Who is Right?: A word identification in noise test for young children using minimal pair distracters

Samuel Evans[1] & Stuart Rosen[2]

[1] Department of Psychology, University of Westminster, 115 New Cavendish St, London, W1W 6UW, UK

[2] Department of Speech, Hearing & Phonetic Sciences, UCL, 2 Wakefield St., London, WC1N 1PF

Abstract (200 words)

Many children have difficulties understanding speech. Reliable tools are needed to identify them in order to provide appropriate interventions. At present, there are relatively few assessments that test for subtle impairments in speech perception that have normative data from UK children. Here we present a new test, which evaluates children's ability to identify target words in background noise by choosing between minimal pair alternatives that differ by a single articulatory phonetic feature. This new test of single word perception is (1) tailored to testing young children, (2) has minimal memory demands, (3) adapts to the child's ability and (4) does not require reading or verbal output. Although designed for young children, it is also readily applicable in adults. Here, we show that speech in noise abilities in this particular task develop rapidly through childhood until they reach maturity at around ten years of age. We make this test freely available, with normative data for listeners aged 4-25 years old, and hope that it will be useful to researchers and clinicians in the assessment of speech perceptual abilities in children with hearing impairments, Developmental Language Disorder (DLD), Dyslexia and Auditory Processing Disorder (APD).

Key words: Speech perception, children, noise, audiological testing

Introduction

Children with speech, language and hearing disorders are at a greater risk of poorer literacy (Anthony & Francis, 2005), psycho-social development (Kilpatrick et al., 2019) and long term prospects (Bryan et al., 2007). Deficits in speech perception, in addition to being a defining feature of hearing impairment and Auditory Processing Disorder (APD) (Moore et al., 2013), are associated with a number of developmental disorders, most notably dyslexia (Noordenbos & Serniclaes, 2015) and Developmental Language Disorder (DLD) (Ferguson et al., 2011). Developing robust methods to identify individuals with speech perceptual deficits is a first step towards better characterising and treating these disorders. At present, there are few tests of speech perception that assess subtle impairments in speech perception and that have appropriate normative data from UK children. Here, we make freely available such a test, which we envisage will be useful to researchers and clinicians in evaluating the perceptual abilities of young children.

Many children find understanding spoken language difficult. In some children, such as those with hearing impairments, these difficulties are obvious and affect perception in both ideal and adverse listening situations. Pure tone thresholds, although important, provide limited information on functional listening abilities (Houtgast & Festen, 2008) and tests of speech perception in noise provide arguably a more valid assessment of day-to-day listening in children (Leibold et al., 2019). Children with developmental language disorders can exhibit subtle speech perceptual deficits, which often only become apparent when listening is made more challenging (Moore et al., 2013; Ziegler et al., 2005). Indeed, deficits are not always readily apparent and are sometimes only found in a minority of individuals, or not at all (Messaoud-Galusi et al., 2011). This may reflect a lack of sensitivity of available

tests, an absence of a true speech perception deficit or significant heterogeneity in the individuals assigned to these groups - only further research will help to uncover which of these explanations is correct. This task is made more difficult by the high co-morbidity between developmental reading, language and auditory processing disorders (Bishop et al., 2016; Moore et al., 2013) and the paucity of tools for assessing speech perception in children. A wider range of speech perception tests are required to better characterise the speech perceptual abilities of hearing impaired children and to further our understanding of developmental language disorders.

Successful speech perception requires the integration of multiple co-varying acoustic features (Kluender & Alexander, 2010; Lisker, 1977). In natural speech, the multiplicity of available features helps to ensure that perception remains relatively robust to acoustic variation and degradation of the speech signal. Speech sounds that differ on the basis of fewer contrastive features are more highly confusable (Miller & Nicely, 1955). Children with language impairments tend to perform more poorly on tasks in which speech tokens differ minimally from one another (Zoubrinetzky et al., 2016) and less so in tasks involving natural speech tokens that differ on the basis of multiple acoustic cues (Coady et al., 2005). Speech perception tasks can also be made more challenging by manipulating extrinsic factors, such as the presence of competing noise. Competing sounds generate overlapping patterns of excitation in the auditory periphery that obscure salient acoustic cues, a phenomena referred to as energetic/modulation masking (Brungart, 2001; Stone et al., 2011). Additional, informational masking effects, those not explained by energetic and modulation masking, are thought to arise at more central, cognitive levels of processing (Shinn-Cunningham, 2008). This form of masking is most often

associated with competing speech and is attributable in part to the difficulty of separating out and attending to the correct speech stream. Difficulties attending to speech in both competing speech (Dole et al., 2012) and non-speech maskers (Ziegler et al., 2005, 2009) have been associated with developmental language disorders.

Unfortunately, there are relatively few tests of speech perception in noise designed specifically for children with normative data from the UK. Tests designed for children need to be made engaging and use appropriate linguistic materials. It is important that tests have normative data from the country in which they are used. Normative data from other English speaking countries is unlikely to be appropriate for use in the UK and can sometimes overestimate the prevalence of perceptual deficits (Dawes & Bishop, 2007). Tests such as the SCAN-C (Keith, 2000) have been adapted for use with British children (Dawes & Bishop, 2007). However, SCAN-C is arguably not ideal for testing children with language impairments as it requires them to repeat back heard words. Many children with language disorders have difficulty planning and producing speech (Bishop et al., 2016) and so tests that require a verbal response may underestimate their true abilities.

For the same reason, tests such as the FAAF that require children to read words (Foster & Haggard, 1987) and those using sentences (e.g. LISN-S, Cameron & Dillon, 2007) that place greater demands on auditory working memory and syntactic processing, may not always be appropriate. Sentence material may be particularly inappropriate given the evidence that sentence repetition in *quiet* appears to be a good way to diagnose DLD (Conti-Ramsden et al., 2001). Tests that use single, early acquired words and that require a non-verbal output response

provide a purer test of speech perceptual ability in those with speech and language impairments.

There are relatively few existing UK tests of single word perception that have a non-verbal output response. The Consonant Confusion Test (CCT) is suitable for very young children and requires them to identify a target word from 4 alternatives presented as pictures. However, in this test the alternatives differ by multiple phonemes, e.g. “cow, owl, house, mouse”, hence the degree of phonemic discrimination required in this task is relatively broad. The Chear Auditory Perception Test (CAPT) is appropriate for slightly older children and includes contrasts that require a finer level of discrimination. However, the normative data for both these tests are derived from presenting the words at an artificially low volume, used as a way of inducing variation in accuracy (Vickers et al., 2018). This is arguably a less ecologically valid approach, compared to using competing noise to bring accuracy ‘off ceiling’.

The McCormick Toy Test (Summerfield et al., 1994) combines phonemic discrimination with concurrent noise presentation. However, the phonemic contrasts between word alternatives are not always minimal (e.g., “man” vs. “lamb”). Vance et al. (2009) includes fine grained phonemic discriminations, such that many of the items differ on a single articulatory phonetic feature, with concurrent noise presentation. However, the use of a fixed rather than an adaptive noise level does not accommodate children performing at the extremes of accuracy. Indeed, this kind of variation in performance is more likely in heterogeneous samples like those with developmental language disorders.

Here, we present a new speech perception test, the Who is Right? (WiR?) test and associated normative data for UK children and young adults. In this computer administered task, the listeners identify a target spoken word from three spoken alternative utterances that are presented against a competing noise. Participants indicate their response non-verbally with a button press. To ensure maximum sensitivity in identifying subtle impairments of speech processing, these alternatives differ by a single articulatory phonetic feature, with the background noise level adjusted adaptively dependent on their trial to trial performance accuracy.

Methods & Materials

Test construction

The WiR consists of 42 trials, all of a similar form. On each trial, the listener is presented with a picture of a target word on a display screen and hears a male speaker produce the name of the target in quiet (see Figure 1). Below the picture of the target are three cartoon faces which then take turns to speak three utterances. These utterances are produced by a different, female voice, to prevent participants using an echoic memory trace to complete the task. The voices are presented against a background of speech-spectrum-shaped noise (see details below). Two of the utterances are non-word foils differing from the target in its initial consonant in a single feature of voicing, place or manner (with the two foils always differing in the contrast used). The other face produces the target. For example, when the target is “bed”, the foils are “med” (differing in manner) and “ped” (differing in voicing). The position of the target and two distracter foils are randomised from trial to trial. The listener’s task is to identify the face that produced the correct target word by clicking on that face using a mouse. A correct response results in the selected cartoon face

smiling, whereas an incorrect response results in the selected face frowning. Every test began with a presentation of 14 familiarisation items followed by 28 test items, with a random permutation of the items within each phase. All stimuli were presented over headphones at a fixed comfortable level of about 65 dB SPL (measured over the frequency range 100 Hz – 5 kHz).

Targets words were early-acquired monosyllabic words mainly of CVC structure (two targets are CVs), that could be presented in an unambiguous pictorial form and whose initial consonant could be altered by a single feature of voicing, manner or place, to create two non-word foils (see Supplementary Materials for full details). All items were early acquired words, and the test items had a mean age of acquisition of 4.0 years, ranging from 2.9 to 5.6 (sd = 0.67), as measured by Kuperman et al. (2012). For the test trials, the distracter foils comprised 14 manner change items, 21 place change items and 21 voicing change items, distributed over the 28 test trials (2 feature changes per target).

During the test, the signal-to-noise ratio (SNR) was varied adaptively using a two-down/one-up adaptive rule tracking 71% correct (Levitt, 1971), which means that the SNR increases after every error, and decreases after two consecutive correct responses. The starting SNR was 20 dB, with a step-size of 7 dB which decreased by 1 dB after every track reversal until it reached 3 dB, at which value it remained for the rest of the test. The Speech Reception Threshold (SRT), here defined as the SNR that led to about 71% correct responses, was calculated from the mean of the track reversals in the test phase. Note that lower values indicate better performance, as this indicates that the listener can tolerate poorer SNRs for the desired accuracy.

The stimulus triplets differed greatly in inherent intelligibility, as would be expected by their variety of acoustic, phonetic and psycholinguistic properties. This is highly undesirable in adaptive testing because it leads to greater variability in the adaptive track. Extensive prior testing on dozens of school-age children allowed the determination of the psychometric functions (relating proportion correct to SNR) for each individual triplet. SRTs were then derived from these functions (through logistic regression) allowing the calculation of a correction factor (the deviation for each triplet from the mean SRT for all triplets) that was applied to the nominal SNR desired during each test. In this way, performance should be similar for all triplets at the same nominal SNR, which leads to more stable estimates of the SRTs.

The foil triplets were presented against a background of speech-spectrum-shaped shaped noise, synthesised to approximate the long-term average speech spectrum for combined male and female voices as estimated from the study of Byrne et al., (1994). This consisted of a low-frequency portion rolling off below 120 Hz at 17.5 dB/octave, and a high-frequency portion rolling off at 7.2 dB/octave above 420 Hz, with a constant spectrum portion in-between. The noise started 450 ms before the utterance triplet and finished 250 ms after, running continuously through the three utterances, with 50 ms rise and fall times. The test, including all materials, and analyses presented in this article are available here: https://github.com/drstuartrosen/WhoIsRight.

Participants

Ethical approval was granted by the UCL Research Ethics Committee. Informed written consent was received from all participants, and their parents, for those aged less than 16 years. None of the children or adults tested had any known

speech, hearing or language impairments and they were all native British English speakers.

The children and young adults were tested in primary and secondary schools in four separate rounds of testing – referred to as SC (n = 30), GY (n =17), RL (n = 54) and HR (n =17) – and were combined in the analysis. In all instances, testing took place in a quiet room either within school, home or in a quiet, distraction free public space, in Southern England. One round of testing (GY) arose from control data from typically developing children as part of a broader study of developmental language disorder (Baird et al., 2011; Loucas et al., 2016).

There were 118 participants who completed the test and for whom there was complete demographic information (mean age = 12.7 years, ranging from 4.0 to 25.1, with a s.d. = 4.63). Gender was well balanced with 53 males and 65 females (55%).

Results

The mean over the reversals in the test phase of the adaptive track was used to estimate a *Speech Reception Threshold* (SRT) for each participant. A plot of the obtained data against age shows a strong developmental trend of improving SRTs up to about age 10 with a significant levelling off after that point. This also suggested that the SRTs from the SC group that mainly included older participants were on average better than the other groups for participants of a similar age. There was one clear outlying low SRT (z = -3) which was deleted.

On the basis of the evidence that SRTs did not improve after age 11, boxplots were made of the SRTs from the 4 studies for all listeners greater than that

age (Figure 1). A one-way ANOVA with a follow-up Tukey post-hoc test confirmed the observation that the mean SRTs were not the same across the 4 testing groups (f (3, 77) = 9.203, p = 2.825e-05). The SRTs for SC were significantly different from RL and GY (both adjusted *p*s < 0.003), but SC and HR were not significantly different from each other (p = 0.152) even though the absolute difference in means was very similar to the other two groups, which did differ.

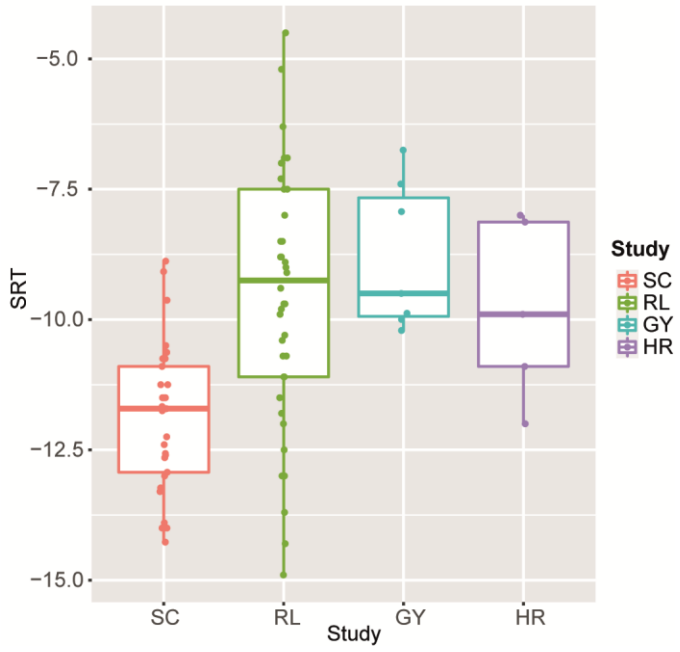

Figure 1: SRTs for children aged 11 years and above in 4 different testing rounds, illustrating the lower SRTs in the SC study.

It is not clear why SRTs were lower in this group and we assume that this reflects random sampling error. As SC only had participants aged 11.6-16.5 years (in secondary school), it seemed undesirable to leave the SRTs as they were, because the overall effect on model fits would not be equal across the age range. Therefore, all SRTs in the SC study were adjusted by the mean difference between the SRTs in that study and the three other studies for children ≥ 11 years old only

(i.e., by 2.4 dB). A one-way ANOVA confirmed that there was no evidence for differences across the groups after the adjustment ($f(3,77) = 0.301$, $p = 0.825$).

On the evidence that SRTs change up to about age 10, and then asymptote, two different models were used to fit the data. One was a segmented, or broken stick regression, in which the model consists of two straight lines which meet at a breakpoint. In this fit, a model in which the upper line had a slope=0 after the breakpoint (implying no change in SRTs after a particular age), was statistically indistinguishable from a model with non-zero slope for the upper segment. The estimated breakpoint was estimated at 9.9 years (95% CI = 8.6 – 11.2). Note that, for completeness, the data were also analysed without the adjustment accounting for the lower SRTs in the SC study and the findings were similar, with a breakpoint at age 10.4 years.

The other model was an asymptotic regression model with the equation:

$$SRT = b_1 + b_2 * \exp(b_3 * age)$$

where b_1 represents the asymptotic value (i.e., the lowest SRT reached through development), as long as $b_3<0$, which was indeed the case; b_3 controls how fast SRTs change over age, and b_2 scales the total range of this change. Note the important interaction between b_2 and b_3 in determining the shape of the curve, whereas b_1 is a simple additive term.

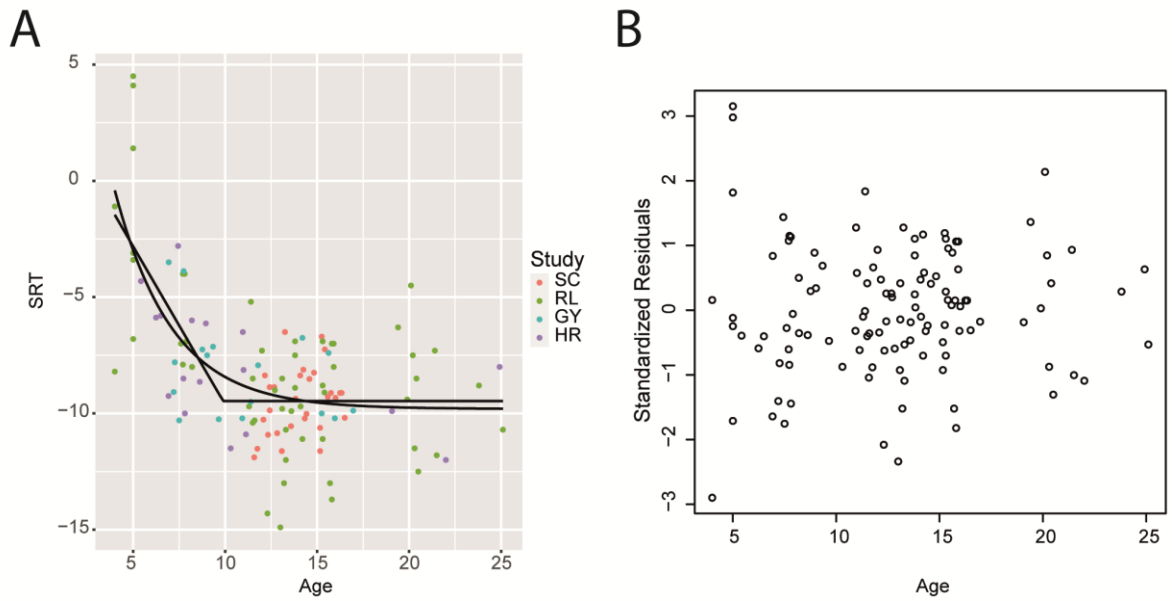

Figure 2: Regression models of SRT with age. A) SRT regressed against age with data sets indicated showing the broken stick and asymptotic regression fits. B) Standardized residuals of the broken stick regression.

The overall fits of the two models were very similar, as shown in Figure 2A, with the broken stick model a slightly better fit with a residual standard error of 2.34 compared to 2.40 on 114 degrees of freedom (as both models have the same number of estimated parameters). Visualisation of the standardised residuals against age for the broken stick regression indicated that variability in measurement of SRT was relatively constant across age after 5 years (Figure 2B).

As for many diagnostic tests, instead of expressing the outcome in a unit that a test directly manipulates (here, SNR in dB), it is often more useful to calculate a z-score, which reflects an individual's level of performance in comparison to their age-matched peers. This is straightforward to do based on the broken stick regression.

First, a predicted SRT must be calculated based on the listener's age, where:

If age ≤ 9.9, Predicted SRT = -1.36 x age + 3.98

If age > 9.9, Predicted SRT = - 9.5

Then, a residual is calculated by subtracting the predicted SRT from the actual SRT. This indicates by how many dB a listener is better or worse than an age-matched peer, with negative numbers again indicating better performance. This is then expressed as a z-score by dividing by an estimate of the standard deviation of the residuals (2.33). From the z-score, a percentile can be calculated.

Suppose, for example, that a child aged 6 obtained an SRT of -0.7 dB. The predicted SRT would be -4.2 dB from the equation above, which means this child is 3.5 dB worse than expected. Dividing through by 2.33 gives z = 1.5, which is to say, 1.5 standard deviations worse than typical 6 year olds. Only about 7% of children of that age would be expected to have an SRT this poor or worse. SRT values in dB and as z-scores relative to this normative data is provided as output for users of the test.

Discussion

We have presented normative data from UK children on a test of word identification in noise using minimal pair distracters. A broken stick regression showed that perceptual abilities on this task continued to improve rapidly until the age of around 10 years, before levelling out. We make this task and associated normative data freely available and hope that this test will be of use to researchers and clinicians in the assessment of speech perception abilities of children with

hearing, speech and language impairments. In the following sections, we discuss future developments and limitations of the task.

Native language speech sound representations are relatively well developed by 24 months of age but continue to be further refined well into later childhood (Kuhl, 2011). However, the point at which they achieve full maturity is still unknown. Changes are observed until at least six years of age (Nittrouer & Studdert-Kennedy, 1987; Nittrouer, 2002) with some studies showing that maturation continues beyond the early teens (Hazan & Barrett, 2000) and into the late teenage years (Davis et al., 2019; McMurray et al., 2018) . In the WiR? test, performance rapidly improves until around 10 years, before reaching a plateau. This break point is very similar to that obtained in a similar open response word recognition task in speech-spectrum-noise in a US sample (Corbin et al., 2016) and is broadly aligned with other studies showing rapid development of speech in noise abilities up until the age of ten for tasks involving competing energetic maskers (Hall et al., 2002; Leibold & Buss, 2013; Nishi et al., 2010; Wightman & Kistler, 2005). The earlier maturation on this task, compared to the tasks described above in which maturation continues into the late teenage years, may be attributed to important task differences. Our task requires participants to discriminate between canonical articulations with perceptual ambiguity arising from an extrinsic source, the presence of competing noise. By contrast, categorical perception paradigms require participants to categorise ambiguous sounds that are synthesised to be intermediate between canonical articulations. This may require a finer level of phonetic discrimination or place differing demands on decision making and executive function that gives rise to a different developmental trajectory.

The early plateau in energetic masking abilities stands in contrast to the more protracted development associated with informational masking. Indeed, children perform worse with speech on speech masking compared to equivalent non-speech maskers. Adult like performance on these tasks is not achieved until much later, often beyond 13 years of age (Corbin et al., 2016; Hall et al., 2002; Leibold & Buss, 2013). Future developments of the WiR could include the incorporation of competing speech maskers to provide sensitivity to the effects of information masking, especially in older children. In addition, the facility to spatially separate target and masking sounds, with associated normative data, may be a useful inclusion to identify those with APD that have a specific difficulty with spatial sound processing (Cameron & Dillon, 2007). Furthermore, given the different minimal pair contrasts available in WiR?, it may be possible to collect normative data on specific phonetic contrasts within the test. The ability to differentiate the contrasts that children find most difficult may provide a perspective on the mechanisms that underlie their speech perceptual weaknesses and allow better targeted interventions for children with hearing impairments. However, it is likely that such tests would require a fixed SNR, rather than an adaptive approach, with the SNR being fixed at a level appropriate for the listener. In this way, it could be assured that listeners would be not performing near floor or ceiling, but obtain intermediate levels of performance which would allow meaningful comparisons across contrast types.

The task in its current form also has limitations. At present we do not have a measure of re-test reliability or an understanding of how performance on the test changes with repetitive testing. We hope that re-test reliability would be relatively high given the efforts made to calibrate the task through the estimation of an SNR correction factor for each item. Visualisation of the standardised residuals of our

372 normative data showed that they are relatively uniformly distributed with few outliers
373 suggesting that the SRT measure is relatively stable across age. We anticipate that
374 learning in the task would be minimal both within a single test session and across
375 multiple tests due to the relatively large number of test words and the fact that they
376 are not repeated. Future work addressing re-test reliability and learning effects will
377 help to clarify whether our intuitions are correct.

378 Another limitation is in the normative data that we have acquired. Our full
379 sample was 117 participants, a sample size roughly in keeping with or better than
380 similar tests (Spyridakou et al., 2020; Vance et al., 2009; Vickers et al., 2018).
381 However, as with most tools of this kind, it would benefit from a larger normative
382 sample, from a broader demographic, as factors like social economic status have
383 been shown to influence speech perceptual ability (Nittrouer, 1996). Indeed, our
384 data was collected from only a small number of settings and likely represents a
385 relatively homogenous demographic sample. In future, normative data from a wider
386 demographic including hard to reach populations is necessary, taking into account
387 the additional time and resources that this would entail (Bonevski et al., 2014). As
388 part of this widening inclusion, it would also be beneficial to consider stratifying by
389 UK region to account for differences in regional accent (Adank et al., 2009). We
390 would hope to address these limitations in the future and to allow others to do so, by
391 making this test freely available, so that others can extend upon our initial work.

392 **Acknowledgements**

393 We are very grateful to Gillian Baird and Vicky Slonims for providing the GY data
394 from their much larger study. More importantly, it was the desire expressed by their
395 team for an appropriate test of speech-in-noise to be used in a study of children with

language disorders that was the prime impetus for the development of WiR? The rest of the data was extracted from the BSc theses of Sarah Cooper, Rebecca Lancaster and Henrietta Louise Roe, to whom we are also grateful. Samuel Evans was funded by a Vacation Scholarship from the charity Deafness Research UK, which was merged with the RNID (Royal National Institute for Deaf People).

Declaration of interests

None

References

Adank, P., Evans, B. G., Stuart-Smith, J., & Scott, S. K. (2009). Comprehension of Familiar and Unfamiliar Native Accents Under Adverse Listening Conditions. *Journal of Experimental Psychology: Human Perception and Performance*, *35*(2), 520–529. https://doi.org/10.1037/a0013552

Anthony, J., & Francis, D. (2005). Development of Phonological Awareness skill. *Current Directions in Psychological Science*, *14*(5), 255–259. http://www.speechpathology.com/articles/article_detail.asp?article_id=342

Baird, G., Slonims, V., Simonoff, E., & Dworzynski, K. (2011). Impairment in non-word repetition: A marker for language impairment or reading impairment? *Developmental Medicine and Child Neurology*, *53*(8), 711–716. https://doi.org/10.1111/j.1469-8749.2011.03936.x

Bishop, D. V. M., Snowling, M. J., Thompson, P. A., Greenhalgh, T., Adams, C., Archibald, L., Baird, G., Bauer, A., Bellair, J., Boyle, C., Brownlie, E., Carter, G., Clark, B., Clegg, J., Cohen, N., Conti-Ramsden, G., Dockrell, J., Dunn, J., Ebbels, S., … Whitehouse, A. (2016). CATALISE: A multinational and multidisciplinary Delphi consensus study. Identifying language impairments in children. *PLoS ONE*, *11*(7), 1–26. https://doi.org/10.1371/journal.pone.0158753

Bonevski, B., Randell, M., Paul, C., Chapman, K., Twyman, L., Bryant, J., Brozek, I., & Hughes, C. (2014). Reaching the hard-to-reach: A systematic review of strategies for improving health and medical research with socially disadvantaged groups. *BMC Medical Research Methodology*, *14*(1), 1–29. https://doi.org/10.1186/1471-2288-14-42

Brungart, D. S. (2001). Informational and energetic masking effects in the perception of two simultaneous talkers. *Journal of the Acoustical Society of America*, *109*(3), 1101–1109. wos:000167369300023

Bryan, K., Freer, J., & Furlong, C. (2007). Language and communication difficulties in juvenile offenders. *International Journal of Language and Communication Disorders*, *42*(5), 505–520. https://doi.org/10.1080/13682820601053977

Cameron, S., & Dillon, H. (2007). Development of the listening in spatialized noise-sentences test (LISN-S). *Ear and Hearing*, *28*(2), 196–211. https://doi.org/10.1097/AUD.0b013e318031267f

Coady, J. A., Kluender, K. R., & Evans, J. L. (2005). Categorical perception of speech by children with specific language impairments. *Journal of Speech, Language, and Hearing Research*, *48*(4), 944–959. https://doi.org/10.1044/1092-4388(2005/065)

Conti-Ramsden, G., Botting, N., & Faragher, B. (2001). Psycholinguistic markers for specific language impairment (SLI). *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *42*(6), 741–748. https://doi.org/10.1111/1469-7610.00770

Corbin, N. E., Bonino, A. Y., Buss, E., & Leibold, L. J. (2016). Development of open-set word recognition in children: Speech-shaped noise and two-talker speech maskers. *Ear and Hearing*, *37*(1), 55–63. https://doi.org/10.1097/AUD.0000000000000201

Davis, M., Evans, S., McCarthy, K., Evans, L., Giannakopoulou, A., & Taylor, J. (2019). Lexical learning shapes the development of speech perception until late adolescence. *PsyArXiv*, 1–52. https://doi.org/10.31234/osf.io/ktsey

Dawes, P., & Bishop, D. V. M. (2007). The SCAN-C in testing for auditory processing disorder in a sample of British children. In *International Journal of Audiology* (Vol. 46, Issue 12). https://doi.org/10.1080/14992020701545906

Dole, M., Hoen, M., & Meunier, F. (2012). Speech-in-noise perception deficit in adults with dyslexia: Effects of background type and listening configuration. *Neuropsychologia*, *50*(7), 1543–1552. http://www.sciencedirect.com/science/article/pii/S0028393212001200

Ferguson, M. A., Hall, R. L., Riley, A., & Moore, D. R. (2011). Communication, listening, cognitive and speech perception skills in children with auditory processing disorder (APD) or specific language impairment (SLI). *Journal of Speech, Language, and Hearing Research*, *54*(1), 211–227. https://doi.org/10.1044/1092-4388(2010/09-0167)

Foster, J. R., & Haggard, M. P. (1987). The Four Alternative Auditory Feature test (FAAF)-linguistic and psychometric properties of the material with normative data in noise. *British Journal of Audiology*, *21*(3), 165–174. https://doi.org/10.3109/03005368709076402

Hall, J. W., Grose, J. H., Buss, E., & Dev, M. B. (2002). Spondee recognition in a two-talker masker and a speech-shaped noise masker in adults and children. *Ear and Hearing*, *23*(2), 159–165. https://doi.org/10.1097/00003446-200204000-00008

Hazan, V., & Barrett, S. (2000). The development of phonemic categorization in children aged 6-12. *Journal of Phonetics*, *28*(4), 377–396.

https://doi.org/10.1006/jpho.2000.0121

Houtgast, T., & Festen, J. M. (2008). On the auditory and cognitive functions that may explain an individual's elevation of the speech reception threshold in noise. *International Journal of Audiology*, *47*(6), 287–295. https://doi.org/10.1080/14992020802127109

Keith, R. W. (2000). Development and standardization of SCAN-C test for auditory processing disorders in children. *Journal of the American Academy of Audiology*, *11*(8), 438–445.

Kilpatrick, T., Leitão, S., & Boyes, M. (2019). Mental health in adolescents with a history of developmental language disorder: The moderating effect of bullying victimisation. *Autism and Developmental Language Impairments*, *4*. https://doi.org/10.1177/2396941519893313

Kluender, K. R., & Alexander, J. M. (2010). Perception of Speech Sounds. In A. I. Basbaum, A. Kaneko, G. M. Shephard, G. Westheimer, T. D. Albright, R. H. Masland, P. Dallos, D. Oertel, S. Firestein, G. K. Beauchamp, C. Bushnell, J. H. Kass, & E. Gardner (Eds.), *The Senses: A Comprehensive Reference*. Elsevier.

Kuhl, P. (2011). Brain Mechanisms in Early Language Acquisition. *Neuron*, *67*(5), 713–727.

Kuperman, V., Stadthagen-Gonzalez, H., & Brysbaert, M. (2012). Age-of-acquisition ratings for 30,000 English words. *Behavior Research Methods*, *44*(4), 978–990. https://doi.org/10.3758/s13428-012-0210-4

Leibold, L. J., & Buss, E. (2013). Children's identification of consonants in a speech-shaped noise or a two-talker masker. *Journal of Speech, Language, and Hearing Research*, *56*(4), 1144–1155. https://doi.org/10.1044/1092-4388(2012/12-0011)

Leibold, L. J., Buss, E., & Calandruccio, L. (2019). Too Young for the Cocktail Party? *Acoustics Today*, *15*(1), 37. https://doi.org/10.1121/at.2019.15.1.39

Lisker, L. (1977). Rapid Versus Rabid - Catalog of Acoustic Features That May Cue Distinction. *Journal of the Acoustical Society of America*, *62*, S77–S78. isi:A1977EA29000377

Loucas, T., Baird, G., Simonoff, E., & Slonims, V. (2016). Phonological processing in children with specific language impairment with and without reading difficulties. *International Journal of Language and Communication Disorders*, *51*(5), 581–588. https://doi.org/10.1111/1460-6984.12225

McMurray, B., Danelz, A., Rigler, H., & Seedorff, M. (2018). Speech categorization develops slowly through adolescence. *Developmental Psychology*, *54*(8), 1472–1491. https://doi.org/10.1037/dev0000542

Messaoud-Galusi, S., Hazan, V., & Rosen, S. (2011). Investigating speech perception in children with dyslexia: Is there evidence of a consistent deficit in individuals? *Journal of Speech, Language, and Hearing Research*, *54*(6), 1682–1701. https://doi.org/10.1044/1092-4388(2011/09-0261)

514 Miller, G., & Nicely, P. (1955). An analysis of perceptual confusions among some
515 English consonant. *Journal of the Acoustical Society of America*, *27*, 338–352.
516 https://doi.org/10.1121/1.1907526

517 Moore, D. R., Rosen, S., Bamiou, D. E., Campbell, N. G., Sirimanna, T., James
518 Bellis, T., Chermak, G., Weihing, J., Musiek, F., Dillon, H., & Cameron, S.
519 (2013). Evolving concepts of developmental auditory processing disorder (APD):
520 A British Society of Audiology APD Special Interest Group "white paper."
521 *International Journal of Audiology*, *52*(1), 3–13.
522 https://doi.org/10.3109/14992027.2012.723143

523 Nishi, K., Lewis, D. E., Hoover, B. M., Choi, S., & Stelmachowicz, P. G. (2010).
524 Children's recognition of American English consonants in noise. *The Journal of*
525 *the Acoustical Society of America*, *127*(5), 3177–3188.
526 https://doi.org/10.1121/1.3377080

527 Nittrouer, S., & Studdert-Kennedy, M. (1987). The role of coarticulatory effects in the
528 perception of fricatives by children and adults. *Journal of Speech and Hearing*
529 *Research*, *30*(3), 319–329. https://doi.org/10.1044/jshr.3003.319

530 Nittrouer, Susan. (1996). The relation between speech perception and phonemic
531 awareness: Evidence from low-SES children and children with chronic OM.
532 *Journal of Speech, Language, and Hearing Research*, *39*(5), 1059–1070.
533 https://doi.org/10.1044/jshr.3905.1059

534 Nittrouer, Susan. (2002). Learning to perceive speech: How fricative perception
535 changes, and how it stays the same. *The Journal of the Acoustical Society of*
536 *America*, *112*(2), 711–719. https://doi.org/10.1121/1.1496082

537 Noordenbos, M. W., & Serniclaes, W. (2015). The Categorical Perception Deficit in
538 Dyslexia: A Meta-Analysis. *Scientific Studies of Reading*, *19*(5), 340–359.
539 https://doi.org/10.1080/10888438.2015.1052455

540 Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends*
541 *in Cognitive Sciences*, *12*(5), 182–186. wos:000256750400006

542 Spyridakou, C., Rosen, S., Dritsakis, G., & Bamiou, D. E. (2020). Adult normative
543 data for the speech in babble (SiB) test. *International Journal of Audiology*,
544 *59*(1), 33–38. https://doi.org/10.1080/14992027.2019.1638526

545 Stone, M. a, Füllgrabe, C., Mackinnon, R. C., & Moore, B. C. J. (2011). The
546 importance for speech intelligibility of random fluctuations in "steady"
547 background noise. *The Journal of the Acoustical Society of America*, *130*(5),
548 2874–2881. https://doi.org/10.1121/1.3641371

549 Summerfield, Q., Palmer, A. R., Foster, J. R., Marshall, D. H., & Twomey, T. (1994).
550 Clinical evaluation and test-retest reliability of the IHR-McCormick Automated
551 Toy Discrimination Test. *British Journal of Audiology*, *28*(3), 165–179.
552 https://doi.org/10.3109/03005369409086564

553 Vance, M., Rosen, S., & Coleman, M. (2009). Assessing speech perception in young
554 children and relationships with language skills. *International Journal of*
555 *Audiology*, *48*(10), 708–717. https://doi.org/10.1080/14992020902930550

Vickers, D. A., Moore, B. C. J., Majeed, A., Stephenson, N., Alferaih, H., Baer, T., & Marriage, J. E. (2018). Closed-Set Speech Discrimination Tests for Assessing Young Children. *Ear and Hearing*, *39*(1), 32–41. https://doi.org/10.1097/AUD.0000000000000528

Wightman, F. L., & Kistler, D. J. (2005). Informational masking of speech in children: Effects of ipsilateral and contralateral distracters. *The Journal of the Acoustical Society of America*, *118*(5), 3164–3176. https://doi.org/10.1121/1.2082567

Ziegler, J. C., Pech-Georgel, C., George, F., Alario, F. X., & Lorenzi, C. (2005). Deficits in speech perception predict language learning impairment. *Proceedings of the National Academy of Sciences of the United States of America*, *102*(39), 14110–14115. http://www.pnas.org/content/102/39/14110.abstract

Ziegler, J. C., Pech-Georgel, C., George, F., & Lorenzi, C. (2009). Speech-perception-in-noise deficits in dyslexia. *Developmental Science*, *12*(5), 732–745. https://doi.org/10.1111/j.1467-7687.2009.00817.x

Zoubrinetzky, R., Collet, G., Serniclaes, W., Nguyen-Morel, M. A., & Valdois, S. (2016). Relationships between categorical perception of phonemes, phoneme awareness, and visual attention span in developmental dyslexia. *PLoS ONE*, *11*(3), 1–26. https://doi.org/10.1371/journal.pone.0151015

Supplementary materials

S1: Full List of targets and Foils for the familiarisation and testing phase. AoA = Age of Acquistion, SAM-PA = SAM-PA machine readable IPA transcription, Feature = Phonological feature change, SNR = SNR adjustment for each word.

	Target X-SAMPA	AoA	SNR	Foil 1 (distracter) X-SAMPA	Feature	Foil 2 (distracter) X-SAMPA	Feature
Familiarisation							
Bike	baIk	4.79	2	waIk	Manner	gaIk	Place
Bin	bIn	4.68	3	mIn	Manner	gIn	Place
Bus	bVs	3.85	-4	wVs	Manner	dVs	Place
Dog	dQg	2.80	2	nQg	Manner	gQg	Place
Doll	dQl	3.68	0	rQl	Manner	bQl	Place
Duck	dVk	3.50	-4	zVk	Manner	gVk	Place
Laugh	lA:f	3.79	-2	zA:f	Manner	wA:f	Place
Leg	leg	3.00	-1	deg	Manner	jeg	Place
One	wVn	3.23	-3	mVn	Manner	lVn	Place
Rain	reIn	3.60	0	neIn	Manner	jeIn	Place

Sea	si:	4.74	-7	zi:	Voicing	Ti:	Place
Sun	sVn	3.40	11	zVn	Voicing	TVn	Place
Watch	wQtS	4.33	-3	qQtS	Manner	rQtS	Place
Wave	weIv	4.26	-1	beIv	Manner	leIv	Place
Test items							
Bed	bed	2.89	-3	med	Manner	ped	Voicing
Book	bUk	3.68	0	wUk	Manner	pUk	Voicing
Boot	bu:t	3.89	5	wu:t	Manner	pu:t	Voicing
Chair	tSe@	3.43	0	se@	Manner	dZe@	Voicing
Boat	b@Ut	3.84	-1	w@Ut	Manner	p@Ut	Voicing
Bag	b{g	4.28	-3	m{g	Manner	p{g	Voicing
Dig	dig	4.19	-3	nIg	Manner	tIg	Voicing
Towel	taUI	3.22	-5	saUl	Manner	paUl	Place
Sing	sIN	3.47	-13	tIN	Manner	SIN	Place
Knife	nalf	4.15	0	dalf	Manner	maIf	Place
Wash	wQS	4.00	-5	bQS	Manner	rQS	Place
Bath	bA:T	3.23	-4	wA:T	Manner	dA:T	Place
Leaf	li:f	4.60	2	ni:f	Manner	wi:f	Place
Road	r@Ud	4.55	-2	z@Ud	Manner	j@Ud	Place
Cough	kAf	4.32	18	pAf	Place	gAf	Voicing
Bite	balt	3.58	-5	dalt	Place	paIt	Voicing
Comb	k@Um	5.50	9	p@Um	Place	g@Um	Voicing
Kite	kalt	4.58	5	palt	Place	gaIt	Voicing
Cow	kaU	3.94	0	taU	Place	gaU	Voicing
Cake	kelk	3.26	3	peIk	Place	geIk	Voicing
Fish	fIS	4.05	1	hIS	Place	vIS	Voicing
Fork	fO:k	3.63	4	sO:k	Place	vO:k	Voicing
Five	s@Up	4.51	4	SaIv	Place	vaIv	Voicing
Fall	fO:l	4.71	0	sO:l	Place	vO:l	Voicing
Soap	s@Up	3.17	2	f@Up	Place	z@Up	Voicing
Foot	fUt	3.44	4	hUt	Place	vUt	Voicing
Suck	sVk	5.58	-8	hVk	Place	zVk	Voicing
Thumb	TVm	4.42	3	SVm	Place	DVm	Voicing

581