Running Head: PIT_Cameron et al.

The Phoneme Identification Test (PIT) for assessment of spectral and temporal discrimination skills in children: Development, normative data and test-retest reliability studies

Sharon Cameron*
Nicky Chong-White *
Kiri Mealings*
Tim Beechey *
Harvey Dillon *
Taegan Young *

* National Acoustic Laboratories, Sydney, Australia

Parts of the research described in this paper were presented at the Audiology Australia National Conference, Melbourne, Australia, May 2016.

Acknowledgments. The authors would like to thank Mark Seeto for statistical advice. Thanks are also due to the NSW Department of Education and the primary school who took part in this research. The participation of the children and their families are also appreciated. This research is funded by the Australian Government through the Department of Health.

Corresponding Author:

Sharon Cameron, PhD
Senior Research Scientist
National Acoustic Laboratories
Australian Hearing Hub
16 University Avenue
Macquarie University NSW 2109
Australia
Phone: +61 2 9412 6851
Fax: +61 2 9412 6769
e-mail: Sharon.Cameron@nal.gov.au

Abstract

Background: Previous research suggests that a proportion of children experiencing reading and listening difficulties may have an underlying primary deficit in the way that the central auditory nervous system analyses the perceptually important, rapidly varying, formant frequency components of speech.

Purpose: The Phoneme Identification Test (PIT) was developed to investigate the ability of children to utilize spectro-temporal cues to perceptually categorize speech sounds based on their rapidly changing formant frequencies. The PIT utilizes an adaptive two-alternative forced-choice procedure whereby the participant identifies a synthesized consonant-vowel (CV) (/ba/ or /da/) syllable. CV syllables differed only in the second formant (F2) frequency along an 11 step-continuum (between 0 and 100% - representing an ideal /ba/ and /da/ respectively). The CV syllables were presented in either quiet (PIT Q) or noise at a 0 dB signal-to-noise ratio (PIT N).

Research Design: Development of the PIT stimuli and test protocols, and collection of normative and test-retest reliability data.

Study Sample: Twelve adults (23, 10 (yr, mo) to 50, 9, mean 32, 5) and 137 typically-developing, primary-school children (6, 0 to 12, 4, mean 9, 3). There were 73 males and 76 females.

Data Collection and Analysis: Data were collected using a touch screen computer. Psychometric functions were automatically fit to individual data by the PIT software. Performance was determined by the width of the continuum for which responses were neither clearly /ba/ nor /da/ (referred to as the Uncertainty Region, or UR). A shallower psychometric function slope reflected greater uncertainty. Age effects were determined based on raw scores. Z-scores were calculated to account for the effect of age on performance. Outliers, and individual data for which the confidence interval of the UR exceeded a maximum

allowable value, were removed. Non-parametric tests were used as the data were skewed toward negative performance.

Results: Across participants, the median value of the F2 range that resulted in uncertain responses was 33% in quiet and 40% in noise. There was a significant effect of age on the width of this uncertainty region (UR) ($p < 0.00001$) in both quiet and noise, with performance becoming adult-like by age nine on the PIT Q and age ten on the PIT N. A skewed distribution towards negative performance occurred in both quiet ($p = 0.01$) and noise ($p = 0.006$). Median UR scores were significantly wider in noise than in quiet ($T = 2041$, $p < 0.0000001$). Performance (z-scores) across the two tests was significantly correlated ($r = 0.36$, $p = 0.000009$). Test-retest z-scores were significantly correlated in both quiet and noise ($r = 0.4$ and 0.37 respectively, $p < 0.0001$).

Conclusions: The PIT normative data shows that the ability to identify phonemes based on changes in formant transitions improves with age, and that some children in the general population have performance much worse than their age peers. In children, uncertainty increases when the stimuli are presented in noise. The test is suitable for use in planned studies in a clinical population.

Key Words: Spectral processing, temporal processing, central auditory processing disorder; categorical perception, speech perception.

Abbreviations: 2AFC = two-alternative forced choice; ANOVA = Analysis of variance; CAPD = Central Auditory Processing Disorder; CI UR = Confidence Interval of the uncertainty region; CV = consonant-vowel; F2 = second formant; GPC = grapheme-to-phoneme conversion; GUI = graphical user interface; ILTASS = international long-term

average speech spectrum; MCL = most comfortable listening level; ms = milliseconds; ParSE = Parsing Syllable Envelopes Test; PIT = Phoneme Identification Test; RMS = root mean square; SNR = signal-to-noise ratio; UR = uncertainty region.

INTRODUCTION

Central auditory processing disorder (CAPD) is an umbrella term for a variety of disorders characterized by poor perceptual processing of auditory information in the central auditory nervous system despite the person having normal hearing thresholds (American Speech-Language-Hearing Association, 2005). Although research into CAPD has been conducted for over 60 years, and several national consensus statements exist (American Academy of Audiology, 2010; American Speech-Language-Hearing Association, 2005; British Society of Audiology, 2011; National Acoustic Laboratories, 2015), there is still some disagreement on the causes, diagnosis, and treatment of the disorder. Certainly there is no universal diagnostic criteria or test battery used for CAPD (Dillon et al., 2012; Vermiglio, 2014, 2016; Wilson and Arnott, 2013). Vermiglio (2016) proposes that new research needs to be conducted and new tests developed that can accurately diagnose more of the unknown specific deficits that are causing children's listening problems, suggesting temporal resolution difficulties as a potential diagnostic test target.

The secondary auditory cortex performs the temporal operations that are important for analyzing rapidly changing acoustic cues, allowing listeners to identify speech sounds based on their phonetic features (Specht, 2014). There are two types of time or "temporal" acoustic cues in an acoustic signal. The first is temporal fine structure, which contains information about fundamental frequency, harmonics and formant transitions. The second acoustic cue is

the more slowly varying amplitude changes, referred to as the temporal envelope of speech (Goswami, 2011; Goswami et al., 2011; Rosen, 1992).

Our goal was to develop tests assessing both fine temporal processing ability, as well as temporal envelope cues, in children with suspected auditory processing deficits. These skills are essential for accurate speech perception and reading development (Goswami, 2011; Stevens and Keyser, 1989; Vandermosten et al., 2011). Specifically, the tests aim to investigate how children perceptually categorize speech sounds based on their rapidly changing formant frequencies (the current paper), and how well they perceive where syllables start and finish based on amplitude modulations at syllable boundaries (Cameron et al., submitted).

Understanding speech relies on a person's auditory system being able to categorically map highly variable acoustic speech signals into discrete phonetic units (Boets et al., 2013; Chang et al., 2010; Liberman et al., 1967). This mapping is essential because people's vocal tracts are capable of producing a wide variety of physical sounds that are intended to represent the same phoneme (Liebenthal et al., 2005). The ability to identify and discriminate speech sounds from this continuum of possible sounds is known as "categorical perception" (Phillips et al., 2000). Categorical perception allows the listener to hear quantal jumps between phonemic categories rather than hearing step-like intra-phonemic variations that correspond to changes in the acoustic signal (Liberman et al., 1967). The human posterior superior temporal gyrus is largely responsible for categorically organizing the neural representation of speech sounds (Chang et al., 2010). For example, this allows listeners to distinguish between the three stop consonants in /ba/, /da/, and /ga/, which differ only in the starting frequency of the second vocal tract resonance formant (F2) transition (Chang et al., 2010).

Not all people are able to accurately categorize speech sounds, however. Phonological dyslexia is a condition where people have difficulty reading regular and/or non-words that follow grapheme-to-phoneme conversion (GPC) rules (e.g. “cat” or “gop”) despite having normal intelligence and access to normal reading instruction (McArthur et al., 2013). Phonological dyslexia is one type of developmental dyslexia. McArthur et al. (2013) utilized word reading profiles to better understand the heterogeneity in dyslexia. Children with phonological dyslexia exhibit poor sub-lexical reading ability (that is, they cannot utilize GPC rules to decipher regular and non-words) but have normal lexical reading ability (that is, they can memorize whole words by site). In contrast, children with surface dyslexia exhibit very poor lexical reading ability and good sub-lexical reading ability, manifesting in difficulties reading irregular words such as “ghost”. The majority of children exhibit a mixed dyslexia, having varying degrees of both poor sub-lexical and lexical reading ability.

A review of 50 papers by Vandermosten et al. (2011) found less consistent stop consonant categorical perception by people with dyslexia compared to those with normal reading skills in 64% of the studies reviewed. In their own categorical perception study, the authors assessed categorical perception in 13 Dutch children with at high family risk of dyslexia aged 11 years and 25 age-matched controls. Children in the dyslexic group scored below the 10th percentile on a standardized reading, non-word reading and spelling task on at least two successive occasions. In a two-alternative forced-choice adaptive (2AFC) ABX procedure, the participant’s task was to identify which of the two preceding sounds the target stimulus (X) was most similar to. The ability to identify both rapidly changing and steady state speech and non-speech sounds was measured. The dyslexic group were significantly less able to categorize the rapidly changing speech and non-speech stimuli ($p = 0.003$). However, there was no difference between groups in perception of steady-state stimuli. Vandermosten et al. (2011) also compared the child data to data collected from dyslexic

adults and controls (Vandermosten et al., 2010) and found that categorical perception improved from child to adulthood in both groups, but the differences between the groups remained, even in adulthood.

Similar to CAPD, the heterogeneous nature of developmental dyslexia makes diagnosis and intervention difficult (McArthur et al., 2013). We need, therefore, to gain a better understanding of the more specific mechanisms driving specific impairments. CAPD and dyslexia are often co-occurring disorders (King et al., 2003). We hypothesize that a proportion of children with both listening and reading difficulties have an underlying primary deficit in the way that the central auditory nervous system analyses the rapidly changing frequency and amplitude components of speech. Continued exposure to low-resolution internal representations of the speech sounds may result in indistinct speech sound templates being created, which may then lead to difficulties mapping speech sounds to orthographic symbols (i.e. sound/letter correspondence). Without these clear auditory building blocks, we believe that listening difficulties and deficits in phonetically-based reading skills will result.

Any such deficits in categorical perception are likely to be most evident in noise. A recent study showed poorer consonant discrimination (/ba/ versus /da/) with increased background noise. Electroencephalography (EEG) analysis also revealed increased mismatch negativity latencies, decreased amplitudes, and decreased power in theta frequency band with decreased signal-to-noise ratio (Koerner et al., 2016). The effects of background noise on the speech-evoked frequency following response in infants was investigated by White-Schwoch et al. (2015). The stimulus was the consonant-vowel syllable /da/. It was found that responses were degraded in noise, being smaller, slower and less stable across trials, with poorer coding of the spectral content and temporal envelope. The authors note that background noise presents a challenge during early childhood, when children are attempting to form precise representations of speech sounds.

The overall aim of our research was to develop a set of clinical assessment tools that could help to identify a subgroup of children whose auditory processing and/or reading deficits related specifically to spectro-temporal resolution deficits impacting analysis of the incoming acoustic signal. The current paper specifically outlines the development of a new categorical perception task – the Phoneme Identification Test (PIT) – that has the potential to become a clinical tool for diagnosing temporal fine structure resolution deficits. The methods used to develop the test, as well as collection of normative test and retest data are described. It was hypothesized that children's categorical perception would be poorer than adults, but that it would improve with age. It was also hypothesized that children's categorical perception would be less accurate in the presence of noise due to the masking effect of the noise on the temporal fine structure features contained in the speech signal.

METHOD

Approval for the study was granted from the Australian Hearing Human Research Ethics Committee and the New South Wales Department of Education.

Participants

A total of 158 participants were initially assessed. There were 12 adults (23, 10 (yr, mo) to 50, 9, mean 32, 5) of which nine were female and three were male. All had normal hearing defined as equal to, or better than, 20 dB HL at all octave frequencies from 250 Hz to 8000 Hz measured bilaterally using an Interacoustics AC40 audiometer (Middelfart, Denmark) with Telephonics TDH 39P audiometric headphones (Huntington, NY) in H7A Peltor cups (3M, St. Paul, MN). The child participants were recruited from a Sydney primary school. Children whose parents reported they had an attention, language, or learning problem in the

study consent form were excluded from participating. Children's hearing was assessed on the day of testing with the PIT and only those who passed the pure tone audiometric screening test participated in the study. Audiometric testing was as for the adult participants, with the exception that an Interacoustics Audio Traveller A222 portable audiometer was used. Data were collected from 146 children. Four children were excluded post-testing on the PIT Q and six on the PIT N due to inconsistent performance, as documented in the following section (Exclusions Based on Confidence Intervals). Further, five outliers were excluded on the PIT Q and four on the PIT N, as documented in the section on calculation of z-scores (i.e. age-adjusted populations standard deviation units). As such, PIT Q data are presented for 137 children (6, 0 (yr, mo) to 12, 3, mean 9, 2), of which 67 were female and 70 were male. PIT N data are presented for 136 children (6, 0 (yr, mo) to 12, 4, mean 9, 3), of which 67 were female and 69 were male.

Software Development

The PIT graphical user interface (GUI) and signal processing application were developed in MATLAB programing language (MathWorks Inc., 2014), and compiled for use on a touchscreen computer. Three screens were developed: a data capture screen for collection of client information and activation of reference tone; an operations screen for activation of practice, familiarization and test materials for PIT Q and N; and a test screen for the participant to respond to the PIT stimuli. An image of the test screen appears as Figure 1.

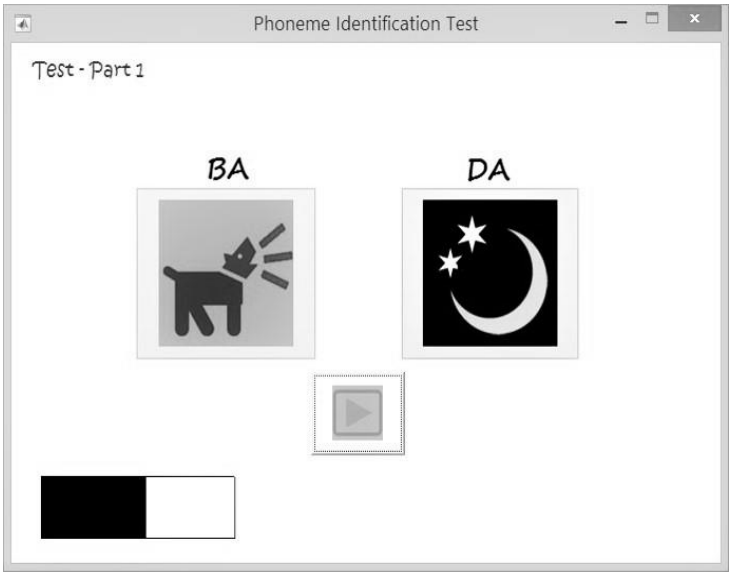


Figure 1. PIT test screen.

Stimuli

The PIT stimuli is comprised of 11 synthetic consonant-vowel (CV) tokens. Each CV token was 315 milliseconds (ms) in length, consisting of a 50 ms portion across which formant 1 and 2 (F1 and F2) transitioned in frequency. This was followed by a 250 ms steady state portion. In all 11 tokens, F1 rises linearly from an initial frequency of 300 Hz to the steady state frequency of 750 Hz. F2 was the only formant manipulated in the experiment. The initial F2 frequency varied across tokens between 1000 and 1500 Hz, and linearly transitioned to a steady state frequency of 1200 Hz, to create the 11 step continuum, so that the end points (tokens 1 and 11) represented an ideal /ba/ (rising FS) and /da/ (falling F2) respectively. The middle (6^{th}) token had an initial F2 frequency of 1250 Hz, midway between the initial F2 frequencies of each of the endpoint tokens. F3, F4 and F5 were steady state throughout the entire duration of each token (see Table 1).

The tokens described above were synthesized with a sampling rate of 44100 Hz using the source-filter model provided by Praat (version 5.4.04) (Boersma and Weenink, 2014). The voicing source consisted of a pulse train with a fundamental frequency (F0) sloping from

110 to 100 Hz. A filter representing the shape of the vocal tract, consisting of five formants with steady state frequencies of 750, 1200, 2350, 3300, and 4000 Hz, was created based on frequencies used by Blomert and Mitterer (2004). Formant bandwidths of 50, 60, 110, 160, and 210 Hz were used, based on bandwidths reported by Fant (1962). The filter was applied to the voicing source to produce a low, central /a/ vowel. The resulting vowel has the temporal properties of the source with the spectral properties of the filter. A low central vowel was selected due to its common use in similar experiments (Blomert and Mitterer, 2004; Goswami, 2011; Serniclaes and Sprenger, 2001).

Using MATLAB (MathWorks Inc., 2014) the end of each token was trimmed at zero crossings and 2.5 ms cosine ramps were applied to the onset and offset to reduce any popping or abruptness. The total length of each token was 315 ms. Each token was then normalized to a total root mean square (RMS) value of -20 dB (ref. full-scale square wave, 50 ms window) in Adobe Audition CS6 (see Figure 2).

A looped, four minute section of international long-term average speech spectrum (ILTASS) filtered noise from the NAL Speech and Noise for Hearing Aid Evaluation CD (National Acoustic Laboratories, 2000) was used as the masking noise for the PIT 0 dB SNR condition. The audio file was level normalized to a total RMS value of -20 dB (ref. full-scale square wave, 50 ms window).

Table 1. Formant frequencies at the start of the formant transition of /b/ and /d/ and during the steady-state vowel.

	Token 1 (/b/) initial frequency (Hz)	Token 11 (/d/) initial frequency (Hz)	Steady-state /a/
F1	300	300	750
F2	1000	1500	1200
F3	2350	2350	2350
F4	3300	3300	3300
F5	4000	4000	4000

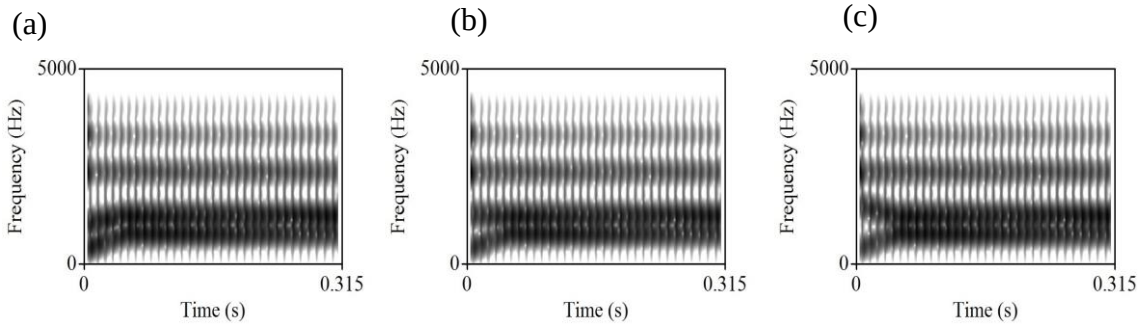


Figure 2. Endpoints and midpoint of the 11 step /ba/-/da/ continuum. (a) Ideal /ba/ with a rising F2; (b) an ambiguous token with a level F2; (c) ideal /da/ with a falling F2.

Task

In an adaptive two-alternative, forced choice (2AFC) procedure the listener's task was to indicate whether each trial contained /ba/ or /da/ by pressing the corresponding button on the

touchscreen. The response buttons became active 300 ms after the offset of the stimulus to ensure that participants did not respond before hearing the complete stimulus. Participants did not receive any feedback following responses. A total of 92 tokens were presented in two blocks for both the PIT Q and the PIT N. The first block contained 48 randomly ordered presentations including 8 presentations of each odd continuum step (0%, 20%, 40%, 60%, 80%, and 100%). At the end of the first block the threshold (inflection point) of the psychometric function was automatically calculated by the PIT software. The second block contained 48 randomly ordered presentations consisting of four presentations of each endpoint token plus eight presentations of each of the five tokens clustered around the threshold of the psychometric function. These five tokens included the tokens nearest to 5% and 95% of the psychometric function and the nearest token above and below the threshold plus the next nearest token (either above or below the threshold). Practice and familiarization conditions were presented orally by the examiner prior to testing, as described below. The practice, familiarization and test instructions are provided in Appendix A. Including practice and familiarization, the PIT took approximately ten minutes to complete for children aged eight to 12 years, and approximately 12 minutes to complete for six and seven year olds.

Practice and Familiarization Procedure

Participants completed a brief practice task prior to each test condition (quiet and 0 dB SNR). The practice task consisted of five repetitions of each endpoint token (i.e. F2 100% /ba/ or F2 100% /da/) presented in random order. Following each response the participant was provided with visual feedback (the words *correct* or *incorrect* appearing on the screen). Following the practice task participants performed a brief training task in quiet to ensure familiarity with the ambiguous tokens and to provide an understanding of how the ambiguous tokens relate to the endpoint tokens. Each continuum step was presented once in order from

step 0 to step 10. After each token the participant selected whether they heard /ba/ or /da/. No feedback was given following responses.

Reporting

Psychometric functions using a logistic curve were automatically fit to individual data by the PIT software. A graph of the results was displayed on the computer screen following testing. Figure 3 provides an image of a typical result for a child on the PIT Q and the PIT N (z-scores of -0.09 and 0.05 respectively).

- The y-axis corresponds to the proportion of /da/ responses (range 0 to 1). Thus, 0 represents /ba/ selected 100% of the time, whereas 1 represents /da/ being selected 100% of the time.
- The x-axis corresponds to percentage of F2 /da/ (range 0% to 100%). Thus 0 represents 100% /ba/, whereas 1 represents 100% /da/.

The following performance criteria were calculated by the software and stored in a spreadsheet.

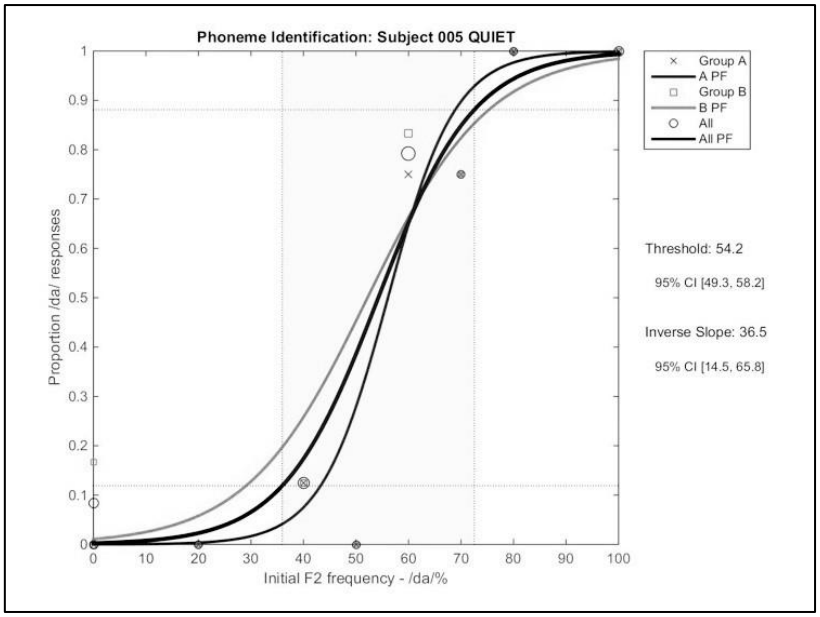
a. Threshold: The threshold is the point of maximum uncertainty. That is, the F2 % /da/ where a participant identifies /da/ 50 percent of the time and /ba/ 50 percent of the time.

b. Uncertainty Region (UR): Performance was determined by the width of the continuum for which responses were neither clearly /ba/ nor /da/. Specifically, the UR equals the inverse of the slope of the calculated parametric function, evaluated at its threshold. This is identical to the change in F2% /da/ between the point at which the psychometric function has a value of 0.12 through to the point where the psychometric function has a value of 0.88, as shown by the dissection lines surrounding the shaded region in Figure 3.

The distance between the corresponding points on the x-axis represent the uncertainty region.

c. Confidence Interval of the uncertainty region (CI UR): The CI UR indicates the reliability with which the UR is estimated for an individual child. The CIUR is obtained using a non-parametric bootstrapping technique with B=400 simulations. The observed proportions correct at each stimulus step (from F2 100% /ba/ to F2 100% /da/) were used directly to generate the bootstrap simulations. The slope parameter of the best-fitting logistic function for each set of simulated responses is sorted in order and the 95% confidence interval is determined using the (B*2.5%)th and (B*97.5%)th values. The width of the confidence interval of the uncertainty region (CI UR) is equal to the range of stimulus percentage values from the corresponding 2.5^{th} percentile UR to the 97.5^{th} percentile UR. Note that although the stimuli can have values only between 0 and 100%, both the UR and CI UR can exceed a range of 100%, and do so whenever the slope of the psychometric function is less than unity (with values on both axes expressed in percent, or both expressed in proportions).

(a)



(b)

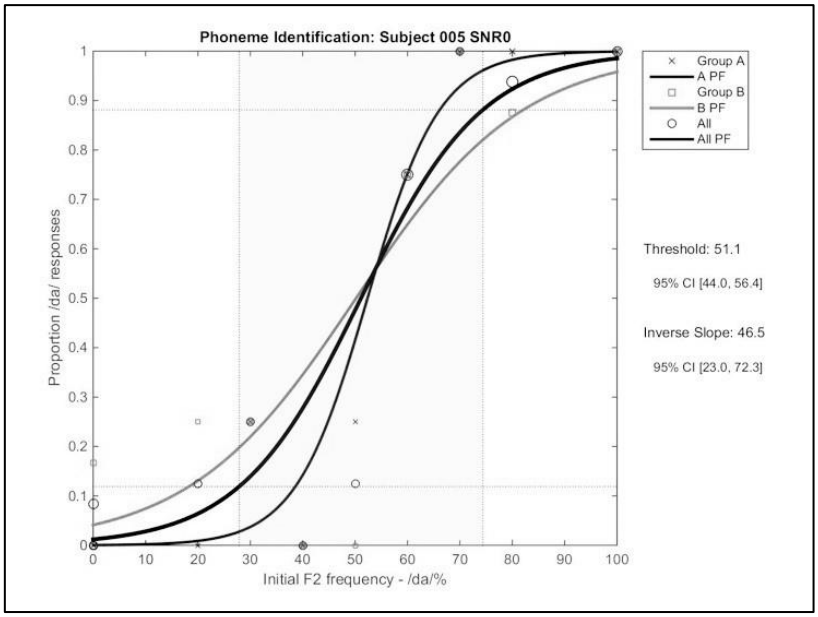


Figure 3. Image of the PIT results screen for a participant 9 yr, 8 mo of age. (a) PIT Q result and (b) PIT N result shows z-scores of -0.09 and 0.05 respectively. The curved lines represent the psychometric functions fitted to the data. The dark-colored central curved line is the curve fitted to all measured data, and the other two lines are the curves fitted to odd and even-numbered responses, respectively. The dashed lines show the upper and lower boundaries of the uncertainty region.

Procedure

Testing for the adult participants was conducted in a sound-attenuated room at the National Acoustic Laboratories using a Sony Vaio Duo 11 touchscreen computer (Sony, Japan). For the child participants, testing was completed in a quiet room at their primary school. Sound levels in the school testing rooms were measured between 45-50 dBA using a Q1362 digital sound level meter (Dick Smith Electronics, Australia). Data was collection using a Microsoft Surface Pro 3 touchscreen computer (Microsoft, China).

The pre-generated stimuli were presented binaurally, using the PIT software, through Sennheiser HD 215 circumaural headphones. All tokens were presented at a volume control setting calibrated to 77 dB SPL during the steady-state /a/ vowel using a Brüel & Kjær Head and Torso Simulator Type 4128C (Naerum, Denmark). This level corresponded to a volume level of 40 on the Sony computer and 18 on the Surface computer. This level was selected based on the average most comfortable listening level (MCL) of the test stimuli chosen by four normal-hearing adult listeners. The quiet condition (PIT Q) was always presented first. All subjects took a one minute break between blocks and a two minute break between the quiet and noise conditions.

Exclusions based on Confidence Intervals

As noted in the participant section, only children whose parents reported no attention or learning deficits on the study consent form were assessed with the PIT. However, an additional inclusion criteria was implemented post-testing based on the width of the confidence interval of the uncertainty region (CI UR) recorded for the PIT Q and PIT N for each participant. If the CI UR was greater than 300, then the result was rejected as invalid as

the reliability of the fitted psychometric function was very poor. In such cases, the slope of the fitted function was extremely flat and the measured data were consistent with the responses being essentially random. Based on this criterion, and as noted in the participant section, of the 146 children assessed, data were excluded for four children on the PIT quiet (Q) condition and six children on the PIT noise (N) condition. The majority of children excluded (50% on the PIT Q and 83% on the PIT N) were aged 6, 0 (yr, mo) to 6, 12.

RESULTS

Statistical analysis was performed using Statistica version 10. For each participant, performance on the PIT Q and PIT N was determined by the uncertainty region (UR) score, which was converted to a z-score, as described below. Effects of age were calculated using raw scores. All other analyses, including correlations between measures, were calculated using the z-scores. Use of z-scores removes the contribution that age makes to correlations between the measures (because all test scores on average improve with age). Correlation coefficients based on z-score data will therefore be smaller than those based on raw scores. As the data were skewed towards negative performance non-parametric analyses were utilized.

Calculation of Z-Scores and Removal of Outliers

The uncertainty region (UR) scores, collected from the children and adults remaining following exclusion based on confidence intervals, were used to create equations that allow the expression of individual scores in age-corrected population standard deviation (SD) units (z-scores).

The raw UR scores were regressed against age using the exponential formula UR=a+b*exp (-age/c) where a, b and c are the coefficients that determine the curve. These three coefficients determine the asymptotic value applicable to adults (a); the rate of change with age (b), and the age above which the effect of age starts to diminish (c). This equation calculated the predicted UR score for participants as a function of age.

Residual scores were calculated as the difference between the each participant's actual score and the predicted score for participants of that age. The squared residual scores were regressed against using the formula described earlier, but with new values of a, b and c. The square root of this second regression formula was used to predict the standard deviation of UR scores at any age. The coefficients are reported in Table 2.

Table 2. PIT Q and PIT N uncertainty region (UR) (% F2 da), regression coefficients used in the creation of z-scores.

	Mean			SD		
Measure	a_1	b_1	c_1	a_2	b_2	c_2
PIT Q	18.3	564	2.621	68.8	14873	2.357
PIT N	18.5	608	3.042	18.9	14625	2.657

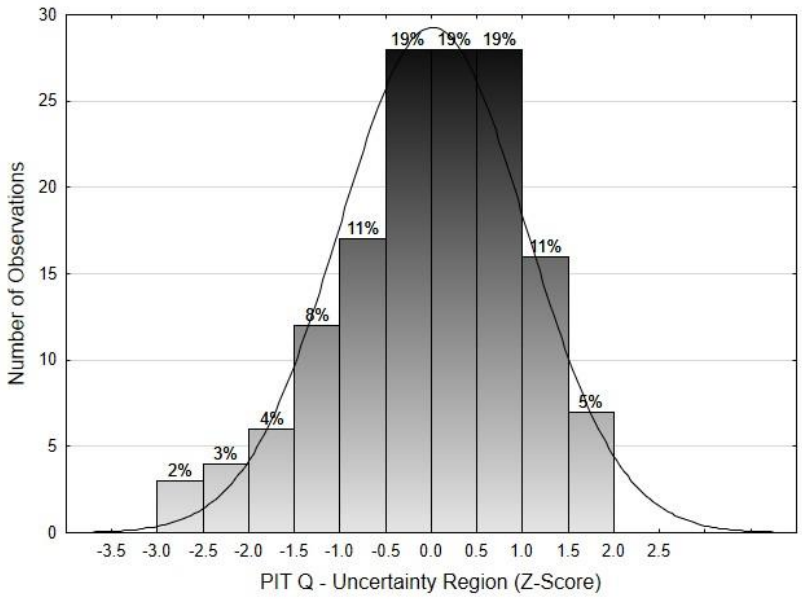
UR scores for each participant were then were standardized to a mean of zero and unity standard deviation and reported as z-scores, using the formula:

$z = (a+b*\exp(-Age/c) - score)/SD_{predicted}$

The UR z-scores were examined to determine if they deviated significantly from a normal distribution. For the PIT Q (n=154), the Shapiro-Wilk *W* value was 0.91 ($p < 0.000001$) with z-scores ranging from 1.5 to -4.5 (mean -0.0006). For the PIT N (n=152), the Shapiro-Wilk *W* value was 0.89 ($p < 0.00001$), with z-scores ranging from 1.9 to -4.6 (mean 0.003). Outliers (z-scores poorer than 2.5 standard deviations below the mean) were removed from the normative data. At this cut-off point it would be expected that approximately one participant would be removed were the distribution to be normal. However there were five children removed from the PIT Q and four from the PIT N, demonstrating a skew towards decreased precision of categorical perception.

Z-scores were recalculated for the remaining participants. For the PIT Q (n=149), the Shapiro-Wilk *W* value was 0.98 ($p = 0.01$), with z-scores ranging from 2.0 to -3.0 (mean = -0.005). For the PIT N (n=148), the Shapiro-Wilk *W* value was 0.97 ($p = 0.006$), with z-scores ranging from 1.9 to -3.0 (mean = 0.001). Histograms are provided as Figure 4.

(a)



(b)

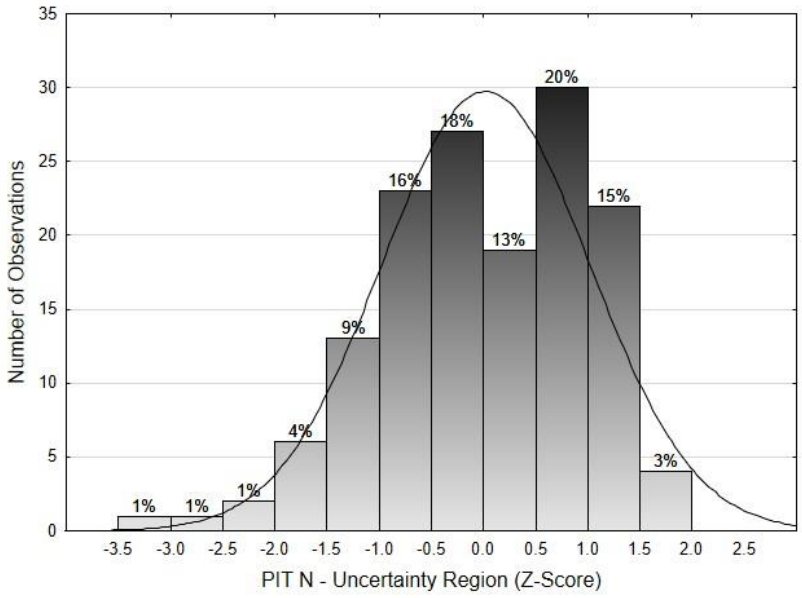


Figure 4. Histograms of uncertainty region z-scores for the (a) 149 participants on the PIT Q and the (b) 148 participants on the PIT N.

Gender Effects

For the PIT Q the median uncertainty region (UR) was 31% for males (n = 73) and 33% for females (n = 76). For the PIT N the UR was 41% for males (n = 72) and 40% for females (n = 76). The Kruskal–Wallis H test (one-way ANOVA on ranks) was used to determine if threshold differed significantly between groups. Age was controlled for by comparing z-scores. There was no significant difference between males and females for either the PIT Q, (H (1) = 0.107, p = 0.74), or the PIT N (H (1) = 0.021, p = 0.88).

Age Effects – Thresholds

The mean and median PIT P and PIT N thresholds, as a function of age, are provided in Table 3. As there were only a small number of 12 year-olds with a maximum age of 12, 4 (yr, mo), data from 11 and 12 year-olds were combined. Across age groups, the median threshold (% F2 /da/) was 51% on both the PIT Q and PIT N. The Kruskal–Wallis H test was used to

determine if threshold differed significantly between age groups. There was no effect of age on PIT Q threshold (H (6) = 6.65, p = 0.59). There was an effect age on PIT N threshold (H (6) = 21. 9, p = 0.001). Seven-year-olds had a lower threshold (median 43%) than 11-year-olds and adults (median 53% and 55% respectively).

Age Effects – Uncertainty Region

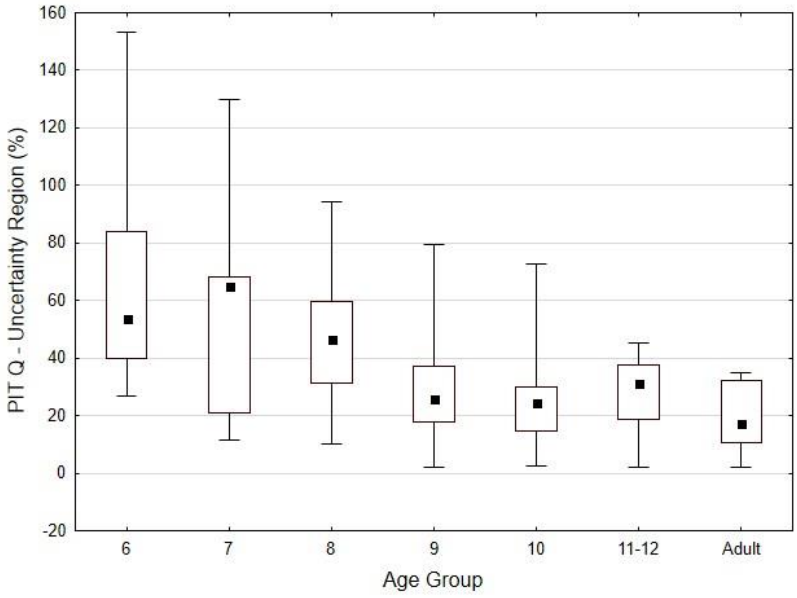
The mean and median PIT P and PIT N UR scores, as a function of age, are provided in Table 3. Children aged 11 and 12 years were combined, as noted above. Across age groups, the median UR width (% F2 /da/) was 33% on the PIT Q and 40% on the PIT N. There was a trend of reduced uncertainty with increased age. The Kruskal–Wallis H test revealed the effect of age was significant on the PIT Q UR (H (6) = 50.4, $p < 0.00001$). Children aged six to eight had significantly wider uncertainty regions than older children and adults. There was also a significant effect of age on the PIT N UR (H (6) = 68.0, $p < 0.00001$), with children aged six to nine having significantly wider uncertainty regions than older children and adults (see Figure 5).

A scatterplot of individual raw uncertainty region data on the PIT Q and PIT N as a function of age is provided as Figure 6. The ± 2 standard deviation limits, calculated from the regression equations, are delineated.

Table 3. Median, mean and standard deviations for PIT Q and PIT N threshold and uncertainty region (UR) (% F2 da), as a function of age.

	PIT Q							PIT N						
		Threshold (%)			Uncertainty Region (%)				Threshold (%)			Uncertainty Region (%)		
Age	N	Median	Mean	SD	Median	Mean	SD	N	Median	Mean	SD	Median	Mean	SD
Overall	149	51	50	9	33	38	26	148	51	49	13	40	50	3
6	21	48	48	11	54	63	31	19	48	45	21	104	89	40
7	19	51	50	10	65	57	32	19	43	38	18	63	73	35
8	20	48	49	7	47	46	22	18	49	48	14	67	61	21
9	23	49	48	10	26	27	17	23	49	51	11	37	43	20
10	23	51	50	7	25	25	14	25	51	51	10	32	40	20
11-12	31	53	52	10	31	28	12	32	53	52	6	28	31	13
Adult	12	51	53	9	17	19	12	12	55	55	7	20	19	5

(a)



(b)

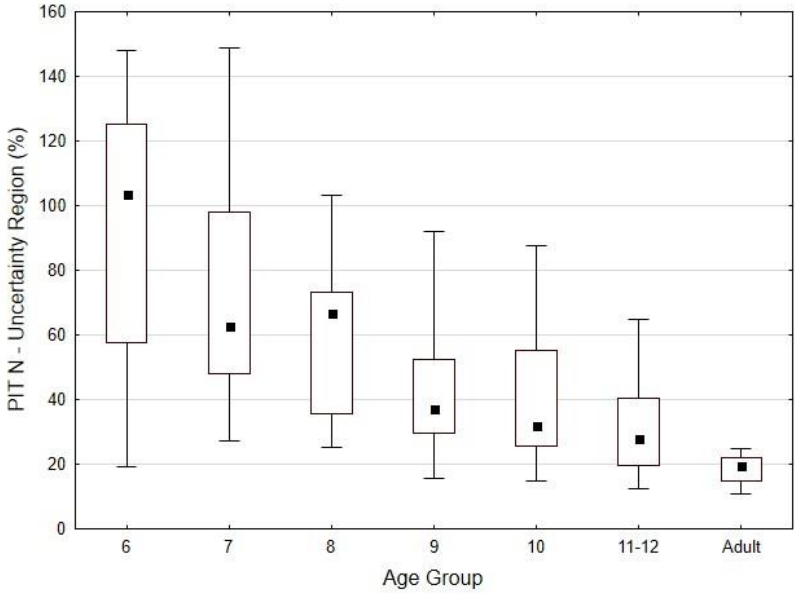
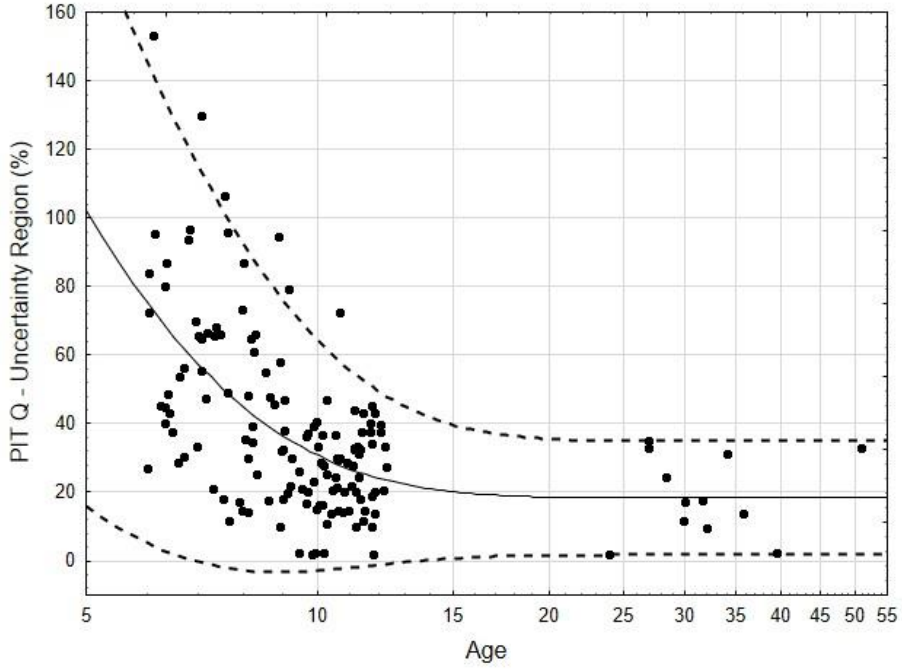


Figure 5. Box and whisker plots of uncertainty region (UR) as a function of age for the (a) 149 participants on the PIT Q and the (b) 148 participants on the PIT N. The filled square represents the median UR, the open boxes represent the 25-75% boundaries, the whiskers represent the minimum and maximum scores.

(a)



(b)

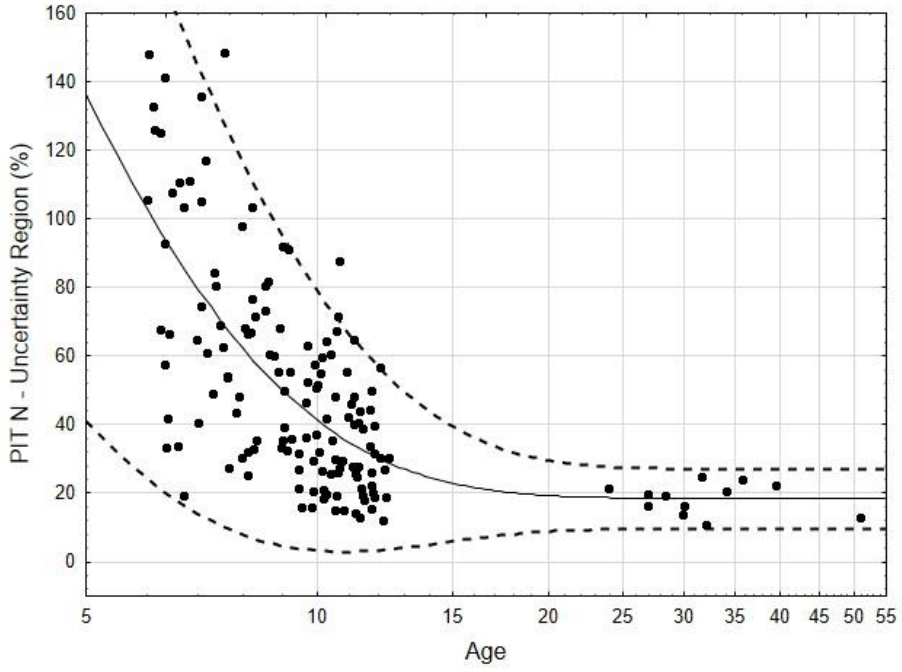


Figure 6. Scatterplots of the individual raw uncertainty region scores for the (a) PIT Q (n = 149) and PIT N (n = 149). The solid line indicates the mean score as a function of age, the dashed line shows the ± 2 SD limits.

Comparisons Between Performance in Quiet and Noise

A total of 143 participants completed both the PIT Q and PIT N. The Wilcoxon Matched Pairs test revealed that the median uncertainty region (UR) in noise (40% F2 /da/) was significantly wider than in quiet (32% FA /da/), (T = 2041, p < 0.000001). Spearman rank order correlations revealed a strong correlation (r = 0.66, p < 0.00001) between PIT Q and PIT N raw UR scores, however the effect of age would contribute to the strength of the correlation. As shown in Figure 7, the relationship between performance (expressed as z-scores) on the PIT Q and PIT N was also significant, although weaker (r = 0.36, p = 0.000009).

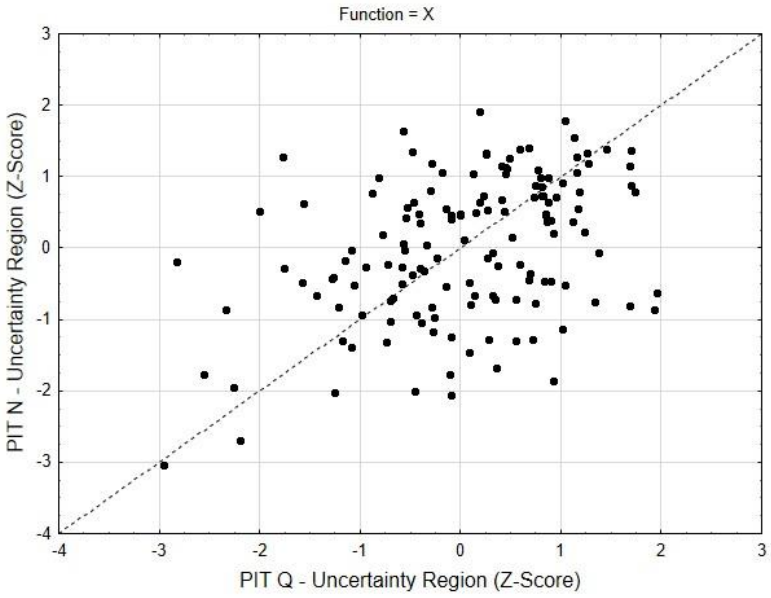


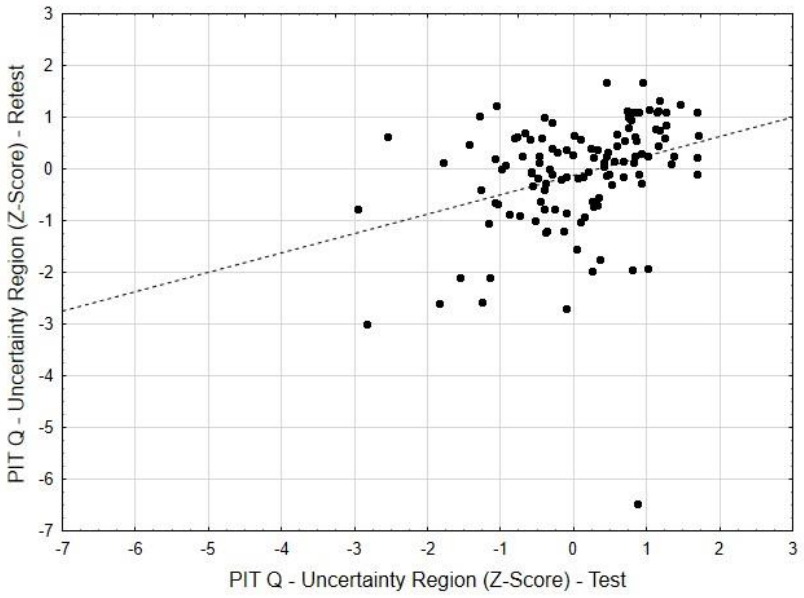
Figure 7. Scatterplot of PIT performance in quiet versus noise, measured as z-scores, for the 143 participants who completed both the PIT Q and PIT N. The dashed line represents equal age-adjusted performance on the two tests.

Test-Retest Reliability

Test-retest reliability data was analysed for 119 children on the PIT Q. Results for an additional four children were rejected prior to statistical analysis due to invalid confidence intervals. Participants were retested between 16 days and 44 days (mean 35 days) after their initial appointment. The median uncertainty region (UR) z-scores were 0.14 SD at test and 0.13 SD at retest (mean 0.05 and -0.1 SD respectively). Repeated measures (Wilcoxon Matched Pairs) test revealed no significant difference between test and retest (T = 2998, p = 0.13). Spearman rank order correlations revealed a moderate correlation (r = 0.4, p = 0.000007) between test and retest UR z-scores (Figure 8a). The mean difference between retest and test z-scores was -0.15 SD. The Pearson's product moment UR z-score test-retest correlation was r = 0.32 (p=0.0004). Consequently the proportion of variance accounted for by measurement error in the test scores is estimated as 68% (equal to 1-r).

Test-retest reliability data was analysed for 117 children on the PIT N. Results for an additional six children were rejected prior to statistical analysis due to invalid confidence intervals. The median uncertainty region (UR) z-scores were 0.19 SD at test and 0.25 SD at retest (mean 0.05 and -0.02 SD respectively). Repeated measures (Wilcoxon Matched Pairs) test revealed no significant difference between test and retest (T = 3347, p = 0.90). Spearman rank order correlations revealed a moderate correlation (r = 0.37, p = 0.000033) between test and retest UR z-scores (Figure 8b). The mean difference between retest and test z-scores was -0.07 SD. The Pearson's product moment UR z-score test-retest correlation was r = 0.42 (p<0.00001). Consequently the proportion of variance accounted for by measurement error in the test scores is estimated as 58%.

(a)



(b)

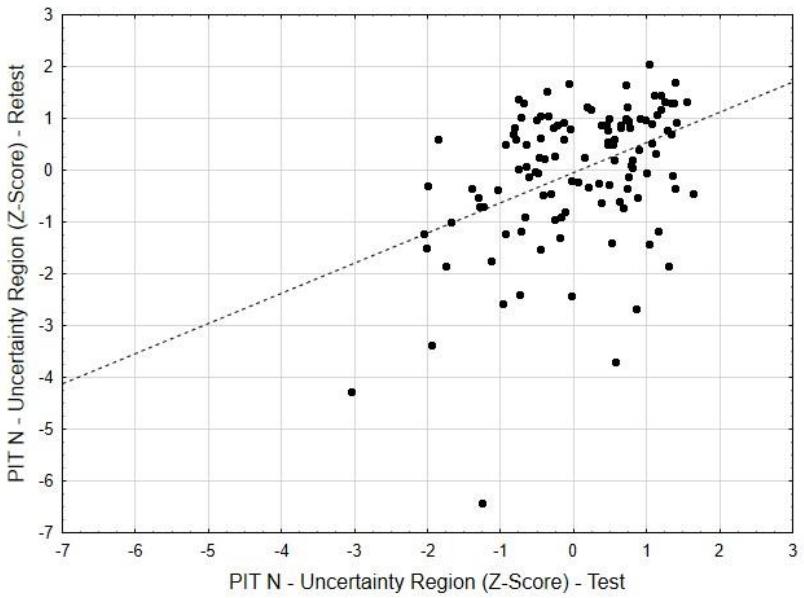


Figure 8. Scatterplots of the uncertainty region scores (z-scores) at test and retest for the (a) 119 children retested on the PIT Q and the (b) 117 children retested on the PIT N. The dashed line represents the least squares regression line.

DISCUSSION

The present paper documents the development of a new test of temporal resolution intended for future clinical use for children with suspected CAPD and/or reading deficits. The Phoneme Identification Test (PIT), uses an adaptive categorical perception task to evaluate an individual's ability to analyse the temporal fine structure of an incoming acoustic signal. The individual hears randomized synthesized speech sounds along a /ba/ and /da/ continuum. The participant's task was to indicate whether /ba/ or /da/ was heard by selecting a corresponding image on a touch screen computer. Performance was determined by the participant's uncertainty region (UR), being the width of the continuum for which responses were not consistently recorded as /ba/ nor /da/. The consonant-vowel tokens were presented in either quiet (PIT Q) or speech-shaped noise at a 0 dB signal-to-noise ratio (PIT N). Normative and retest reliability data was collected from over 140 children and adults.

It was found that the median threshold – that is, the % /da/ where an individual perceives that they heard /ba/ half the time and /da/ half the time - was 51% /da/ in both quiet and noise. This is not an unexpected result as the stimuli were synthesized along an 11 step continuum so that 50% /da/ would be the most ambiguous token. However, it does provide substantiation that the stimuli were perceptually balanced, not biased towards one end point or the other (that is, listeners were on average, no more likely to hear /da/ than /ba/).

As found in previous research (Vandermosten et al., 2011) categorical perception boundaries are smaller in adults than in children. Our study found that the PIT uncertainty region decreased with increasing age. In quiet the median UR was 33% and children six to eight years had significantly wider URs than older children and adults. When masking noise was added the median width of the UR increased to 40% and performance did not become adult-like until age ten. There was a strong correlation between performance in quiet and

noise, however when the effect of age was removed using z-scores that correlation was weaker, although still moderate in strength.

It is clear that the distribution of normative data test scores deviated from a normal distribution, with a skew towards below-average performance. That is, the worst performers were further below the mean performance than the better performers were above the mean. It is not possible to say whether this reflects an intrinsic skew in the range of abilities of phoneme perception, or is a product of the particular measure used to describe performance – in this case the range of second formant values over which the children were uncertain about the identity of the phoneme. We could, for example, have chosen to express performance in terms of the slope of the fitted psychometric curve, which is inversely proportional to the width of the uncertainty region. The distribution of slopes is also non-normal, but in this case the distribution is skewed towards children who perform much better than normal. While we could have performed a mathematical transformation that normalises the distribution (with either metric), we have chosen not to, as there is no a priori reason why the ability to categorically perceive phonemes should be normally distributed. Importantly, any such transforms do not change the rank order of children's performance on the task, so the only effect of such a transform is to change the z-score at which one considers performance is sufficiently different from the mean to represent a problem in real life perception of speech.

There was no significant difference between test and retest scores on either the PIT Q or the PIT N, indicating that increased familiarity with the test on the second testing occasion did not significantly affect the test scores. Further, there was a moderate correlation between test and retest z-scores. Compared to using raw scores to determine test-retest relationships, the use of z-scores lowers the correlation co-efficient by removing the effect of age. Using z-scores is, however, a more valid metric for such analyses.

Test-retest correlations were used to determine the impact of random measurement error. At first sight, the finding that 68% of the variance in test scores for the PIT test in quiet, and 58% of the variance in PIT scores in noise, is accounted by measurement error might make the value of the test seem low. This is not the case. These results were obtained on a typically developing population, so the implication is certainly that a single administration of the test is not adequate to reliably order children from such a population in terms of their categorical perception along the /ba-da/ continuum. Importantly, however, it is well known that the correlation between any two measures containing random measurement error increases as the true range of values present in each of the measures increases. Thus, were children with markedly lower than normal categorical perception ability to be included along with typically developing children, the test-retest correlation would correspondingly increase.

The purpose of the test is, of course, to identify children with categorical perception ability outside the normal range. The normal ranges identified in this experiment, extending somewhat arbitrarily from two standard deviations below the mean to two standard deviations above it, already include the variance due to the estimated measurement error. Consequently, when a child is found to perform poorer than approximately two standard deviations below the mean, we can be confident that the child has categorical perception ability poorer than is typical for his or her age. Because the test includes an inbuilt calculation of the confidence interval surrounding the test result, it should often be possible to identify children whose results have been influenced by inconsistent responses, and for whom retesting or additional assessments would be appropriate.

It is evident from Figures 5 and 6 that the greatest variation in performance occurred in the youngest children. Confidence intervals were calculated on the slope parameter used to calculate the uncertainty region and results were deemed invalid for individuals whose confidence intervals were substantially wider than those of the remaining children. Of the

children excluded, the majority were in the six year-old age group. In clinical trials currently in progress participation has been restricted to children 7 yrs, 6 mths to 11 yrs, 6mths. If it is found during these studies that the PIT has clinical validity in older children we may re-examine performance on 6 to 7.5 year olds using a verbal or pictorial response method whereby the child indicates to the audiologist the phoneme category perceived and the audiologist inputs the data. Younger children may be more motivated to attend to the test stimuli, and less distracted by outside influences, if an authority figure is more actively involved in the test administration.

Finally, even having accounted for any invalid results by excluding children based on UR confidence interval parameters, prior to analysis of the normative data, five outliers were removed from the PIT Q results and four outliers were removed from the PIT N data. Based on the sample size, the number of outliers would be predicted to be around one. Thus the number of children who exhibited decreased precision of categorical perception outside the normal range was unexpected, and may be an indication of a distinct clinical population.

The present study forms just the first steps toward developing a new and innovative test of temporal resolution ability suitable for clinical use. Studies are currently being undertaken in children who present with phonological dyslexia and/or children with suspected CAPD who are experiencing difficulties listening in the classroom. Patterns of performance across a range of standardized assessment tasks and measures of cortical auditory evoked potentials will be documented. These results will be correlated with PIT results, as well as another test – the Parsing Syllable Envelopes (ParSE) test - developed for the study to investigate syllable boundary detection (Cameron et al., submitted). Should these clinical studies support the hypothesised relationships between dyslexia and performance on either the PIT or ParSE tests, then we will recommend addition of these new tests to the clinical APD assessment

battery for children who present with poor reading ability. Such an addition would expand the diagnostic targets for auditory processing assessment noted by (Vermiglio, 2016).

REFERENCES

American Academy of Audiology. (2010) *Clinical practice guidelines: Diagnosis, treatment and management of children and adults with central auditory processing disorder*.

American Speech-Language-Hearing Association. (2005) *(Central) auditory processing disorders [Technical statement]*.

Blomert, L, Mitterer, H. (2004) The fragile nature of the speech-perception deficit in dyslexia: Natural vs synthetic speech. *Brain Lang* 89(1):21–26.

Boersma, P, Weenink, D. (2014) Praat: doing phonetics by computer (version 5.4.04).

Boets, B, Op De Beeck, HP, Vandermosten, M, Scott, SK, Gillebert, CR, Mantini, D, Bulthé, J, Sunaert, S, Wouters, J, Ghesquière, P. (2013) Intact but less accessible phonetic representations in adults with dyslexia. *Science (80-)* 342(6163):1251–1254.

British Society of Audiology. (2011) Position Statement: Auditory processing disorder (APD). *Br Soc Audiol* (March 2011):1–9.

Cameron, S, Chong-White, N, Mealings, KT, Beechey, T, Dillon, H, Young, T. The Parsing Syllable Envelopes (ParSE) test for assessment of amplitude modulation discrimination skills in children: Development, normative data and test-retest reliability studies. *J Am Acad Audiol*.

Cameron, S, Chong-White, N, Mealings, KT, Beechey, T, Dillon, H, Young, T. The Parsing Syllable Envelopes (ParSE) test for assessment of amplitude modulation discrimination skills in children: Development, Normative Data and Test-Retest Reliability Studies. *J Am Acad Audiol*.

Chang, EF, Rieger, JW, Johnson, K, Berger, MS, Barbaro, NM, Knight, RT. (2010) Categorical speech representation in human superior temporal gyrus. *Nat Neurosci* 13(11):1428–1432.

Dillon, H, Cameron, S, Glyde, H, Wilson, W, Tomlin, D. (2012) An opinion on the assessment of people who may have an auditory processing disorder. *J Am Acad Audiol* 105(23):97–105.

Fant, G. (1962) Formant bandwidth data. Speech Transmission Laboratory Quarterly Progress and Status Report. Stockholm: Royal Institute of Technology, 1–2.

Goswami, U. (2011) A temporal sampling framework for developmental dyslexia. *Trends Cogn Sci* 15(1):3–10.

Goswami, U, Fosker, T, Huss, M, Mead, N, Szucs, DD. (2011) Rise time and formant transition duration in the discrimination of speech sounds: The Ba-Wa distinction in developmental dyslexia. *Dev Sci* 14(1):34–43.

King, WM, Lombardino, LJ, Crandell, CC, Leonard, CM. (2003) Comorbid auditory processing disorder in developmental dyslexia. *Ear Hear* 24(5):448–456.

Koerner, TK, Zhang, Y, Nelson, PB, Wang, B, Zou, H. (2016) Neural indices of phonemic discrimination and sentence-level speech intelligibility in quiet and noise: A mismatch negativity study. *Hear Res* 339:40–49.

Liberman, AM, Cooper, FS, Shankweiler, DP, Studdert-Kennedy, M. (1967) Perception of the speech code. *Psychol Rev* 74(6):431–461.

Liebenthal, E, Binder, JR, Spitzer, SM, Possing, ET, Medler, DA. (2005) Neural substrates of phonemic perception. *Cereb Cortex* 15(10):1621–1631.

MathWorks Inc. (2014) MATLAB (Release 2014b). [Computer Software]. Natick,

Massachusetts: The MathWorks Inc.,.

McArthur, G, Kohnen, S, Larsen, L, Jones, K, Anandakumar, T, Banales, E, Castles, A. (2013) Getting to grips with the heterogeneity of developmental dyslexia. *Cogn Neuropsychol* 30(1):1–24.

National Acoustic Laboratories. (2000) Speech and Noise for Hearing Aid Evaluation (CD).

National Acoustic Laboratories. (2015) *NAL Position Statement on Auditory Processing Disorders*.

Phillips, C, Pellathy, T, Marantz, A, Yellin, E, Wexler, K, Poeppel, D, McGinnis, M, Roberts, T. (2000) Auditory Cortex Acceses Phonological Categories: An MEG Mismatch Study. *J Cogn Neurosci* 12(6):1038–1055.

Rosen, S. (1992) Temporal information in speech: Acoustic, auditory, and linguistic aspects. *Philos Trans* 336(1278):367–373.

Serniclaes, W, Sprenger, L. (2001) Perceptual discrimination of speech sounds in developmental dyslexia. *J Speech, Lang Hear Res* 44(2):384–399.

Specht, K. (2014) Neuronal basis of speech comprehension. *Hear Res* 307(378):121–135.

Stevens, KN, Keyser, SJ. (1989) Primary features and their enhancement in consonants. *Language (Baltim)* 65(1):81–106.

Vandermosten, M, Boets, B, Luts, H, Poelmans, H, Golestani, N, Wouters, J, Ghesquière, P. (2010) Adults with dyslexia are impaired in categorizing speech and nonspeech sounds on the basis of temporal cues. *Proc Natl Acad Sci U S A* 107(23):10389–10394.

Vandermosten, M, Boets, B, Luts, H, Poelmans, H, Wouters, J, Ghesquière, P. (2011)

Impairments in speech and nonspeech sound categorization in children with dyslexia are driven by temporal processing difficulties. *Res Dev Disabil* 32(2):593–603.

Vermiglio, AJ. (2014) On the clinical entity in audiology: (Central) auditory processing and speech recognition in noise disorders. *J Am Acad Audiol* 25(9):904–917.

Vermiglio, AJ. (2016) On diagnostic accuracy in audiology: Central site of lesion and central auditory processing disorder studies. *J Am Acad Audiol* 27(2):141–156.

White-Schwoch, T, Davies, EC, Thompson, EC, Woodruff Carr, K, Nicol, T, Bradlow, AR, Kraus, N. (2015) Auditory-neurophysiological responses to speech during early childhood: Effects of background noise. *Hear Res* 328:34–47.

Wilson, WJ, Arnott, W. (2013) Using different criteria to diagnose (central) auditory processing disorder: How big a difference does it make? *J Speech Lang Hear Res* 56(1):63–70.

APPENDIX A

Instructions given to participants prior to undertaking the Phoneme Identification Test (PIT) practice, familiarization and test conditions.

Practice (quiet and 0dB SNR conditions)

"When you are ready to start press the play button. You will hear a voice saying either 'ba' or 'da'. Press the button on the screen that matches the sound you hear each time. The left button is for 'ba' and has a picture of a barking puppy. The right button is for 'da' and has a picture of a dark sky. If you are not sure which sound you hear please guess."

Familiarisation (quiet condition only)

"Now you will hear the voice again but the sound will start as 'ba' and then it will start to sound more like 'da'. Press the button that matches what you hear. If you're not sure, just guess."

Test (quiet and 0dB SNR conditions)

Prior to each test condition participants were given the following instructions (including the sentence in parentheses before the 0dB SNR trial only):

"Now you will hear the same type of sounds but some sounds may be less clear. (You will also hear a whooshing noise. Just try to ignore the whooshing noise and listen to the voice.) If you are not sure whether you hear 'ba' or 'da' please choose the sound you think it was more likely to be. Half-way through you can have a break. Try to press the button as soon as you hear the sound but it is more important to be accurate than fast".

FIGURE CAPTIONS

Figure 1. PIT test screen.

Figure 2. Endpoints and midpoint of the 11 step /ba/-/da/ continuum. (a) Ideal /ba/ with a rising F2; (b) an ambiguous token with a level F2; (c) ideal /da/ with a falling F2.

Figure 3. Image of the PIT results screen for a participant 9 yr, 8 mo of age. (a) PIT Q result and (b) PIT N result shows z-scores of -0.09 and 0.05 respectively. The curved lines represent the psychometric functions fitted to the data. The dark-colored central curved line is the curve fitted to all measured data, and the other two lines are the curves fitted to odd and even-numbered responses, respectively. The dashed lines show the upper and lower boundaries of the uncertainty region.

Figure 4. Histograms of uncertainty region z-scores for the (a) 149 participants on the PIT Q and the (b) 148 participants on the PIT N.

Figure 5. Box and whisker plots of uncertainty region (UR) as a function of age for the (a) 149 participants on the PIT Q and the (b) 148 participants on the PIT N. The filled square represents the median UR, the open boxes represent the 25-75% boundaries, the whiskers represent the minimum and maximum scores.

Figure 6. Scatterplots of the individual raw uncertainty region scores for the (a) PIT Q (n = 149) and PIT N (n = 149). The solid line indicates the mean score as a function of age, the dashed line shows the ± 2 SD limits.

Figure 7. Scatterplot of PIT performance in quiet versus noise, measured as z-scores, for the 143 participants who completed both the PIT Q and PIT N. The dashed line represents equal age-adjusted performance on the two tests.

Figure 8. Scatterplots of the uncertainty region scores (z-scores) at test and retest for the (a) 119 children retested on the PIT Q and the (b) 117 children retested on the PIT N. The dashed line represents the least squares regression line.