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

Corresponding author:

Dr Samuel Evans
Department of Psychology
University of Westminster
115 New Cavendish St
London
W1W 6UW
UK

S.Evans1@westminster.ac.uk

Abstract

Purpose: Many children have difficulties understanding speech. At present, there are few assessments that test for subtle impairments in speech perception with normative data from UK children. We present a new test that 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 task is (1) tailored to testing young children, but also readily applicable to adults, (2) has minimal memory demands, (3) adapts to the child's ability and (4) does not require reading or verbal output.

Method: We tested 155 children and young adults aged from 5 to 25 years of age on this new test of single word perception.

Results: Speech in noise abilities in this particular task develop rapidly through childhood until they reach maturity at around nine years of age.

Conclusions: We make this test freely available and provide associated normative data. We hope that it will be useful to researchers and clinicians in the assessment of speech perception abilities in children that are hard of hearing, have Developmental Language Disorder (DLD), dyslexia or Auditory Processing Disorder (APD).

Key words: Speech perception, development, noise, audiology, auditory processing disorder, dyslexia, hard of hearing, developmental language disorder

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 perception deficits is a first step towards better characterising and treating these disorders. At present, there are few tests 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 children that are hard of hearing, 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 often exhibit subtle speech perception deficits. However, 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 perception abilities of children who are hard of hearing 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 such as when categorising synthetic continua that differ on a single acoustic parameter (Collet et al., 2012; Zoubrinetzky et al., 2016). Deficits in these groups have been shown to be less pronounced in tasks involving natural speech tokens that differ on the basis of multiple acoustic cues (Blomert & Mitterer, 2004; 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 or destroy salient acoustic cues, phenomena referred to as energetic and/or modulation masking (Brungart, 2001; Stone et al., 2011). White noise and steady-state speech-spectrum-shaped noise (as used in this study) are expected to interfere with speech perception predominantly through masking of this type. 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.

Speech perception deficits are not always observed in children with developmental language disorders when tested in ideal listening conditions. Performance is often at ceiling and the addition of competing noise is needed to provide a perceptual stressor that more reliably reveals subtle perceptual deficits (Calcus et al., 2015, 2018; Inoue et al., 2011; Ziegler et al., 2005, 2009). These deficits have been observed in the context of both competing speech (Dole et al., 2012) and competing non-speech (Ziegler et al., 2005, 2009). Most frequently, deficits have been observed when participants are required to identify and categorise non-word syllables, suggesting a locus of deficit originating at the phonetic and/or phonemic levels (Calcus et al., 2015; Varnet et al., 2016; Ziegler et al., 2005, 2009). Studies have shown weaknesses discriminating specific kinds of phonetic contrasts in children with language impairment (Cornelissen et al., 1996; Ziegler et al., 2005, 2009). Results from these studies suggest that different language impairments might be associated with deficits in specific phonetic contrasts; for example, children with dyslexia have been shown to have greater difficulty with voicing contrasts whilst those with developmental language disorder have problems with place and manner (Ziegler et al., 2005, 2009). Some studies have also found evidence for generalised deficits, rather than difficulties for specific classes of phonetic contrasts (Calcus et al., 2015).

In typical development, the encoding in the auditory periphery of basic sound features matures early and is thought to be broadly complete by around 6 months of age (Leibold & Buss, 2019). Despite this early maturation, perception in noise abilities continue to mature over a long period. Adult-like perceptual ability does not emerge until 9-10 years of age for speech in steady-state speech-shaped noise (Nishi et al., 2010) and matures even later, around 13-14 years, for speech in speech masking (Corbin et al., 2016). This slow development likely reflects the maturation of central auditory and cognitive abilities that relate to sound segregation, dip-listening, selective attention, working memory and language skills (Leibold et al., 2019; Leibold & Buss, 2019). Young children are easily distracted by additional sound streams, even when the target and masker sounds do not overlap in frequency (Youngdahl et al., 2018). Over time, children learn to deal with distraction and begin to exploit the acoustic distinctions that adults use to improve speech in noise performance, such as spatial cues to location (Litovsky, 2005) and differences in pitch and speaker characteristics (Flaherty et al., 2019). Improvements in auditory abilities may also be underpinned by developments in vocabulary and working memory, which have been positively associated with differences in speech in noise abilities (McCreery et al., 2017), while noting that these associations have not always been observed (Nittrouer et al., 2013).

Charting the development of speech in noise ability in UK children is difficult as there are relatively few tests designed for children with normative data. 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, the 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). Children with language learning impairments such as DLD and dyslexia often have difficulties in reading, syntactic processing, working memory and vocabulary development (Cowan et al., 2017; Laws et al., 2015; Van Der Lely, 2005). Tests that use single, early acquired words and that require a non-verbal output response, allow better assessment of speech perception abilities (especially in young children and those with language learning impairments) as they minimise extraneous syntactic, vocabulary and working memory demands.

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 the same single 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 three utterances are produced by the same single female speaker. Note that the target voice presented in quiet and the voices that participants choose between are from different talkers, intentionally of different sex, so as to prevent participants using an echoic memory trace to perform the task. The voices are presented against a background of steady-state 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 utterance is 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 (over which a Speech Reception Threshold (SRT) was calculated), 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).

Target words were 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, S1, 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 SNR was adapted during both the familiarisation and test phase. The Speech Reception Threshold was defined as the SNR that led to about 71% correct responses, calculated from the mean of the track reversals during the test phase only. Note that lower values indicate better performance, as this indicates that the listener can tolerate poorer SNRs for the desired accuracy. Younger children (under age 9) took more time to complete the test, with a median completion time of about 7 minutes, but everyone older took only about 6 minutes.

Each test consisted of the same 42 trials (14 familiarisation and 28 test items) presented in a different order. The response options on each trial included the target word and the same two unique non-word distracter foils – a stimulus triplet. These stimulus triplets differed greatly in inherent intelligibility, as would be expected by their variety of acoustic, phonetic and psycholinguistic properties, not to mention the exact choice of foils as being an important determinant of performance. 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 (using a combination of adaptive and fixed-SNR testing) allowed the determination of the psychometric functions (relating proportion correct to SNR) for each individual triplet. SRTs for each word 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 (see the Supplementary Materials, S1). This correction factor was used in an additive way to adjust the SNR level up or down for each individual triplet/trial. 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 three response alternatives were presented against a background of speech-spectrum-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.

[Insert Figure 1 here]

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. These criteria were confirmed by the caregiver during the consent process.

The children and young adults were tested in primary and secondary schools in six separate rounds of testing – referred to as SC (n = 30), GY (n =17), RL (n = 54), HR (n = 17), HW (n = 18) and CR (n = 19) – 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, e.g. a room in a community centre. The majority of testing took place in Southern England. Participants for 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). Further details concerning the age composition and testing environment for each data set are described in supplementary materials, S2.

There were 155 participants who completed the test (with 2 exclusions during analysis) and for whom there was complete demographic information (following data exclusions: mean age = 11.7 years, ranging from 4.9 to 25.1, s.d. = 4.6). Gender was well balanced with 63 males and 73 females (54%). There was a mix of genders in all testing rounds. Due to tester error, there was no gender data retained for the CR group, but it was of mixed gender.

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. Listeners varied considerably in the total number of reversals that were obtained, from 4 – 15 (mean = 9.6), with 94% of the listeners having 7 or more reversals, and no difference on average between younger (under 9) and older listeners (within 0.06). There was also no relationship between the number of reversals and age or the SRT. Also of interest is the level of performance observed over the test phase of 28 trials, which should be near the targeted value of 71%. In fact, observed performance levels varied from 61% - 82% (mean= 70%) and 95% of listeners had levels within the range of 64 - 75%. Again, there was no difference on average between younger and older listeners (within 0.5%) and no relationship between performance and age or the SRT. In short, it appears that the adaptive procedure worked equally well across the age range, so any differences in SRT with age likely reflect genuine differences in ability to do the task.

A plot of the obtained data against age showed a strong developmental trend of improving SRTs up to about age 9 or 10, 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.

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 2). 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 ($f(3, 78) = 9.978$, $p = 1.22 \times 10^{-5}$). 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 = 0.086) even though the absolute difference in means was very similar to the other two groups, which did differ. This is likely due to the fact that there are only 5 older listeners in the HR group.

[Insert Figure 2 here]

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 (by 2.74 dB). A one-way ANOVA confirmed that there was no evidence for differences across the groups after the adjustment ($f(3, 78) = 0.256$, $p = 0.857$).

On the evidence that SRTs change up to about age 9 or 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. Two participants were removed from the data set as they contributed a residual with z-scores > 3. Once those points were excised, all other z-scores were within ± 3. 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 ($p > 0.4$). Also, the broken stick was a much better fit than that provided by a simple linear relationship of SRT with age ($p = 3.7 \times 10^{-12}$). The breakpoint was estimated at 9.2 years (95% CI = 8.3 – 10.2). 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 10.1 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.

[Insert Figure 3 here]

The overall fits of the two models were identical, as shown in Figure 3, with a residual standard error of 2.42 on 150 degrees of freedom (as both models have the same number of estimated parameters). We prefer the broken stick model because it gives an unambiguous estimated age for which performance in this task is adult-like. 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 4).

[Insert Figure 4 here]

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 ≤ 9.2, Predicted SRT = -1.64 x age + 5.57
If age > 9.2, Predicted SRT = - 9.6

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 (2.41). From the z-score, a percentile can be calculated.

Suppose, for example, that a child aged 6 years obtained an SRT of -0.6 dB. The predicted SRT would be -4.2 dB from the equation above, which means this child is 3.6 dB worse than expected. Dividing through by 2.41 gives $z \approx 1.5$, 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. The test software outputs SRT values in dB, with an option of an extra step to calculate z-scores based on specifying the listener's age.

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 9 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 language impairments and those that are hard of hearing. 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 9-10 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/modulation 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 (Davis et al., 2019; Hazan & Barrett, 2000; McMurray et al., 2018), 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 give 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, with adult-like performance on these tasks not achieved until much later, often beyond 13 years of age (Corbin et al., 2016; Hall et al., 2002; Leibold & Buss, 2013). There is also, albeit weak evidence, that SRTs for speech-on-speech masking are a better predictor than equivalent noise masking thresholds for the everyday listening challenges that children that are hard of hearing face (Hillock-Dunn et al., 2015). Such notions may make it seem desirable to implement our task with informational maskers like speech. At present there is not a speech-on-speech task for children that has normative data from UK children. Although it would be possible to construct such a task based on the WiR?, there seems little point to using such carefully constructed stimuli (with the emphasis on the perception of fine phonetic detail), in a version of the task in which higher order abilities like resistance to distraction and auditory scene analysis are important factors. An approach based on simple closed-set targets (e.g., as in Brungart, 2001) might be more appropriate in this instance.

What might be a more promising avenue for these materials, given the different minimal pair contrasts available in WiR?, is to collect normative data on the perception of specific phonetic contrasts. The ability to identify the contrasts that children find most difficult may provide a perspective on the mechanisms that underlie their speech perception weaknesses and allow better targeted interventions for children who are hard of hearing or have developmental language disorders. 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 a sufficient number of errors for 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 normative data show that they are relatively uniformly distributed with few outliers suggesting that the SRT measure is relatively stable across age. We anticipate that learning in the task would be minimal both within a single test session and across multiple sessions due to the relatively large number of test words and the fact that they are not repeated. Future work addressing re-test reliability and learning effects will help to clarify our intuitions. As part of that investigation, it would be useful to know whether it is better to take the first attempt or to average over multiple SRT estimates to attain a truer estimate of speech perception abilities. Indeed, there is some noticeable individual variation in SRT scores (around 5-10 dB range) and greater reliability might be attained by averaging over three measurements (cf. Calandruccio et al., 2020).

Another limitation is that we did not test the pure tone thresholds for our children and so do not have an objective measure of hearing thresholds for the children in our normative sample. However, all parents reported that their children were without hearing difficulties or speech and language impairments and we have no reason to think that our sample is unrepresentative of typically developing children. Our full sample (excluding outliers) was 153 participants, a sample size roughly in keeping with or larger than similar tests (Spyridakou et al., 2020; Vance et al., 2009; Vickers et al., 2018). As with most tools of this kind, it would benefit from a larger normative sample and from a broader demographic; factors like social economic status have been shown to influence speech perception ability (Nittrouer, 1996). Our data was collected from only a small number of settings and likely represents a relatively homogenous demographic sample. In future, normative data from a wider demographic including hard to reach populations is necessary, taking into account the additional time and resources that this would entail (Bonevski et al., 2014). As part of this widening inclusion, it would also be beneficial to consider stratifying by UK region to account for differences in regional accent (Adank et al., 2009).

Finally, these normative data apply to quiet listening environments, as might be found in a quiet room within a school or a community clinic. In the future, it would be useful to generate equivalent normative data from children tested in an audiological setting. We hope to address these limitations in the future and allow others to do so, by making this test freely available. We hope that the community will make use of and extend upon our initial work. Only further work will show whether it will be a useful tool in clarifying the speech perception difficulties experienced by listeners with various clinical disorders.

Acknowledgements

We are very grateful to Gillian Baird and Vicky Slonims for providing the GY data from their much larger study. More importantly, it was the desire expressed by their 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, Henrietta Louise Roe, and Hannah Williams and the PhD thesis of Csaba Redey-Nagy, to whom we are also grateful. Sam Eaton-Rosen also assisted by drawing some of the cartoon faces. 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, UK).

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

Blomert, L., & Mitterer, H. (2004). The fragile nature of the speech-perception deficit in dyslexia: Natural vs. synthetic speech. *Brain and Language*, *89*(1), 21–26. https://doi.org/10.1016/S0093-934X(03)00305-5

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

Calandruccio, L., Porter, H. L., Leibold, L. J., & Buss, E. (2020). The Clear-Speech Benefit for School-Age Children: Speech-in-Noise and Speech-in-Speech Recognition. *Journal of Speech, Language, and Hearing Research*, *63*(12), 4265–4276. https://doi.org/10.1044/2020_JSLHR-20-00353

Calcus, A., Colin, C., Deltenre, P., & Kolinsky, R. (2015). Informational masking of speech in dyslexic children. *The Journal of the Acoustical Society of America*, *137*(6), EL496–EL502. https://doi.org/10.1121/1.4922012

Calcus, A., Hoonhorst, I., Colin, C., Deltenre, P., & Kolinsky, R. (2018). The “Rowdy Classroom Problem’” in Children with Dyslexia: A Review.” In T. Lachmann & T. Weis (Eds.), *Reading and Dyslexia: From Basic Functions to Higher Order Cognition* (pp. 183–211). Springer International Publishing. https://doi.org/10.1007/978-3-319-90805-2_10

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)

Collet, G., Colin, C., Serniclaes, W., Hoonhorst, I., Markessis, E., Deltenre, P., & Leybaert, J. (2012). Effect of phonological training in French children with SLI: Perspectives on voicing identification, discrimination and categorical perception. *Research in Developmental Disabilities*, *33*(6), 1805–1818. https://doi.org/10.1016/j.ridd.2012.05.003

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

Cornelissen, P. L., Hansen, P. C., Bradley, L., & Stein, J. F. (1996). Analysis of perceptual confusions between nine sets of consonant-vowel sounds in normal and dyslexic adults. *Cognition*, *59*(3), 275–306. https://doi.org/10.1016/0010-0277(95)00697-4

Cowan, N., Hogan, T. P., Alt, M., Green, S., Cabbage, K. L., Brinkley, S., & Gray, S. (2017). Short-term Memory in Childhood Dyslexia: Deficient Serial Order in Multiple Modalities. *Dyslexia*, *23*(3), 209–233. https://doi.org/10.1002/dys.1557

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)

Flaherty, M. M., Buss, E., & Leibold, L. J. (2019). Developmental Effects in Children's Ability to Benefit From F0 Differences Between Target and Masker Speech. *Ear and Hearing*, *40*(4). https://journals.lww.com/ear-hearing/Fulltext/2019/07000/Developmental_Effects_in_Children_s_Ability_to.15.aspx

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

Hillock-Dunn, A., Taylor, C., Buss, E., & Leibold, L. J. (2015). Assessing Speech Perception in Children With Hearing Loss: What Conventional Clinical Tools May Miss. *Ear and Hearing*, *36*(2). https://journals.lww.com/ear-hearing/Fulltext/2015/03000/Assessing_Speech_Perception_in_Children_With.19.aspx

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

Inoue, T., Higashibara, F., Okazaki, S., & Maekawa, H. (2011). Speech perception in noise deficits in Japanese children with reading difficulties: Effects of presentation rate. *Research in Developmental Disabilities*, *32*(6), 2748–2757.

https://doi.org/10.1016/j.ridd.2011.05.035

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

Laws, G., Briscoe, J., Ang, S. Y., Brown, H., Hermena, E., & Kapikian, A. (2015). Receptive vocabulary and semantic knowledge in children with SLI and children with Down syndrome. *Child Neuropsychology*, *21*(4), 490–508. https://doi.org/10.1080/09297049.2014.917619

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. (2019). Masked Speech Recognition in School-Age Children. *Frontiers in Psychology*, *10*(September). https://doi.org/10.3389/fpsyg.2019.01981

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

Levitt, H. (1971). Transformed Up- Down Methods in Psychoacoustics. *The Journal of the Acoustical Society of America*, *49*(2B), 467–477. https://doi.org/10.1121/1.1912375

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

Litovsky, R. Y. (2005). Speech intelligibility and spatial release from masking in young children. *The Journal of the Acoustical Society of America*, *117*(5), 3091–3099. https://doi.org/10.1121/1.1873913

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

McCreery, R. W., Spratford, M., Kirby, B., & Brennan, M. (2017). Individual differences in language and working memory affect children's speech recognition in noise. *International Journal of Audiology*, *56*(5), 306–315. https://doi.org/10.1080/14992027.2016.1266703

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)

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

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

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

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

Nittrouer, S, Caldwell-Tarr, A., Tarr, E., Lowenstein, J., Rice, C., & Moberly, A. (2013). Improving speech-in-noise recognition for children with hearing loss: Potential effects of language abilities, binaural summation, and head shadow. *International Journal of Audiology*, *52*(8), 513–525. https://doi.org/10.3109/14992027.2013.792957

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

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

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

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

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

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

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

Van Der Lely, H. K. J. (2005). Domain-specific cognitive systems: Insight from Grammatical-SLI. *Trends in Cognitive Sciences*, *9*(2), 53–59. https://doi.org/10.1016/j.tics.2004.12.002

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

Varnet, L., Meunier, F., Trollé, G., & Hoen, M. (2016). Direct viewing of dyslexics' compensatory strategies in speech in noise using Auditory Classification Images. *PLoS ONE*, *11*(4), 1–17. https://doi.org/10.1371/journal.pone.0153781

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

Youngdahl, C., Healy, E., Yoho, S., Apoux, F., & Rachael, H. (2018). The Effect of Remote Masking on the Reception of Speech by Young School-Age Children. *Journal of Speech, Language, and Hearing Research*, *61*(2), 420–427. https://doi.org/10.1044/2017_JSLHR-H-17-0118

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

Figures & Legends

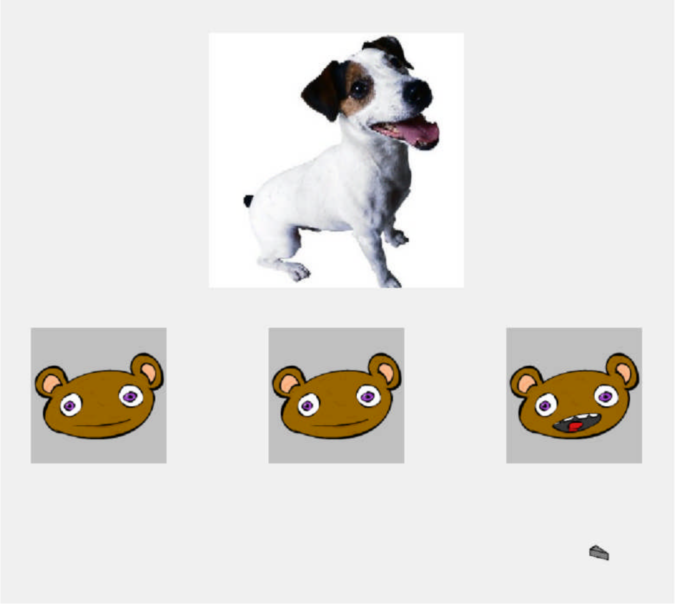

Figure 1: The WiR? task. On each trial, the listener sees a picture of a target word and hears the same single male speaker produce the name of the target in quiet. Below, three cartoon faces take turns to speak three utterances presented against a background of steady-state speech-spectrum-shaped noise. Two of the utterances are non-word foils differing from the target in a single phonetic feature. The other utterance is the target. Participants select the face that said the “right” word by clicking it with a mouse. A pie chart at bottom right displays the participant’s progress.

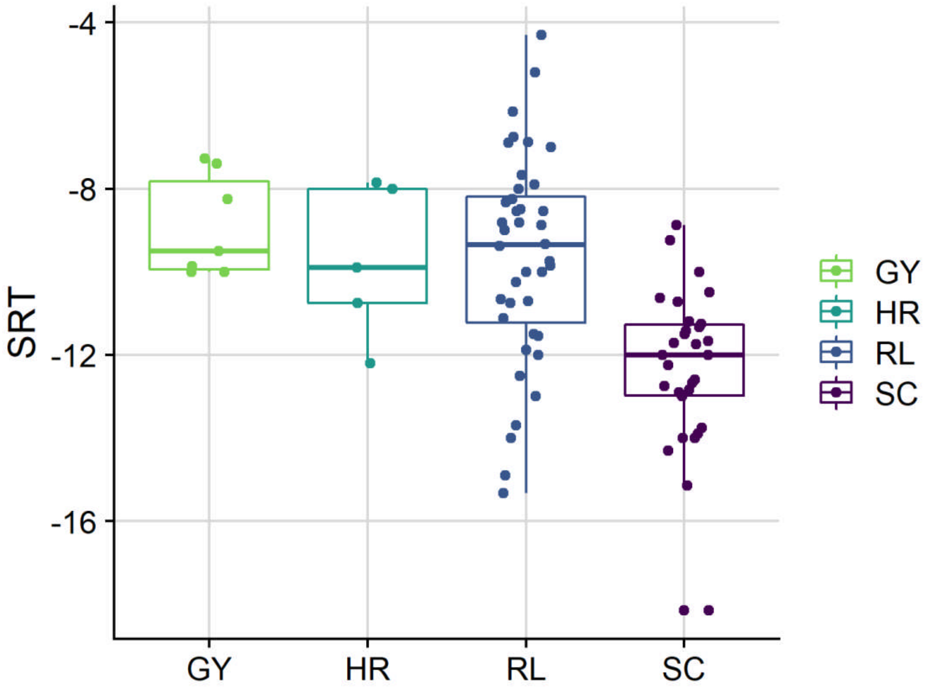

Figure 2: SRTs for children aged 11 years and above, illustrating lower SRTs in the SC study. The individual data points are jittered horizontally so as to minimise overlap.

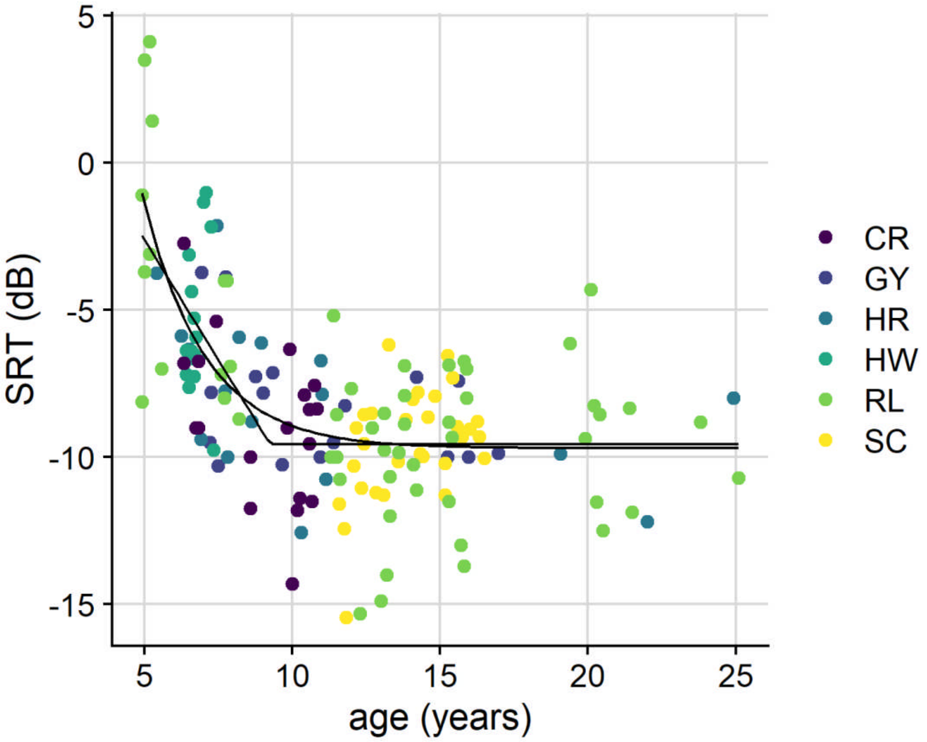

Figure 3: Regression models of SRT with age. The colour of the data points indicates which data set they arise from. The two continuous black lines show the predictions of an asymptotic regression (the curved line) and the broken stick regression (‘broken’ line).

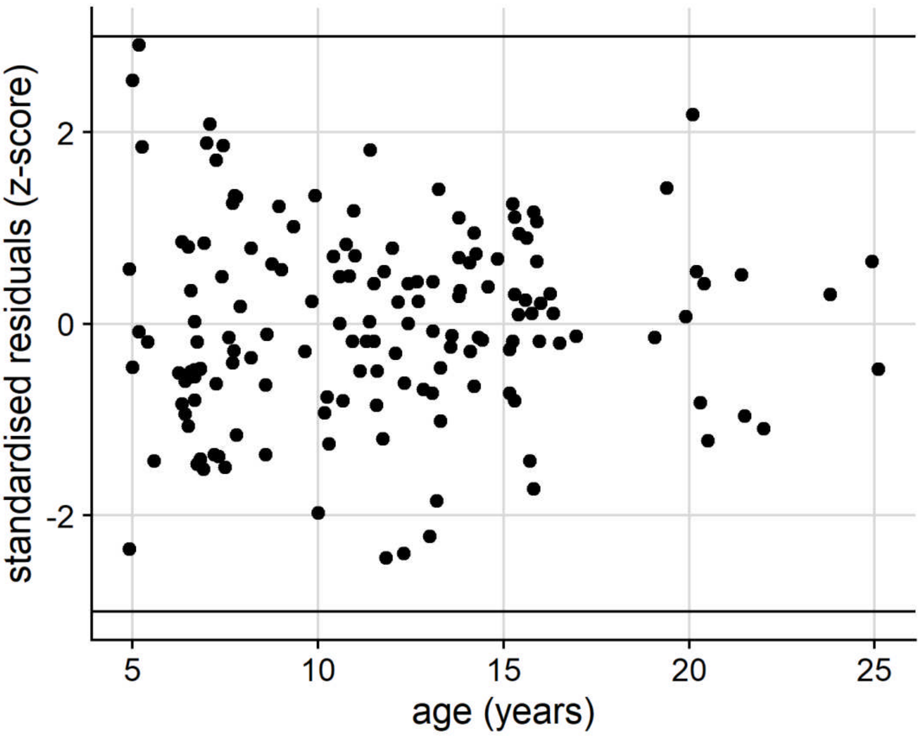

Figure 4: The standardized residuals from the broken stick regression.

Supplementary Material

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

Orthographic	Target			Foil 1 (distracter)		Foil 2 (distracter)	
	IPA	AoA	SNR	IPA	Feature	IPA	Feature
Familiarisation							
Bike	baɪk	4.79	2	waɪk	Manner	gaɪk	Place
Bin	bɪn	4.68	3	mɪn	Manner	gɪn	Place
Bus	bʌs	3.85	-4	wʌs	Manner	dʌs	Place
Dog	dɒg	2.80	2	nɒg	Manner	gɒg	Place
Doll	dɒl	3.68	0	ɹɒl	Manner	bɒl	Place
Duck	dʌk	3.50	-4	zʌk	Manner	gʌk	Place
Laugh	lɑːf	3.79	-2	zɑːf	Manner	wɑːf	Place
Leg	leg	3.00	-1	deg	Manner	jeg	Place
One	wʌn	3.23	-3	mʌn	Manner	lʌn	Place
Rain	ɹeɪn	3.60	0	neɪn	Manner	jeɪn	Place
Sea	siː	4.74	-7	ziː	Voicing	θiː	Place
Sun	sʌn	3.40	11	zʌn	Voicing	θʌn	Place
Watch	wɒtʃ	4.33	-3	gɒtʃ	Manner	ɹɒtʃ	Place
Wave	weɪv	4.26	-1	beɪv	Manner	leɪv	Place
Test items							
Bed	bed	2.89	-3	med	Manner	ped	Voicing
Book	bʊk	3.68	0	wʊk	Manner	pʊk	Voicing
Boot	buːt	3.89	5	wuːt	Manner	puːt	Voicing
Chair	tʃeə	3.43	0	seə	Manner	dʒeə	Voicing
Boat	bəʊt	3.84	-1	wəʊt	Manner	pəʊt	Voicing
Bag	bæg	4.28	-3	mæg	Manner	pæg	Voicing
Dig	dɪg	4.19	-3	nɪg	Manner	tɪg	Voicing
Towel	taʊl	3.22	-5	saʊl	Manner	paʊl	Place
Sing	sɪŋ	3.47	-13	tɪŋ	Manner	ʃɪŋ	Place
Knife	naɪf	4.15	0	daɪf	Manner	maɪf	Place
Wash	wɒʃ	4.00	-5	bɒʃ	Manner	ɹɒʃ	Place
Bath	bɑːθ	3.23	-4	wɑːθ	Manner	dɑːθ	Place
Leaf	liːf	4.60	2	niːf	Manner	wiːf	Place
Road	ɹəʊd	4.55	-2	zəʊd	Manner	jəʊd	Place
Cough	kɒf	4.32	18	pɒf	Place	gɒf	Voicing
Bite	baɪt	3.58	-5	daɪt	Place	paɪt	Voicing
Comb	kəʊm	5.50	9	pəʊm	Place	gəʊm	Voicing

Kite	kaɪt	4.58	5	paɪt	Place	gaɪt	Voicing
Cow	kaʊ	3.94	0	taʊ	Place	gaʊ	Voicing
Cake	keɪk	3.26	3	peɪk	Place	geɪk	Voicing
Fish	fɪʃ	4.05	1	hɪʃ	Place	vɪʃ	Voicing
Fork	fɔːk	3.63	4	sɔːk	Place	vɔːk	Voicing
Five	faɪv	4.51	4	ʃaɪv	Place	vaɪv	Voicing
Fall	fɔːl	4.71	0	sɔːl	Place	vɔːl	Voicing
Soap	səʊp	3.17	2	fəʊp	Place	zəʊp	Voicing
Foot	fʊt	3.44	4	hʊt	Place	vʊt	Voicing
Suck	sʌk	5.58	-8	hʌk	Place	zʌk	Voicing
Thumb	θʌm	4.42	3	ʃʌm	Place	ðʌm	Voicing

S2: Participants characteristics and testing environments

Group	**Adults (> 18 years)**	**Children (< 18 years)**	**Children's testing site**	**Total**
SC	0	30 children (age range=11.6-16.5, mean =14.0, sd = 1.5)	1 state secondary school in North London	30
RL	11 adults (age range: 19.4-25.1, mean = 21.1, sd = 1.8)	43 children (age range: 4.9-15.9, mean = 11.3, sd = 3.8)	2 state primary schools in North London 1 secondary school in South East England	54
GY*	0 adults	17 children (age range: 6.9-17, mean = 10.9, sd = 3.5)	Recruited widely from the UK	17
HR	3 adults (age range: 19.0-24.9, mean = 22.0, sd=3.9)	14 children (age range: 5.4 – 11.1, mean = 8.4, sd = 1.9)	1 state primary school in Devon 1 private primary school in London	17
HW	0 adults	18 children (age range: 6.4-7.3, mean = 6.8, sd = 0.3)	1 primary school in North London	16 (2 excluded)
CR	0 adults	19 children (age range: 6.3-10.8, mean = 9.0, sd=1.7)	South London primary schools	19
				153

*Participants in the GY group were control participants recruited as part of a study of children with developmental language disorder. See Baird et al., (2010) and Loucas et al. (2016) for full details.