

1 **Closed Set Speech Discrimination Tests for Assessing Young Children**

2

3 Deborah A. Vickers<sup>1</sup>, Brian C.J. Moore<sup>2</sup>, Arooj Majeed<sup>3</sup>, Natalie Stephenson<sup>3</sup>, Hala Alferaih<sup>3</sup>,

4 Thomas Baer<sup>2</sup>, and Josephine E. Marriage<sup>4</sup>

5

6

7 <sup>1</sup>University College London, Department of Speech Hearing and Phonetic Sciences, 2

8 Wakefield Street, London. WC1N 1PF, UK

9

10 <sup>2</sup>Dept of Experimental Psychology, University of Cambridge, Downing St, Cambridge CB2

11 3EB, UK

12

13 <sup>3</sup>University College London, Ear Institute, 332 Grays Inn Road, London. WC1X 8EE, UK

14

15 <sup>4,2</sup>Chear, 30 Fowlmere Road, Shepreth, Royston, Herts SG8 6QS, UK

16

17

18 Send correspondence to:

19

20 Deborah A. Vickers

21 email: [d.vickers@ucl.ac.uk](mailto:d.vickers@ucl.ac.uk)

22

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

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

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

39 **Conclusions:** The CCT is appropriate for children with developmental age in the range 2 to  
40 4.5 years and the CAPT is appropriate as a follow on test from the CCT. If a child scores  
41 80% or more on the CCT they can be further tested using the CAPT, which contains more  
42 advanced vocabulary and more difficult contrasts. This allows the assessment of consonant  
43 perception ability and of changes over time or following an intervention.

44

45

## INTRODUCTION

46

47

48

49

50

51

52

53

54

55

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.

56

57

58

59

60

61

62

63

64

65

66

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.

67

68

69

70

71

72

73

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

74 appropriately assessed, but makes it difficult to assess long-term developmental trends for  
75 individual children or to compare results for different age groups.

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

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

93         Parental-response questionnaires are typically used with children under four years of  
94 age. The subjectivity of these can make them insensitive to small changes. However, a  
95 validated questionnaire can be useful for monitoring relatively large changes in auditory  
96 perception over time. The Infant Toddler Meaningful Auditory Integration Scale (IT-MAIS,  
97 McConkey Robbins et al. 2004) is a validated measure, with known normative ranges, that  
98 has been shown to be sensitive to changes in perception with age. The Categories of Auditory  
99 Performance test (CAP; Archbold et al. 1995) uses a hierarchical rating scale with eight  
100 levels of auditory perception from “no awareness of environmental sounds” to “uses the  
101 telephone”. Although this appears to be a fairly gross measure, it has been shown to be  
102 sensitive to differences in performance over time, as exemplified over the first 12 months of

103 hearing experience for children receiving cochlear implants at an early age (Zhou et al. 2013).  
104 However, for both the IT-MAIS and the CAP, the variability of the outcomes, the limited  
105 number of discrete scores, and the subjective nature of the responses, prevent these measures  
106 from being viable in assessing the impact of small changes in sound delivery, for example,  
107 changes in the frequency-gain characteristic of a hearing aid. They do, however, have a role  
108 in detecting gross changes in perception.

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

118 Thornton and Raffin (1978) calculated theoretical critical differences for the CID W-  
119 22 word test and compared them with obtained critical differences. The CID W-22 test is an  
120 open-set monosyllabic speech perception test. They showed that both the theoretical and  
121 obtained critical difference values were greatly affected by the number of items used to  
122 evaluate each condition. This leads to a dilemma when using speech tests with children: for  
123 example, presenting ten words per condition would not provide a sufficiently reliable  
124 measure of any change in performance across conditions, but presentation of many more  
125 items to increase reliability could make the test too time consuming for clinical practice or  
126 could lead to loss of attention of the child. Probably because of the limited reliability of the  
127 speech scores obtained with young children, many studies on early intervention for hearing-  
128 impaired children do not report speech recognition scores for children younger than about six  
129 years (Davidson & Skinner 2006; Strauss & van Dijk 2008).

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

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

141 Other closed-set tests for young children, using American English, are the pediatric  
142 speech intelligibility (PSI) test (Jerger et al. 1980) and the online imitative test of speech  
143 pattern contrast perception (OLIMSPAC) (Boothroyd et al. 2005). The critical differences for  
144 these tests have not been reported, making the results difficult to interpret on a case-by-case  
145 basis. However, the tests have been demonstrated to be effective measures for group level  
146 data; see, for example, Sininger et al. (2010). Holt and Lalonde (2012) described a test  
147 assessing toddler speech sound discrimination, for two different contrasts, using a change/no  
148 change paradigm. They measured test-retest reliability for normal-hearing 2- and 3-year old  
149 children and found a strong correlation between scores for two successive runs ( $r = 0.886$ ,  $p$   
150  $= 0.037$ ). However, critical difference values were not presented.

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

161 age range.

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

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

170

171

## METHOD

172

### **Ethical Approval**

174 Ethical approval was obtained from the University College London ethics committee  
175 (4059/001).

176

### **Participants**

178 Thirty one children aged between 38 and 107 months (mean age = 74 months; 19  
179 females and 12 males) were assessed with the CCT and 55 children aged between 48 and 107  
180 months (mean age = 81 months; 31 females and 24 males) were assessed with the CAPT. All  
181 children were screened with pure-tone audiometry at the beginning of the session to have  
182 hearing thresholds at 0.5, 1.0, 2.0 and 4.0 kHz that were less than 20 dB HL. The following  
183 demographic characteristics were collected: chronological age, whether English was their  
184 only language or one of two or more languages spoken, and results of the Renfrew word-  
185 finding vocabulary scale (Renfrew 1995). The latter provides a quick assessment of  
186 expressive vocabulary based on a picture-naming task, giving gender-appropriate age-  
187 equivalent scores. These scores are used as a proxy for English language development. To be

188 included children were required to have sufficient speech and language skills to be able to  
189 understand and participate in the tests. This was evaluated by showing them cards of the  
190 pictures used in the tests to ensure that they understood what word was associated with each  
191 picture. Of the children tested with the CCT, 13 had English as their only language (E1L  
192 group) and 18 spoke more than one language (English as additional language, EAL, group).  
193 Of the children tested with the CAPT, 22 fell in the E1L group and 33 in the EAL group. All  
194 children were attending English speaking nurseries or schools and had intelligible spoken  
195 English.

196

### 197 **Consonant Confusion Test (CCT)**

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

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



215 **TABLE 1. Word groups for the consonant confusion test (CCT). Note that the vowel**  
 216 **sound remains approximately the same within each group but word-initial and word-**  
 217 **final consonants can change**  
 218

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

219

### 220 **Chear Auditory Perception Test (CAPT)**

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

230 The CAPT contains different sections to assess: (1) discrimination of consonants,

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

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

248         Only the long form of the consonant discrimination section was used here because  
249 that is the most critical section for assessing the effect of spectral changes (e.g. changes in the  
250 frequency-gain characteristic of a hearing aid) and it is the section of the CAPT that is most  
251 similar in nature to the CCT for the purpose of the comparison between measures. There were  
252 48 words in total (12 word groups, each containing 4 words; see table 2 for the words in the  
253 test). The test was conducted twice within the same test session but with a break in between  
254 and a different order of presentation of items for each run.

255

256

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

260

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

261

### 262 **Speech Test Delivery**

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

267 Within a session, each child was tested with the CCT or CAPT alone or with both  
 268 tests. This was decided at the beginning of the session, based on the child's developmental  
 269 language age and time commitments. Five children were tested with the CCT alone, 29 with  
 270 the CAPT alone, and 26 with both tests. The test and the re-test for a given test (CCT or  
 271 CAPT) were run consecutively, with a ten-minute break in between. If both tests were  
 272 administered, there was a 15-minute break between administration of the two tests. The CCT

273 was always administered first. If a child had performed poorly on the CCT, they would not  
274 have been further tested using the CAPT. However, this did not happen for any child.

275         Stimuli for both tests were generated via the built-in sound card of a laptop PC  
276 (sampling rate = 44100 Hz, 16-bit precision) and presented via Sennheiser HD600  
277 headphones. These headphones have a diffuse-field response and stimuli were presented  
278 diotically at an equivalent diffuse-field level of 30 dB SPL (the actual level at the eardrum  
279 was higher, especially for frequencies around 3 kHz, because of the diffuse-field response of  
280 the headphones). This low level was selected to avoid ceiling effects. In theory, ceiling  
281 effects can also be avoided by using background noise, but the speech reception threshold in  
282 noise is hardly altered by substantial variations in frequency-gain response (van Buuren et al.  
283 1995), and this test was intended to be sensitive to such variations. For speech with a diffuse-  
284 field level of 30 dB SPL, the mean level at the eardrum in a 1/3 octave band around 3 kHz  
285 would be about 25 dB SPL, with speech peaks reaching levels of about 37 dB SPL (Moore et  
286 al. 2008). Hence, the 30 dB SPL level would have led to a sensation level (SL) of about 25-  
287 30 dB. Stimuli with similar SLs are often used in studies with hearing-impaired people, since  
288 loudness recruitment precludes the use of high SLs. However, we acknowledge that a child  
289 would only rarely have to try to understand speech with a level as low as 30 dB SPL. The  
290 implications of the use of this low level are discussed later.

291         The sensitivity and frequency response of the headphones were checked by mounting  
292 them on a KEMAR Type 45DA head assembly, fitted with G.R.A.S. RA0045 ear simulator,  
293 40AG microphone, and 26 AC preamplifier. The input signals were 1 volt 0.125-, 0.25-, 1.0,  
294 2.0-, 4.0-, and 8.0-kHz pure tones. The output of the preamplifier was analyzed using a  
295 Hewlett-Packard HP3561A dynamic signal analyzer. Since the headphones have a sensitivity  
296 at low frequencies (where the diffuse-field-to-eardrum transfer function has a value close to 0  
297 dB) of 102 dB SPL/V, the sound level of 30 dB SPL was obtained by setting the root-mean-

298 square voltage of the speech at the input to the headphones to  $10^{(30-102)/20}$ , i.e., 0.25 mV.

299 Calibration procedures for sound field delivery can be found in the companion paper

300 (Marriage et al., 2017).

301

### 302 **Conversion of Scores to d' Values**

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

313

## RESULTS

### 314 **Consonant Confusion Test (CCT)**

315 The mean percent correct score for the CCT was 2.1% higher for the first than for the  
316 second run, but a paired t-test showed that the difference was not significant ( $t = 0.39$ ,  $df =$   
317  $30$ ,  $p = 0.70$ ).

318 To determine the test-retest reliability of the CCT, an inter-class correlation (ICC)  
319 using a two-way random-effects model, with type absolute agreement, with averaged  
320 measures was calculated based on the d' scores for each run of the CCT (Bland & Altman  
321 1986). The ICC showed a very strong agreement of 0.80 between the two runs. A within-  
322 subject  $s_{\omega}$  (Bland & Altman, 1996) was calculated to derive the 95% confidence interval of  
323 the score for an individual. The quantity  $s_{\omega}$  is the square root of the mean group variance  
324 (mean across individuals of the variance calculated for each individual). An individual's  
325 observed score is expected to lie within  $\pm 1.96s_{\omega}$  of their "true" score (for 95% of

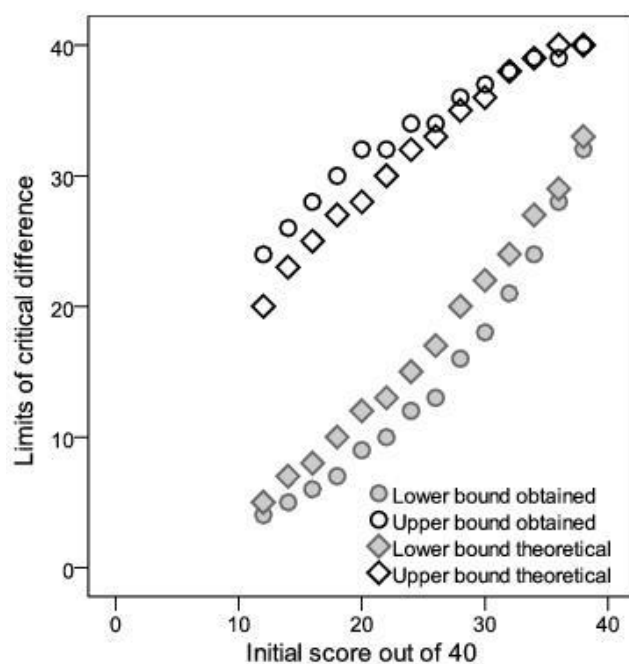
326 observations; the confidence interval). The critical difference is calculated as  $\sqrt{2} * 1.96s_{\sigma}$ . If  
 327 scores obtained on two different occasions differ by  $\sqrt{2} * 1.96s_{\sigma}$  or more, then they differ  
 328 significantly at  $p < 0.05$ . The mean values obtained in this way were  $s_{\sigma} = 0.35d'$ ,  $1.96s_{\sigma} =$   
 329  $0.69d'$ , and  $\sqrt{2} * 1.96s_{\sigma} = 0.97d'$ . When calculated as a percentage, the mean critical difference  
 330 was 14.2% (the exact value varies across the performance range).

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

339 **TABLE 3. Critical differences for the CCT expressed as  $d'$  values and**  
 340 **converted back to scores out of 40. The upper and lower values for the**  
 341 **critical difference are indicated. Equivalent percentage values are shown in**  
 342 **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 (%))
	<b>x</b>	<b>x-0.96</b>	<b>x+0.96</b>	
38 ( <b>95</b> )	2.92	1.95	3.89 (max=3.80)	32 ( <b>80</b> )
36 ( <b>90</b> )	2.45	1.48	3.42	28 ( <b>70</b> )
34 ( <b>85</b> )	2.14	1.17	3.11	24 ( <b>60</b> )
32 ( <b>80</b> )	1.89	0.92	2.86	21 ( <b>53</b> )
30 ( <b>75</b> )	1.68	0.71	2.65	18 ( <b>45</b> )
28 ( <b>70</b> )	1.49	0.52	2.46	16 ( <b>40</b> )
26 ( <b>65</b> )	1.22	0.25	2.19	13 ( <b>33</b> )
24 ( <b>60</b> )	1.15	0.18	2.12	12 ( <b>30</b> )
22 ( <b>55</b> )	0.99	0.02	1.96	10 ( <b>25</b> )
20 ( <b>50</b> )	0.84	-0.13	1.81	9 ( <b>23</b> )
18 ( <b>45</b> )	0.68	-0.29	1.65	7 ( <b>18</b> )
16 ( <b>40</b> )	0.52	-0.45	1.49	6 ( <b>15</b> )
14 ( <b>35</b> )	0.36	-0.61	1.33	5 ( <b>13</b> )
12 ( <b>30</b> )	0.19	-0.78	1.16	4 ( <b>10</b> )
10 ( <b>25</b> )	0	chance level		

343



344

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

350

351 It is of interest to compare the critical difference values in Figure 1 with those that would be  
 352 expected for a 40-item test, based on the binomial distribution. This was done using the  
 353 following steps: (1) The initial proportion correct,  $P$ , was arcsine transformed ( $=$   
 354  $2\arcsin(\sqrt{P})$ ); (2) The expected standard deviation,  $SD_e$  of the test scores on the same  
 355 transformed frequency scale ( $= 1/\sqrt{(N+1)}$ ) (Thornton & Raffin 1978) was calculated, where  
 356  $N$  is the number of test items (40 in this case); (3) The value of  $SD_e$  was multiplied by  
 357  $1.96*\sqrt{2}$  to calculate the critical difference in the transformed variable; (4) This critical  
 358 difference was added to and subtracted from the transformed initial score; (5) The upper and  
 359 lower bounds of the transformed score were converted back to proportions  $(\sin(\text{value}/2))^2$ ,

360 and from that to the corresponding number of items, rounded to the nearest whole number.  
361 The outcomes are shown as diamonds in Figure 1. The theoretical critical differences are  
362 consistently slightly smaller than the obtained critical differences, by a factor of about 1.3,  
363 indicating that the children were not entirely consistent across the two tests, probably  
364 reflecting fatigue or boredom, or an increase in proficiency due to practice.

365

### 366 **Chear Auditory Perception Test (CAPT)**

367 The mean percent correct score for the CAPT was 1.8% lower for the second than for  
368 the first run, but a paired t-test showed that the difference was not significant ( $t = -1.69$ ,  $df =$   
369  $54$ ,  $p = 0.10$ ).

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

378

379



380 **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
	<b>x</b>	<b>x-0.82</b>	<b>x+0.82</b>		
46 ( <b>96</b> )	3.05	2.23	3.87 (max=3.80)	42 ( <b>88</b> )	
43 ( <b>90</b> )	2.45	1.63	3.27	36 ( <b>75</b> )	
41 ( <b>85</b> )	2.14	1.32	2.96	31 ( <b>65</b> )	
38 ( <b>79</b> )	1.89	1.07	2.71	27 ( <b>56</b> )	
36 ( <b>75</b> )	1.68	0.86	2.5	24 ( <b>50</b> )	
34 ( <b>71</b> )	1.49	0.67	2.31	22 ( <b>46</b> )	
31 ( <b>65</b> )	1.22	0.4	2.04	18 ( <b>38</b> )	
29 ( <b>60</b> )	1.15	0.33	1.97	17 ( <b>35</b> )	
26 ( <b>54</b> )	0.99	0.17	1.81	14 ( <b>29</b> )	
24 ( <b>50</b> )	0.84	0.02	1.66	12 ( <b>25</b> )	
22 ( <b>46</b> )	0.68	-0.14	1.5	11 ( <b>23</b> )	
19 ( <b>40</b> )	0.52	-0.3	1.34	9 ( <b>19</b> )	
17 ( <b>35</b> )	0.36	-0.46	1.18	7 ( <b>15</b> )	
14 ( <b>29</b> )	0.19	-0.63	1.01	6 ( <b>13</b> )	
12 ( <b>25</b> )	0	chance level			

381

382

383

384

385

386

387

388

389

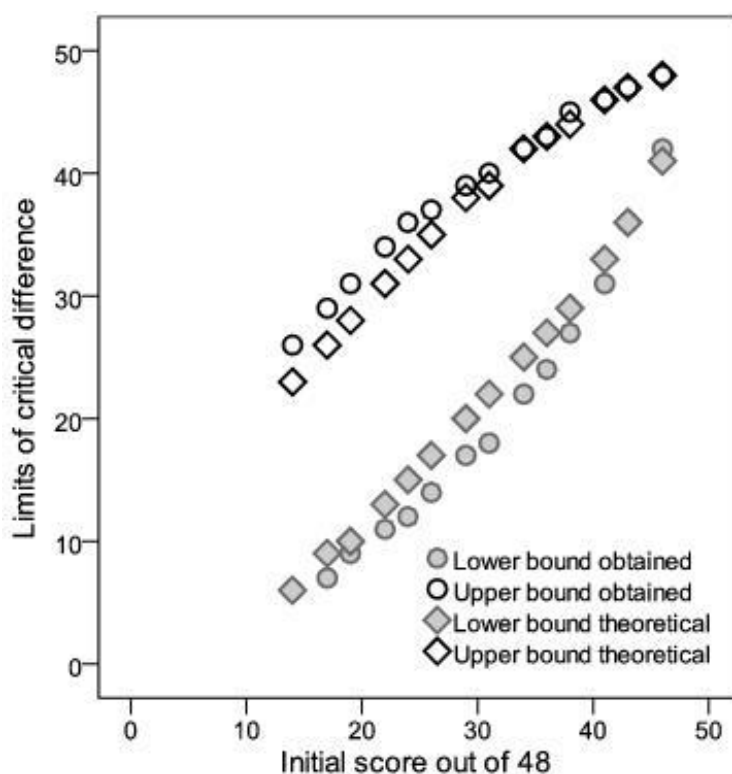
390

391

392

393

394



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

396

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

401

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

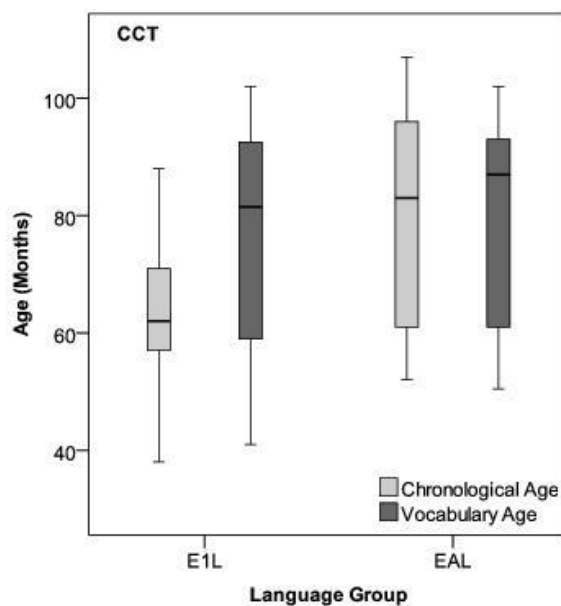
403 An analysis of the linear relationship between age and scores for the CCT and CAPT  
404 was conducted using the Pearson product-moment correlation for children for whom both  
405 chronological and expressive vocabulary age were available. This was done because a high  
406 proportion of the participants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

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

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

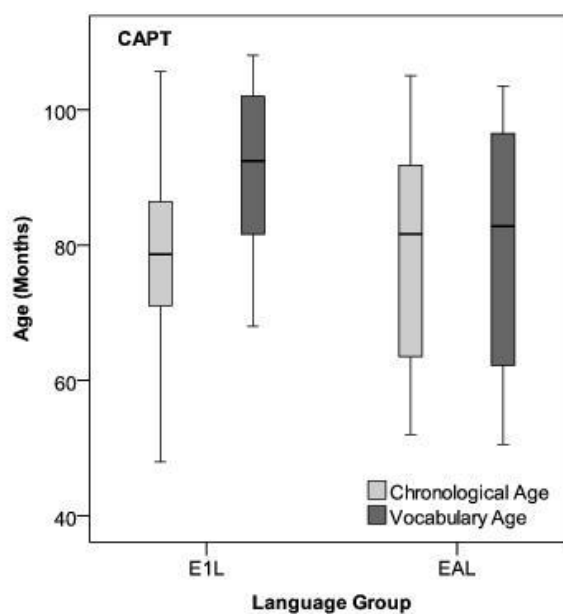
421

422  
423  
424  
425  
426  
427  
428  
429  
430  
431  
432



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

438  
439  
440  
441  
442  
443  
444  
445  
446  
447  
448



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

450

451 Figure 5 shows a scatter plot of performance on the CCT versus age for the 17  
 452 children whose scores were not excluded. There were significant correlations between  $d'$   
 453 scores and both chronological age ( $r = 0.68, n = 17, p < 0.002$ ; the equation for the  
 454 relationship was  $d' = 0.028\text{age} + 0.50$ , where age is in months) and vocabulary age ( $r = 0.64, n$   
 455  $= 17, p = 0.006$ ; the equation for the relationship was  $d' = 0.026\text{age} + 0.64$ ).  
 456

457

458

459

460

461

462

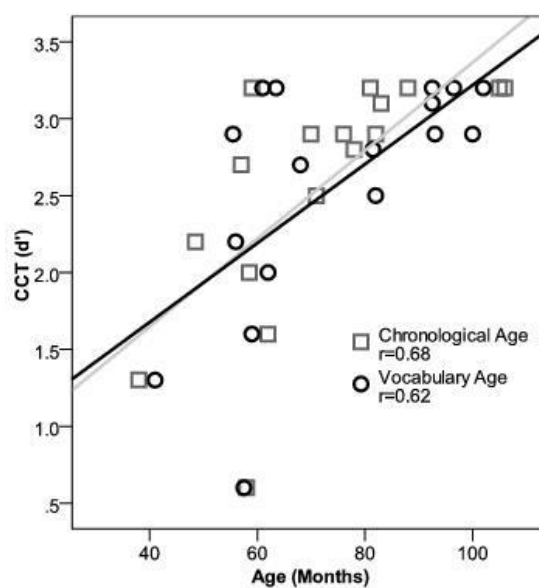
463

464

465

466

467



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

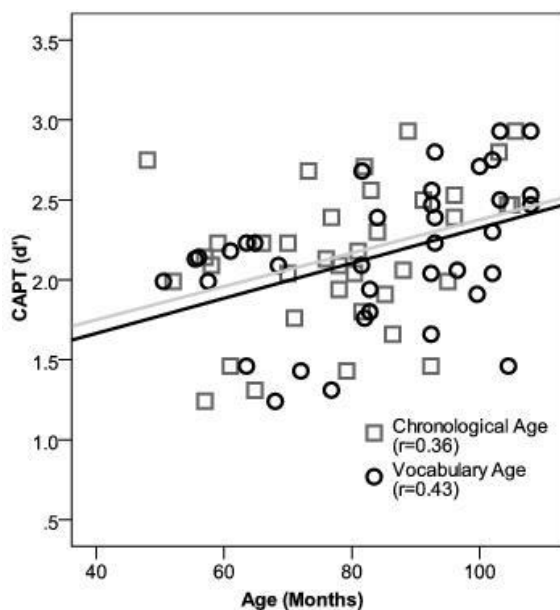
471

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

477

478

479



480

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

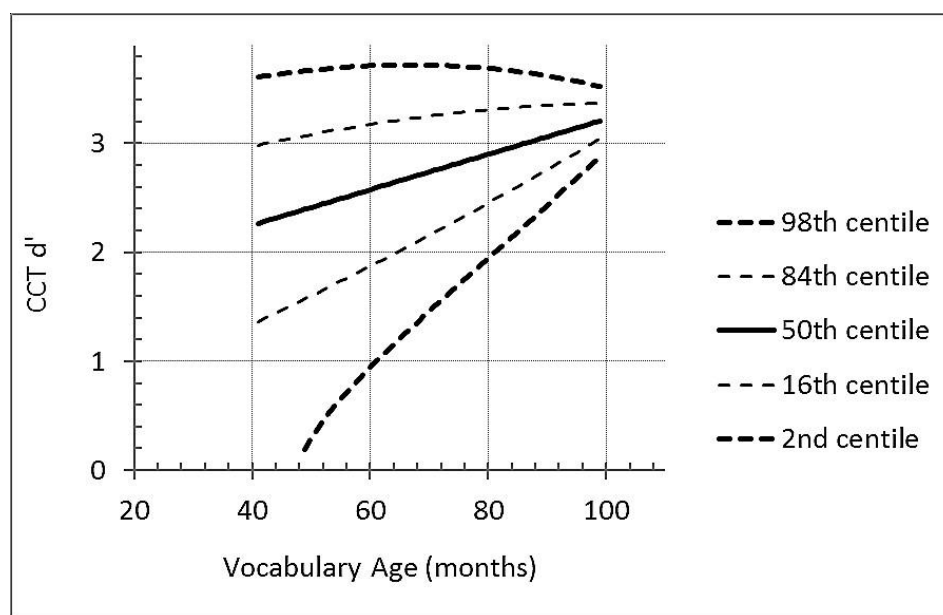
482

483 For both the CCT and the CAPT the correlation was similar for vocabulary and for  
484 chronological age. For the CCT the relationship with age accounted for approximately 43%  
485 of the variance in the scores (chronological = 46%; vocabulary age = 41%) whereas for the  
486 CAPT the relationship with age only accounted for approximately 15% of the variance  
487 (chronological age = 13%; vocabulary age = 19%). This finding of a lower strength of the  
488 relationship between age and score for the CAPT than for the CCT occurred partly because  
489 the vocabulary level requirement for the CAPT prevented very young children from taking  
490 the test, so the spread of ages was larger for the CCT. For the CCT the age range of the  
491 children tested was from 38 to 107 months and performance ranged from 32.5 to 100%. This  
492 enabled a rough estimate of the appropriate age range of the test to be derived based on the  
493 regression equation relating vocabulary age in months to  $d'$  score. Assuming that for a test to  
494 be sensitive to change the child should score between 60 and 85% ( $d'$  values of 1.15 and 2.14,  
495 respectively), the appropriate age range for the CCT is approximately 23 to 59 months. This  
496 estimate is approximate because it involves some extrapolation for the lower age limit. If a  
497 child scores at the upper limit of the CCT test, the child should be tested with the CAPT if it

498 is desired to track changes in performance over time or to assess the effect of an intervention.

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

507



508

509 **Fig. 7. Percentiles for CCT score as a function of vocabulary age, generated using the**  
 510 **LMS method. The 16% and 84% percentiles represent – 1 SD and + 1 SD, respectively.**  
 511

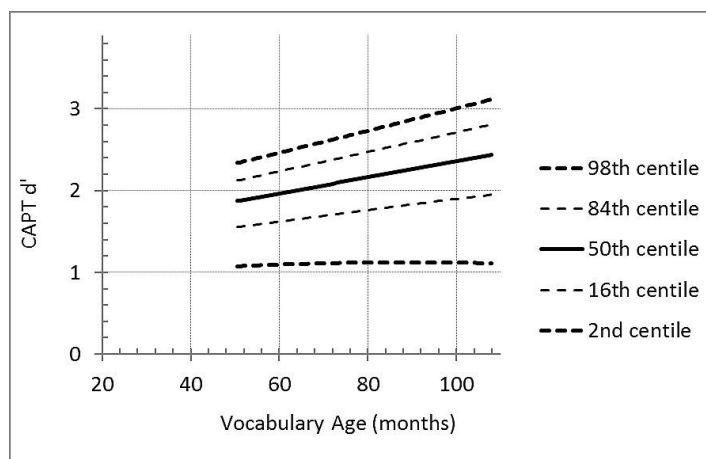
512 The relationship between developmental age and performance in the CCT is shown in  
 513 figure 7, and this figure should be used to determine if a child’s performance is within 1  
 514 standard deviation (16<sup>th</sup> and 84<sup>th</sup> percentile) of the mean for their vocabulary age. Once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

516 testing with the CAPT.

517 The relationship with developmental age was weaker for the CAPT than for the CCT;

518 the percentiles are shown in figure 8. There was a smaller range of performance than for the

519 CCT.



520

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

522

523

## DISCUSSION

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

534 The CCT was developed to have vocabulary items that are appropriate for children  
535 from two years old, while the CAPT was developed to be appropriate for slightly older

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

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

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

563 The analysis of the relationship between chronological and vocabulary age and  
564 performance showed a significant correlation for both tests, for both chronological and



565 vocabulary age. The correlations were higher for the CCT, probably because the vocabulary  
566 for the CCT was more challenging for the young children and because younger children  
567 could not be tested using the CAPT as their vocabulary was inadequate. The correlation of  
568 performance with age on both tests is consistent with previous results showing that the ability  
569 to understand speech improves with increasing age up to the early teens (Stelmachowicz et al.  
570 2000; Vance et al. 2009). The improvement presumably reflects the combined effects of  
571 maturation of auditory and cognitive skills and greater experience of the language.

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

594

595

## CONCLUSIONS

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

## ACKNOWLEDGEMENTS

613

614

615

616

617

618

## REFERENCES

619

620

621

622

- 623 Archbold, S., Lutman, M. E., & Marshall, D. H. (1995). Categories of auditory performance.  
624 *Ann Otol Rhinol Laryngol*, *166*, 312-314.
- 625 Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between  
626 two methods of clinical measurement. *Lancet*, *1*, 307-310.
- 627 Boothroyd, A., Eisenberg, L. S., & Martinez, A. (2005). *OLIMSPAC. Version 3.1d*. Los  
628 Angeles, CA: House Ear Institute.
- 629 Byrne, D., Dillon, H., Ching, T., et al. (2001). NAL-NL1 procedure for fitting nonlinear  
630 hearing aids: characteristics and comparisons with other procedures. *J Am Acad Audiol*,  
631 *12*, 37-51.
- 632 Cole, T. J., & Green, P. J. (1992). Smoothing reference centile curves: the LMS method and  
633 penalized likelihood. *Stat Med*, *11*, 1305-1319.
- 634 Cornelisse, L. E., Seewald, R. C., & Jamieson, D. G. (1995). The input/output formula: A  
635 theoretical approach to the fitting of personal amplification devices. *J Acoust Soc Am*,  
636 *97*, 1854-1864.
- 637 Cullington, H., Bele, D., Brinton, J., et al. (2013). United Kingdom national paediatric  
638 bilateral cochlear implant audit: preliminary results. *Cochlear Implants Int*, *14 Suppl 4*,  
639 S22-26.
- 640 Govaerts, P. J., Daemers, K., Yperman, M., et al. (2006). Auditory speech sounds evaluation  
641 (A $\S$ E $\text{\textcircled{R}}$ ): a new test to assess detection, discrimination and identification in hearing  
642 impairment. *Cochlear Implants Int*, *7*, 92-106.
- 643 Green, P. (1987). Penalized likelihood for general semi-parametric regression models. *Int*  
644 *Stat Rev*, *55*, 245-259.
- 645 Hacker, M. J., & Ratcliff, R. (1979). A revised table of d' for M-alternative forced choice.  
646 *Percept Psychophys*, *26*, 168-170.
- 647 Holt, R. F., & Lalonde, K. (2012). Assessing toddlers' speech-sound discrimination. *Int J*  
648 *Pediatr Otorhinolaryngol*, *76*, 680-692.
- 649 Jerger, S., Lewis, S., Hawkins, J., et al. (1980). Pediatric speech intelligibility test. I.  
650 Generation of test materials. *Int J Pediatr Otorhinolaryngol*, *2*, 217-230.
- 651 Kirk, U. (Ed.). (2012). *Neuropsychology of Language, Reading and Spelling*. New York:  
652 Academic Press.
- 653 Lovett, R., Summerfield, Q., & Vickers, D. (2013). Test-retest reliability of the Toy  
654 Discrimination Test with a masker of noise or babble in children with hearing  
655 impairment. *Int J Audiol*, *52*, 377-384.

- 656 Lovett, R. E., Vickers, D. A., & Summerfield, A. Q. (2015). Bilateral cochlear implantation  
657 for hearing-impaired children: criterion of candidacy derived from an observational  
658 study. *Ear Hear*, *36*, 14-23.
- 659 Macmillan, N. A., & Creelman, C. D. (2005). *Detection Theory: A User's Guide, 2nd Ed.*  
660 New York: Erlbaum.
- 661 Marriage, J. E., & Moore, B. C. J. (2003). New speech tests reveal benefit of wide-dynamic-  
662 range, fast-acting compression for consonant discrimination in children with moderate  
663 to severe hearing loss. *Int J Audiol*, *42*, 418-425.
- 664 Marriage, J. E., Vickers, D. A., Baer, T., et al. (2017). Comparison of different hearing aid  
665 prescriptions for children. *Ear Hear*, (submitted).
- 666 McConkey Robbins, A., Koch, D. B., Osberger, M. J., et al. (2004). Effect of age at cochlear  
667 implantation on auditory skill development in infants and toddlers. *Arch Otolaryngol*  
668 *Head Neck Surg*, *130*, 570-574.
- 669 Moore, B. C. J., Stone, M. A., Füllgrabe, C., et al. (2008). Spectro-temporal characteristics of  
670 speech at high frequencies, and the potential for restoration of audibility to people with  
671 mild-to-moderate hearing loss. *Ear Hear*, *29*, 907-922.
- 672 Reed, M. (1959). A verbal screening test of hearing. In *Proceedings of III World Congress of*  
673 *the Deaf*. Wiesbaden, Germany: Deutscher Gehorlosen.
- 674 Renfrew, C. (1995). *Word Finding Vocabulary Test*. Oxford: Winslow Press.
- 675 Stiles, D. J., Bentler, R. A., & McGregor, K. K. (2012). The Speech Intelligibility Index and  
676 the pure-tone average as predictors of lexical ability in children fit with hearing aids. *J*  
677 *Speech Lang Hear Res*, *55*, 764-778.
- 678 Thornton, A. R., & Raffin, M. J. (1978). Speech-discrimination scores modeled as a binomial  
679 variable. *J Speech Hear Res*, *21*, 507-518.
- 680 Tomblin, J. B., Harrison, M., Ambrose, S. E., et al. (2015). Language outcomes in young  
681 children with mild to severe hearing loss. *Ear Hear*, *36 Suppl 1*, 76S-91S.
- 682 van Buuren, R. A., Festen, J. M., & Plomp, R. (1995). Evaluation of a wide range of  
683 amplitude-frequency responses for the hearing impaired. *J Speech Hear Res*, *38*, 211-  
684 221.
- 685 Vickers, D. A., Backus, B. C., Macdonald, N. K., et al. (2013). Using personal response  
686 systems to assess speech perception within the classroom: an approach to determine the  
687 efficacy of sound field amplification in primary school classrooms. *Ear Hear*, *34*, 491-  
688 502.

689 Zhou, H., Chen, Z., Shi, H., et al. (2013). Categories of auditory performance and speech  
690 intelligibility ratings of early-implanted children without speech training. *PLoS One*, 8,  
691 e53852.  
692  
693

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

700

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

702

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

709

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

711

712 Fig. 5. Scatter plots showing the relationship between age (in months) and  $d'$  score for the  
713 CCT. Grey squares indicate chronological age and dark circles indicate vocabulary age.

714 Pearson correlation coefficients are shown in brackets.

715

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

717

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

720

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

722