the test, so the spread of ages was larger for the CCT. For the CCT the age range of the children tested was from 38 to 107 months and performance ranged from 32.5 to 100%. This enabled a rough estimate of the appropriate age range of the test to be derived based on the regression equation relating vocabulary age in months to d′ score. Assuming that for a test to be sensitive to change the child should score between 60 and 85% (d′ values of 1.15 and 2.14, respectively), the appropriate age range for the CCT is approximately 23 to 59 months. This estimate is approximate because it involves some extrapolation for the lower age limit. If a child scores at the upper limit of the CCT test, the child should be tested with the CAPT if it is desired to track changes in performance over time or to assess the effect of an intervention.

To create percentile charts to provide guidance on the normative ranges for the CCT and the CAPT, smoothed reference percentile curves were generated using the LMS method (Cole & Green 1992). The method summarizes the age dependence of three variables: L – the coefficient of variation; M - the median; and S – the skewness. This is done using a method called “penalized maximum likelihood” (Green 1987). The percentile curves were generated based on vocabulary age, so that they would be applicable to children whose chronological and vocabulary ages were different. The outcomes for the CCT are shown in figure 7. Curves were created based on d′ scores and converted to percent correct for ease of interpretation.

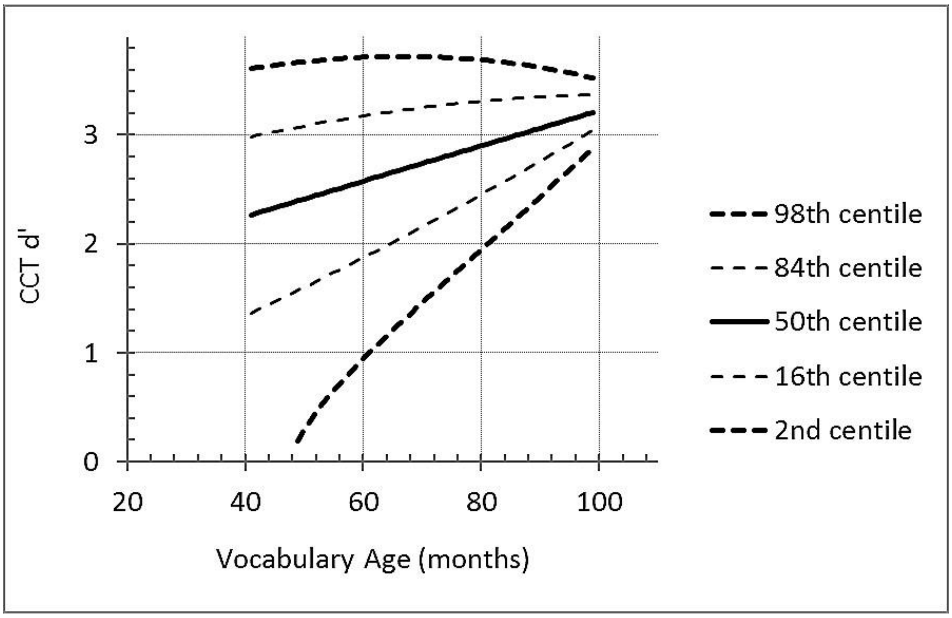

Fig. 7. Percentiles for CCT score as a function of vocabulary age, generated using the LMS method. The 16% and 84% percentiles represent – 1 SD and + 1 SD, respectively.

The relationship between developmental age and performance in the CCT is shown in figure 7, and this figure should be used to determine if a child’s performance is within 1 standard deviation (16th and 84th percentile) of the mean for their vocabulary age. Once the child reaches a performance level of 80%, it would be appropriate to transfer the child to testing with the CAPT.

The relationship with developmental age was weaker for the CAPT than for the CCT; the percentiles are shown in figure 8. There was a smaller range of performance than for the CCT.

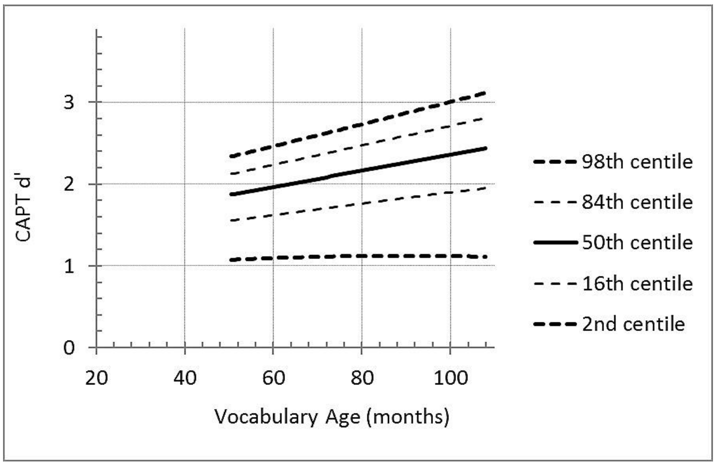

Fig. 8. As for figure 7 but for the CAPT scores.

DISCUSSION

We have presented data for two monosyllabic closed-set consonant discrimination tests that can be used with young children. The goals were to present normative data, determine the reliability of the tests, determine if there was an effect of age on performance, and assess whether the two tests could be used to evaluate consonant perception across the age range 2-8 years, avoiding ceiling and floor effects. One reason for the choice of this age range was that we required a suitable test battery for assessing the performance of young hearing-impaired children in study comparing different gain prescriptions for hearing aids (Marriage et al. 2017), and there were not any validated British English measures of consonant discrimination that could be used for young children. Also, the tests available for different ages were different in nature, so comparisons across age groups was not possible.

The CCT was developed to have vocabulary items that are appropriate for children from two years old, while the CAPT was developed to be appropriate for slightly older children. The words in both tests are nouns and can be represented by pictures. The items of the CCT are easier to discriminate, because differences occur in both the initial and final consonants for each of the words in a group of four. This restriction arose because of the small pool of vocabulary-appropriate word for children with ages of two years. For the CAPT, the consonant contrasts were in word-initial or word-final position, but not both.

Both the CCT and CAPT demonstrated strong agreement between the two test runs, with ICC values of 0.80 and 0.84 for the CCT and CAPT, respectively. The critical differences for the two tests, presented in Table 3 and 4, can be used to determine whether or not changes in performance that are observed for an individual child are significant. The critical difference values are slightly higher than the obtained values reported by Thornton and Raffin (1978) for monosyllabic words (50 word version of the CID word test W-22) presented to adults. The larger values found here are probably due to the respondents being young children and to the smaller number of independent test items (10 groups of four words for the CCT and 12 groups of four words for the CAPT). Critical difference values are seldom provided for paediatric speech measures (except for the McCormick Toy Test, which gives an estimate of the speech reception threshold rather than a percentage correct speech score), so the present critical differences cannot be compared to those for other paediatric speech tests.

Reliability can be increased by conducting multiple runs of a test, but this can be unrealistic when testing young children, because of their limited attention span. The problem of limited attention span can be partly overcome by using a variety of assessment materials, which also ensures that a full picture of abilities is determined. However, a method is then needed to derive a single composite score from the multiple measures. Such a composite score might have greater reliability than the score for any single test. When combining data across groups to compare different conditions (e.g. comparing two hearing-aid signal-processing schemes) it is probably sufficient to use a single run of each test with each child, provided that an appropriate number of children are assessed.

The analysis of the relationship between chronological and vocabulary age and performance showed a significant correlation for both tests, for both chronological and vocabulary age. The correlations were higher for the CCT, probably because the vocabulary for the CCT was more challenging for the young children and because younger children could not be tested using the CAPT as their vocabulary was inadequate. The correlation of performance with age on both tests is consistent with previous results showing that the ability to understand speech improves with increasing age up to the early teens (Stelmachowicz et al. 2000; Vance et al. 2009). The improvement presumably reflects the combined effects of maturation of auditory and cognitive skills and greater experience of the language.

Some limitations of the tests and of our study should be noted. Firstly, the critical differences are larger than would be desired for both tests, making it difficult to use the tests to identify small changes in performance of an individual child, for example, as a result of changing the fitting of a hearing aid. Secondly, the tests were conducted using sounds presented at 30 dB SPL, which is lower than the typical levels of speech encountered in everyday life. While the speech sounds were clearly audible to the normal-hearing children used here, sounds with such low levels would often not be audible to children with hearing loss, even when amplification was provided. It is unclear whether the critical differences found here are applicable to children with hearing loss tested at higher levels, although we do not know of any theoretical reason why this should be the case. Thirdly, several of the older children performed at ceiling, despite the fact that all of our testing was conducted using a low level of 30 dB SPL. It would be unrealistic to test at an even lower level in an attempt to completely avoid ceiling effects. An alternative approach is to use background noise, such as speech-shaped noise, to limit performance, but, as noted earlier, the ability to understand speech in noise at moderate overall levels is hardly affected by quite large changes in frequency-gain response (van Buuren et al. 1995), and this would make the tests insensitive to interventions such as changes in the fitting of a hearing aid. However, background noise could be used if the goal were to evaluate other types of interventions, such as the effectiveness of a noise-reduction algorithm. Further work is needed to determine the test-retest reliability and critical differences for the CCT and CAPT under conditions where background noise is present, although we do not know of any theoretical reason why they should differ from those found here.

CONCLUSIONS

Both the CCT and CAPT can be used to assess consonant perception by young children. When applied to a group of normal hearing children ranging in age from 2 to 9 years both tests have high test re-test reliability, with intra-class correlations ≥ 0.8. Performance on both tests improved with increasing age, the correlation of performance with age being greater for the CCT than for the CAPT. The consonants used in the tests contain acoustic cues with a wide range of spectro-temporal characteristics, so the tests should be sensitive to the effects of different patterns of hearing loss and to changes in the characteristics of hearing aids. The children tested were all able to use a mouse or a touchscreen to select response alternatives on a screen, which avoided any influence of the child's speech production abilities or of the tester's ability to understand poorly articulated speech. The CCT is appropriate for children up to the age of 54 months, or for older children whose vocabulary age is below their chronological age. If a child scores 80% or more on the CCT they can be further tested using the CAPT, which contains more advanced vocabulary and more difficult contrasts. This allows the assessment of consonant perception ability and changes over time or following an intervention.

ACKNOWLEDGEMENTS

We thank Deafness Research UK (now part of Action on Hearing Loss) for funding, John Deeks for help with calibration, and the children and their families. We also thank the editorial board of Ear and Hearing and three reviewers for helpful comments.

REFERENCES

Archbold, S., Lutman, M. E., & Marshall, D. H. (1995). Categories of auditory performance. *Ann Otol Rhinol Laryngol*, *166*, 312-314.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, *1*, 307-310.

Boothroyd, A., Eisenberg, L. S., & Martinez, A. (2005). *OLIMSPAC. Version 3.1d*. Los Angeles, CA: House Ear Institute.

Byrne, D., Dillon, H., Ching, T., et al. (2001). NAL-NL1 procedure for fitting nonlinear hearing aids: characteristics and comparisons with other procedures. *J Am Acad Audiol*, *12*, 37-51.

Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and penalized likelihood. *Stat Med*, *11*, 1305-1319.

Cornelisse, L. E., Seewald, R. C., & Jamieson, D. G. (1995). The input/output formula: A theoretical approach to the fitting of personal amplification devices. *J Acoust Soc Am*, *97*, 1854-1864.

Cullington, H., Bele, D., Brinton, J., et al. (2013). United Kingdom national paediatric bilateral cochlear implant audit: preliminary results. *Cochlear Implants Int*, *14 Suppl 4*, S22-26.

Govaerts, P. J., Daemers, K., Yperman, M., et al. (2006). Auditory speech sounds evaluation (A§E®): a new test to assess detection, discrimination and identification in hearing impairment. *Cochlear Implants Int*, *7*, 92-106.

Green, P. (1987). Penalized likelihood for general semi-parametric regression models. *Int Stat Rev*, *55*, 245-259.

Hacker, M. J., & Ratcliff, R. (1979). A revised table of d' for M-alternative forced choice. *Percept Psychophys*, *26*, 168-170.

Holt, R. F., & Lalonde, K. (2012). Assessing toddlers' speech-sound discrimination. *Int J Pediatr Otorhinolaryngol*, *76*, 680-692.

Jerger, S., Lewis, S., Hawkins, J., et al. (1980). Pediatric speech intelligibility test. I. Generation of test materials. *Int J Pediatr Otorhinolaryngol*, *2*, 217-230.

Kirk, U. (Ed.). (2012). *Neuropsychology of Language, Reading and Spelling*. New York: Academic Press.

Lovett, R., Summerfield, Q., & Vickers, D. (2013). Test-retest reliability of the Toy Discrimination Test with a masker of noise or babble in children with hearing impairment. *Int J Audiol*, *52*, 377-384.

Lovett, R. E., Vickers, D. A., & Summerfield, A. Q. (2015). Bilateral cochlear implantation for hearing-impaired children: criterion of candidacy derived from an observational study. *Ear Hear*, *36*, 14-23.

Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide, 2nd Ed.* New York: Erlbaum.

Marriage, J. E., & Moore, B. C. J. (2003). New speech tests reveal benefit of wide-dynamic-range, fast-acting compression for consonant discrimination in children with moderate to severe hearing loss. *Int J Audiol*, *42*, 418-425.

Marriage, J. E., Vickers, D. A., Baer, T., et al. (2017). Comparison of different hearing aid prescriptions for children. *Ear Hear*, (submitted).

McConkey Robbins, A., Koch, D. B., Osberger, M. J., et al. (2004). Effect of age at cochlear implantation on auditory skill development in infants and toddlers. *Arch Otolaryngol Head Neck Surg*, *130*, 570-574.

Moore, B. C. J., Stone, M. A., Füllgrabe, C., et al. (2008). Spectro-temporal characteristics of speech at high frequencies, and the potential for restoration of audibility to people with mild-to-moderate hearing loss. *Ear Hear*, *29*, 907-922.

Reed, M. (1959). A verbal screening test of hearing. In *Proceedings of III World Congress of the Deaf*. Wiesbaden, Germany: Deutscher Gehorlosen.

Renfrew, C. (1995). *Word Finding Vocabulary Test*. Oxford: Winslow Press.

Stiles, D. J., Bentler, R. A., & McGregor, K. K. (2012). The Speech Intelligibility Index and the pure-tone average as predictors of lexical ability in children fit with hearing aids. *J Speech Lang Hear Res*, *55*, 764-778.

Thornton, A. R., & Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial variable. *J Speech Hear Res*, *21*, 507-518.

Tomblin, J. B., Harrison, M., Ambrose, S. E., et al. (2015). Language outcomes in young children with mild to severe hearing loss. *Ear Hear*, *36 Suppl 1*, 76S-91S.

van Buuren, R. A., Festen, J. M., & Plomp, R. (1995). Evaluation of a wide range of amplitude-frequency responses for the hearing impaired. *J Speech Hear Res*, *38*, 211-221.

Vickers, D. A., Backus, B. C., Macdonald, N. K., et al. (2013). Using personal response systems to assess speech perception within the classroom: an approach to determine the efficacy of sound field amplification in primary school classrooms. *Ear Hear*, *34*, 491-502.

Zhou, H., Chen, Z., Shi, H., et al. (2013). Categories of auditory performance and speech intelligibility ratings of early-implanted children without speech training. *PLoS One*, *8*, e53852.

Fig. 1. The circles show the upper and lower bounds of the critical difference for the CCT when 40 items are presented. The diamonds show theoretical values based on the binomial distribution (see text). The x-axis shows the score obtained on the first test session. The dark open symbols show the upper bound and the light filled symbols the lower bound within which a score for a second test would not be considered to be significantly different from that for the first test.

Fig. 2. As figure 1 but for the CAPT when 48 items are presented.

Fig. 3. Boxplots of scores for the children tested with the CCT to show the distribution of vocabulary age and chronological age (in months), separated into those with English as first language (E1L) and those with English as an additional language (EAL). The light and dark boxes indicate chronological and vocabulary age, respectively. The boxes show the inter-quartile range, the lines in the boxes shows the medians, and the whiskers indicate the range of values.

Fig. 4. As for figure 3, but for the children tested with the CAPT.

Fig. 5. Scatter plots showing the relationship between age (in months) and d′ score for the CCT. Grey squares indicate chronological age and dark circles indicate vocabulary age. Pearson correlation coefficients are shown in brackets.

Fig. 6. As for figure 5 but for the CAPT scores.

Fig. 7. Percentiles for CCT score as a function of vocabulary age, generated using the LMS method. The 16% and 84% percentiles represent – 1 SD and + 1 SD, respectively.

Fig. 8. As for figure 7 but for the CAPT scores.