Improving the understanding and diagnosis of Auditory Processing Disorder (APD) in Children

Shiran Koifman

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy at University College London

March 15 2021

Contents

List of Figures

List of Tables

3

APD study

Contents

3.2 Methods

3.2.1 Participants

Forty-four primary school children who ~~are~~ native British English speakers with normal hearing acuity participated in the study. Amongst them twenty-one belonged to the APD clinical group (5 females) with an average age of 11.04 ± 1.42 years (range: 7.8 - 12.9 years). The remaining twenty-three (12 females) comprised of typically developing control children (TD) with no reported concerns or diagnosis of an auditory, language or other cognitive developmental disorder. The TD group average age was 9.47 ± 1.58 years and ranged between 7 to 12.1 years. Since not all the measurement equipment was easily portable and in order to maintain the same environment during the assessment across the complete sample, the children and their caregivers were required to travel to central London for the testing. In order to maximise the number of children taking part in the study, 8 out of the 23 TD children (35%) had an APD sibling which took part in a parallel study that took place on the same day of testing. All the children who participated in the study were required to have normal hearing acuity, defined as thresholds $\leq$ 25 dB HL at the octave frequency bands between 0.25 to 8 kHz and their eardrum had to be visible, healthy and intact in both ears following otoscopic inspection. One APD participant was excluded from the analysis due to raised thresholds predominantly in the right ear, ranging between 30 to 45 dB HL ($\mathrm{PTA}_{Right} = 36.25$ dB HL; $\mathrm{PTA}_{Left} = 13.75$ dB HL), thus resulting in a final APD group size of twenty. Otoscopic inspection of the child's ear canal revealed a large accumulation of cerumen in both ears with an occluded right ear. Two additional children (x1 APD, x1 TD) had slightly raised thresholds at 8 kHz in one ear of 35 and 30 dB HL, respectively. However, since thresholds at all other frequency bands were well within the $\leq$ 25 dB HL criteria they were not excluded.

APD children were recruited in two ways. Children diagnosed with APD at Great Ormond Street Hospital (GOSH) or at the London Hearing and Balance

Centre (LHBC), London, UK, were identified based on their clinical records and were contacted by a clinical team member. The caregivers were provided with information about the study and means of contact to express interest in participation. Others, including the TD group were recruited by advertisements on social networks (e.g., APD Support UK Facebook group), science events, local information boards and UCL staff newsletter email, where parents were requested to fill-out an online interest form with short screening questions to ensure that the child met the participation requirements. Most of the children in the APD group (85%, 17/20) were reported to undergo an APD assessment at GOSH, about a third were directly recruited from the clinic. The remaining three were reported to be assessed at LHBC, at the University of Southampton Auditory Implant Service ~~and~~ Chime Audiology Royal Devon & Exeter Hospital (screening only).

Our initial aim was to take a conservative stance on inclusion criteria by including only those who met a clinical APD criteria (2 SD below the norms on two or more tests during the assessment). Moreover, being aware of the high prevalence of APD children with additional co-occurring developmental disorders, we strived to recruit children who display a "pure" form of APD without reported diagnosis or concerns for additional developmental disorder/s. However, very few APD children met these strict criteria~~,~~ ~~only~~ 75% (15/20) met the clinical criteria of APD, out of which 60% (9/15) ~~where~~ diagnosed with spatial processing disorder (SPD) due to abnormal SRM in the LiSN-S task (see Table 3.1 for descriptives of the APD group). Of the remaining children in the APD group, four did not meet the diagnos~~is~~ criteria for various reasons (e.g., young age, lack of psychological educational evaluation report and the need to exclude other deficits), however their assessment report acknowledged some "auditory processing difficulties", whereas the fifth child awaited an APD assessment following an APD screening. Due to the small sample-size these children were included in the APD group for the analysis, nevertheless, they were subdivided as children with Listening Difficulties (LiD) and differences in performance LiD and APD children were later explored. Furthermore, half of the

Table 3.1: APD group demographics and APD-related history background.

School type	85% (17/20) Mainstream, (1 child in a special ASD unit, 2 in a private school), 15% (3/20) non-mainstream school
Assessment location	85% (17/20) GOSH, 15% (3/20) other
APD Diagnosis	75% (15/20) APD, 25% (5/20) LiD
SPD subtype	60% (9/15) SPD
Additional disorder (diagnosed)	50% (10/20) secondery developmental disorder/s
Additional disorder (undergoing assessment)	25% (5/20)
MEHx	60% (12/20)
PET history	25% (5/20)
FM-device usage	55% (11/20)
Auditory training	35% (7/20)

MEHx: History of middle ear problem
PET: Pressure equalisation tube

APD group (10/20) were reported for being diagnosed with one or more secondary developmental disorder/s (x6 Dyslexia, x3 HF-ASD, x3 DLD, x1 ADHD, x1 ADD, x1 Dyspraxia, x1 visual stress, x1 sensory integration disorder, and x1 poor short-term working-memory). Nonetheless, several caregivers reported that their motivation for seeking additional diagnosis was to get more help from the school, rather than a real concern, after feeling that their support for their child's APD was lacking.

Caregivers from both groups completed a comprehensive background questionnaire, similar to the one that is typically given prior to an APD assessment, concerning the caregiver/s educational level, child and family history of hearing, listening problems and developmental disorders, child history of otitis media with effusion (OME), pressure-equalisation tubes (PET / grommets), pregnancyrelated questions (e.g., complications, prematurity, etc.), APD-related (e.g., date of diagnosis, location, use of FM device and auditory training), any diagnosis or concerns regarding the child's speech, language, educational and/or cognitive skills, speech and language therapy, medication taken, musical training and the type of school the child attends.

Children in the APD group were on average 1.5 years older than children in the TD group. Difference in age between the two groups was tested with a one-way ANOVA using *anova()* function (parametric assumption of normal distribution and homoscedasticity were met). The test revealed a significant difference in age between the groups [$F(1,41) = 11.58$, $p < 0.01$]. Nonetheless, since age is often reported as a strong ~~indicator~~ for performance in other similar behavioural studies, analysis of the results obtained in the current study was conducted for age-independent scaled scores and should not affect the comparison between the two groups. The project was approved by the UCL Research Ethics Committee (Project ID Number 0544/006) and the NHS Health Research Authority (REC reference: 18/LO/0250). The testing commenced once an informed consent was given by both the caregiver and the child.

3.2.2 Measurements

The test battery used in the present study is described in the following section and summarised in Table 3.2.

Auditory evaluation

Standard & extended high-frequency (EHF) audiometry

Otoscopic inspection was performed prior to the audiometric test to ensure the ear was clear from cerumen and to avoid harming the eardrum when inserting the ear probe. Both standard and extended high-frequency (EHF) audiometry thresholds were measured using the Hughson-Westlake manual procedure, starting from 1 kHz. Standard air conduction pure-tone audiometry was carried out at six octave frequency bands ranging between 0.25 to 8 kHz using ???? audiometer and

Table 3.2: Summary of the study test battery.

Task	Information	Measure
Standard & extended high-frequency (EHF) audiometry	Pure-tone~~s~~ detection thresholds measured at the octave frequency ~~bands~~ between 0.25 and 8 kHz (standard), and 8 to 20 kHz (EHF).	Detection threshold in dB HL
Switching task (ST)	Adaptive speech-on-speech listening task that involves perception of interrupted and periodically segmented speech that is switched between the two ears out-of-phase with an interrupted distractor. ST assesses the ability to switch attention and integration of binaural information.	Proportion of speech required to understand 50% of the keywords, Speech Reception duty cycle Threshold (SRdT)
Listening in Spatialised Noise Sentences UK (LiSNS-UK)	Locally developed version of the LiSN-S (Cameron & Dillon, 2007), an adaptive speech-on-speech listening task that assesses the ability to use spatial release from masking (SRM), measured as the difference in perception between collocated and separated speech distractors.	Signal-to-noise-ratio (SNR) yielding 50% speech intelligibility, Speech Reception Threshold (SRT)
Speech-shaped-noise (SSN)	Conventional adaptive speech in noise task that asseses speech perception of ASL sentences (MacLeod & Summerfield, 1990) in a speech-shaped-noise with a spectrum matched to the ASL material.	SRT
The Environmental Auditory Scene Analysis task, ENVASA (Leech et al., 2009)	Non-linguistic self-administered task involves detection of everyday environmental sounds presented in naturalistic auditory scenes and can be used to asses IM effects as well as sustained selective auditory attention skills.	%-correct
Recalling sentences, CELF-RS (Wiig et al., 2017)	A subtest from the Clinical Evaluation of Language Fundamentals UK 5^{th} edition (CELF-5-UK)~~,~~ assess expressive language skills, measured by the ability to repeat in verbatim sentences with varying length and complexity. Standardised for children aged 5 to 16 years.	Age-corrected scaled scores
The Evaluation of Children's Listening and Processing Skills, ECLiPS (Barry & Moore, 2014)	Standardised questionnaire comprise~~s~~ of 38 statements grouped into five categories designed to identify listening and communication difficulties in children aged 6 to 11 years. Respondent~~s~~ agreement is expressed using a five-point Likert scale ("*strongly agree*" - "*strongly disagree*").	Age-corrected scaled scores
The Children's Communication Checklist 2^{nd} edition, CCC-2 (Bishop, 2003)	Standardised questionnaire comprisng ~~of~~ 70 items designed to screen language and/or communication problems in children aged 4 to 16 years. Item ~~comprises~~ of a behaviour statement (e.g., "*Mixes up words of similar meaning*") with respondents asked to judge how often the behaviours occur using a four-point Likert scale (0-3).	Age-corrected scaled scores

??? headphones.

Extended high-frequency pure-tone detection thresholds were measured at four octave frequenc~~y bands~~ 8, 11, 16, & 20 kHz using locally written MATLAB based software which generated the stimul~~us~~ and collected the data. Target tones were pulsed (3 repetitions) with a duration ~~time~~ of 700 ms and 50 ms rise/fall time. EHF measurements took place in a ~~designated~~ sound attenuated chamber with the child sitting in the centre of the chamber while the examiner was situated outside. Communication during the testing was carried out via a video-audio intercom system. The child was instructed to raise his/her~~s~~ hand each time s/he heard a tone. The MATLAB script was executed using a Windows PC which which was connected via USB to an RME FireFace UC sound card (Audio AG, Haimhausen Germany) and an ER10X Extended-Bandwidth Acoustic Probe System (Etymōtic Research, Elk Grove Village, IL, USA). ~~Stimulus was~~ presented via an otoacoustic emission probe with silicon tips in variable sizes (between 8 to 13 mm), depending on the size of the child's ear.

Standing waves in the ear canal produces spatially non-uniform sound pressure at frequencies above 2-3 kHz, introducing calibration errors when estimating the sound pressure level arriving at the eardrum (Lee et al., 2012; Richmond et al., 2011; Siegel, 1994). Together with other factors such as individual variations in the ear canal length and differences in depth in which the ear probe is inserted into the ear canal, these factors can introduce up to 20 dB calibration error (Siegel, 1994). To account for that, a sound pressure level calibration was applied using ARLas MATLAB-based software package (Goodman, n.d.), using a similar technique as described by Lee et al. (2012). *accurate??* The first half-wave resonance of the ear canal was measured for each frequency using chirp noise, estimating the distance between the ear-probe and the eardrum. The target stimulus was then ~~filtered~~ to the desired output level that corresponds to 0 dB HL using the in-situ calibration forward-pressure level data (FPL) and EHF-specific weighting thresholds (in dB SPL) measured across 84 NH listeners aged 10 to 21 years (see Table 1 in Lee et al., 2012).

Switching task (ST)

Estimating the effect of IM while minimising peripheral EM on speech perception was measured using the switching task (ST) which is believed to assess the listeners ability to switch attention and integrat~~ion~~ of binaural information. The ~~exact~~ same test procedure and equipment was used as described in Chapter 1. Listeners were presented with both test versions using the ASL and the CCRM speech material. As for the stimuli, the ASL target sentences, spoken by a single male talker, were taken from the final sentences selected following the normalisation study. In addition, a level correction was applied to each sentence using the sentence-specific weighing factors estimated in the normalisation study (see Chapter 2). The first five test lists out of the eight phonetically-balanced normalised test lists (~~à~~ 25 sentences each) were used, whereby their order was quasi-randomised to account for order, masker combinations, and fatigue effect. The target CCRM sentences were the same as described in Chapter 1, spoken by three different male talkers~~,~~ were selected at random every trial and always began with the priming animal 'dog'. The target speech material was presented either without a distractor (Quiet) with and without switching (NoAlt / Alt) or with a distractor. A selection of four distractors were used (see Chapter 1 for detailed description): English (ENG_F) and Mandarin (MDR_F) unrelated connected-speech, each spoken by ten different female talkers, and a non-speech amplitude-modulated speech-spectrum-noise (AMSSN) with the envelope of a single talker out of 40 talkers (20 females). The fourth distractor was presented only with the CCRM speech material and comprised of CCRM target-like sentences (CCRM_F) with a different priming animal, colour and digit, spoken by ten different female talkers. Each participant was presented with a total of 11 runs, one for each test condition, with 5 conditions for the ASL (Quiet-NoAlt,Quiet-Alt, MDR_F-Alt, ENG_F-Alt), and 6 for the CCRM (with the additional CCRM_F-Alt condition). Testing started following a practice phase, where four trials of each of the eleven test conditions were presented. Practice runs started at an easy-to-moderate DC ~~rate~~ of 0.8 in order to expose the listeners to the adaptive procedure. In addition, every test run started with two practice sentences (initial DC = 0.97) to orient the listeners to the test condition that ~~is~~ about to be presented.

Listening in Spatialised Noise Sentences UK (LiSNS-UK)

The locally developed Listening in Spatialised Noise Sentences UK (LiSNS-UK) assesses the ability to use binaural cues in speech-on-speech listening conditions. The test development, speech material normalisation, and norms standardisation followed Cameron and Dillon (2007) ~~development steps~~ and are described in detail in Chapter 2. The test uses virtualisation techniques to create spatial distribution of sound sources in space for headphone~~s~~ presentation where target sentences (ASL; MacLeod & Summerfield, 1990) are presented in two simultaneous speech distractors (unrelated children’s stories spoken by the target talker). The LiSNS-UK comprises ~~of~~ two main listening conditions, differing in their availability of spatial cues. The target sentences are configured to always appear in front of the listener~~’s head~~, at 0° azimuth on the horizontal plane, with the two streams of speech distractors either collocated in space with the target (S0N0), resulting in relatively poor speech perception, or offset in space, with one distractor to either side of the target at ± 90°. The spatial separation in the later condition results in an improvement in speech perception of circa 13 dB (Cameron et al., 2011), typically termed as spatial release from masking (SRM). This SRM advantage is calculated by taking the difference between performance in the collocated and the separated condition.

Speech distractors were presented continuously throughout a run at a fixed 65 dB SPL output level and comprised of a combination of two out of three available passages. A 1-up/1-down adaptive procedure was used, varying the level of the target talker relative to the distractors depending on the listener’s response to measure their speech reception threshold (SRT), i.e., the signal-to-noise-ratio (SNR) yielding 50% speech intelligibility. A 2 ms long reference cue (1 kHz pure-tone) was presented 500 ms before the target sentence onset at 65 dB SPL. The initial target output level was 75 dB SPL for the collocated condition and 70 dB SPL for the separated condition with an initial step-size of 4 dB SNR. The step-size was reduced after ~~every~~ reversal, reaching a minimum step-size of 2 dB SNR ~~after three practice reversals~~. The adaptive procedure ended once all 25 test trials were presented and stopped in case a maximal output level of 89 dB SPL was reached more than three times. Nonetheless, such event did not occur in the present study. Since each listener was only presented once with each condition, it was decided not to introduce any other stopping rules that could have expedited the testing time but may as well introduced an estimation error for the SRTs in some cases. The SRT was calculated by averaging the test reversals SNRs, whereby test reversals were defined as any reversals following three practice reversals.

The order of the listening condition, test lists, sentences within a run, and distractor~~s~~ combinations was fixed across all the participants and started with the collocated condition. Each test list consisted of 25 sentences taken from the 8-phonetically-balanced ASL test lists which were constructed following the normalisation study ~~and~~ a sentence-specific level correction ~~was applied~~ (see Chapter 2). Spatialisation was applied by convolving each stimuli with head-related transfer functions (HRTFs) at the corresponding azimuthal direction separately for the left and the right channel. The HRTFs were measured with a Knowles Electronics Manikin for Acoustic Research (KEMAR) with a small pinnae taken from the CIPIC HRTF database[1] (see Algazi et al., 2001, "special" HRTF data). A post-equalisation step was applied in order to flatten the magnitude of the headphone~~s~~ frequency response. Headphone-to-ear Transfer Functions (HpTFs) measured with KEMAR manikin for HD-25 supraaural headphones were extracted from Wierstorf et al. (2011) HRTF database. The final mixed stimulus was filtered with the inverse HpTFs

[1]The database is available online in: https://www.ece.ucdavis.edu/cipic/spatial-sound/hrtf-data/

separately for the left and the right channel before being combined together as a final step. Every participant was presented with two runs, one for each listening condition (collocated / separated). Testing started following a practice phase of two runs, one for each of the test conditions with five BKB sentences each (Bench et al., 1979). Listeners were instructed to verbally repeat the target sentences to the experimenter who was situated alongside in a sound treated chamber. The experimenter scored the response by selecting the correctly repeated keywords on the screen. Listeners were encouraged to guess if unsure while no feedback was given at any time. A loose keyword scoring method was used, whereby errors of case or declension were considered as correct responses, e.g., a repetition of the keywords '<clown**s**> <funny> <face**s**>' to the stimulus 'The <clown> had a <funny> <face>'.

Speech-shaped-noise (SSN)

A speech-in-noise test was used as a more conventional listening task that is widely used in the clinic as opposed to the more complex listening conditions measured by the ST or the LiSNS-UK. The normalised ASL sentences were presented in a speech-spectrum-noise (SSN) with spectrum matched to the ASL corpus. The SSN onset was 500 ms before the target sentence began. The ~~exact~~ same adaptive procedure as for the LiSNS-UK was used with the same stopping-rules and SRT calculation. Each listener was presented with a single run of 25 sentences following a practice phase with seven BKB sentences. The same test list and sentences order was used across all the listeners.

The Environmental Auditory Scene Analysis task (ENVASA)

In analogy to the classic 'cocktail-party' scenario, ENVASA is a non-linguistic paradigm (Leech et al., 2009) that measures detection of everyday environmental sounds presented in naturalistic auditory scenes and can be used to asses IM effects as well as sustained selective auditory attention skills. In the task, short environmental target sounds (e.g., a dog's bark, a door knock, or a bouncing ball) were presented in a dichotic background scene (i.e., the target sound is presented only in one ear), consisting of either a single background scene, presented in both ears, or two background scenes, each presented in a different ear. The number of targets, the onset time and the ear of presentation varied across trials. Four SNRs were employed split into two categories ‘low’ (-6 and -3 dB) and ‘high’ (0 and +3 dB). Target-background contextual agreement was manipulated by embedding the target sound in a *congruent* background scene that is in agreement with the listener’s expectations (e.g., a cow’s ‘moo’ in a farmyard scene) or in an *incongruent* background scene which violate these expectations (e.g., a cow’s ‘moo’ in a traffic scene). A schematic illustration of a single test sequence is shown in Figure 3.1.

The experiment was carried out using the original setup as described by Leech et al. (2009). Sounds were presented via Sennheiser HD-25 headphones (Wedemark, Germany) and the participants response was recorded using ???? gamepad. The output level was adjusted to a comfortable level before the test started. The participants were situated in front of a laptop and were instructed to hold the gamepad. Prior to the test, the listeners were presented with a short child-friendly demonstration video with audio instructions. Next, a short recap was given verbally by the examiner and an exemplary trial was simulated together with the child to ensure that the child fully understood the task’s instructions. The task began with three short practice trials with provided feedback, while no further feedback was given during the test phase.

Every trial was made of two parts, starting with a target audio and visual familiarisation phase before the main target detection phase. Target identification was recorded by pressing one of the three buttons on the gamepad which corresponded to the location of the target objects on the screen. A response was counted as correct only if the participants pushed the corresponding button within 2 seconds, 300 ms after the target onset. The outcome measure was calculated as the percentage of target sounds correctly identified within a condition (%-correct). In total there were 115 target sounds presented over 40 trials, where 46 target sounds were presented in

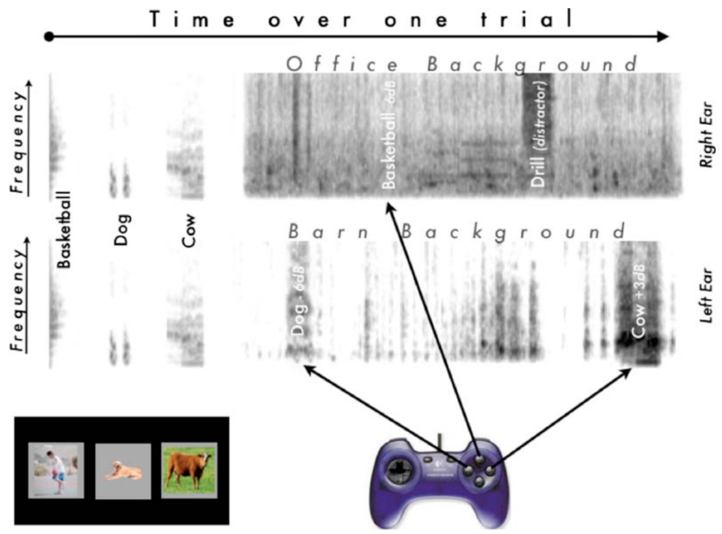

Figure 3.1: Schematic of the ENVASA experimental paradigm (taken from Leech et al., 2009)

a single background condition and another 46 in a dual-background condition. The 23 remaining target sounds served as foil items which were played at 0 dB SNR without a corresponding picture on the screen. The order of the foil items was quasi-randomised and was used to estimate the quality of the participants performance.

CELF-RS

The Recalling Sentences (RS) sub-test of the Clinical Evaluation of Language Fundamentals UK fifth edition (CELF-5-UK; Wiig et al., 2017) was administered to assess the listeners expressive language skills, measuring the ability to repeat in verbatim sentences with varying length and complexity. Standardised norms are available for children aged 5 to 16 years. The CELF-RS is simple and quick to administr~~ate~~ and has been shown to be a good psycholinguistic marker for children with Developmental Language Disorder (DLD) and to provide high levels of sensitivity and specificity (Conti-Ramsden et al., 2001), thus making it a good screening tool. Scoring ~~were marked~~ by hand by the examiner as instructed by the test manual. The sentences were presented using a local MATLAB program via headphones using the same experimental equipment as listed above at a comfortable output level of 70 dB HL. The sentences were spoken by a female speaker with a standard southern British English accent and were recorded in a sound-treated recording booth at the Speech, Hearing and Phonetics Sciences (SHaPS), UCL laboratory, London. The task began with two practice sentences while the number of test items varied depending on the child's age and performance. No repetitions or feedback ~~was~~ given during the testing and the test was discontinued in case the child failed to score any points for four consecutive items. Age-scaled score were calculated based on the test norms with a mean score of 10 and SD of 3. Scaled scores within $\pm$ 1 SD from the norms mean (between 8 to 12) are classified as average scores, whereas performance beyond $\pm$ 1 SD are classified as above / below the average score, with scaled-scores < 7 considered as abnormally poor.

Questionnaires

The Evaluation of Children's Listening and Processing Skills (ECLiPS)

The ECLiPS questionnaire (Barry & Moore, 2014) comprises ~~of~~ 38 items, where the respondents are asked to express their agreement on simple statements about the child's listening and other related skills or behaviours using a five-point Likert scale (from "strongly agree" to "strongly disagree"). The ECLiPS was design to identify listening and communication difficulties in children aged 6 to 11 years. Nonetheless, in their evaluation study, Barry and Moore (2014) found little to no age effect in many of the scale items, suggesting that testing age could be extended below and beyond the population used for the development. Based on factor analysis the items were grouped into five subcategories: 1. Speech & Auditory Processing (SAP), assessing ability to interpret speech and non-speech input, 2. Environmental & Auditory Sensitivity (EAS), estimating the ability to cope with environmentally challenging conditions, 3. Language, literacy & laterality (L/L/L), assessing different abilities that are known to be coupled with language and literacy difficulties, 4. Memory & Attention (M&A), covering short-term and serial memory as well as attention, 5. Pragmatic & Social skills (PSS), assessing pragmatic language or non-normative social behaviours. Aggregated measures were calculated for *Listening* (SAP, M&A, & PSS), *Language* (L/L/L & M&A), *Social* (PSS & EAS), and a *Total* aggregate, calculated by taking the mean of scores across all the sub-scales. Individual age- and sex-scaled scores were computed using the test excel scorer. A score below the 10^{th} percentile (corresponding to a scale score of circa 6) is generally considered clinically significant.

The Children's Communication Checklist 2^{nd} edition (CCC-2)

Communication abilities were assessed using the Children's Communication Checklist second edition questionnaire (CCC-2; Bishop, 2003) which is designed to screen communication problems in children aged 4 to 16 years ~~and~~ comprises ~~of~~ 70 checklist items each ~~comprising~~ of a behaviour statement, like "*Mixes up words of similar meaning*". The respondents are asked to judge how often the behaviours occur using a four-point Likert rating scale: 0. *less than once a week (or never)*, 1. *at least once a week, but not every day*, 2. *once or twice a day*, 3. *several times (more than twice) a day (or always)*. The items are grouped into ten sub-scales of behaviours tapping into different skills (A. Speech, B. Syntax, C. Semantics, D. Coherence, E. Inappropriate initiation, F. Stereotyped language, G. Use of context, H. Non-verbal communication, I. Social relations, J. Interests). Taking the sum of scores for the sub-scales A to H are used to derive the General Communication Composite (GCC) which is used to identify clinically abnormal communication competence. A GCC score < 55 was found to well ~~separate~~ between control and clinical groups, ~~identifying children with~~ scores at the bottom 10% (Norbury & Bishop, 2005). Another proposed composite~~,~~ is the SIDC (Social-Interaction Deviance Composite) which is calculated by taking the difference in sum of subscales E, H, I, and J (tapping into pragmatic language and social skills) from the sum of scales of A to D (describe~~s~~ structural language skills). Abnormal GCC (< 55) combined with a negative SIDC score has been shown to be indicative of an autistic spectrum disorder profile (Bishop, 2003). The CCC-2 scaled and composite scores were computed using the test scorer.

3.2.3 Procedure

Testing took place at the SHaPS laboratory (UCL, London) in a sound-attenuated chamber. Unfortunately, since many of the APD children had to travel from outside London and because of difficulties in recruitment, all the testing had to be ~~made~~ in a single session, lasting in total circa 2.5 to 3 hours (including breaks). To minimise possible fatigue effect, the session was carefully designed to ensure several planned and unplanned ~~brakes~~. The participants were encouraged to request ~~for~~ a break between test runs whenever they required and were observed for any signs of fatigue by the examiner. The different tasks were gathered into short blocks and different measures were scattered throughout the session to keep the session fun and engaging for the child. At the end of the session, each child received a certificate and an Amazon voucher as a token of appreciation for taking part in the study ~~and~~ travel costs of the family were reimbursed.

Participants from both the TD and the APD group completed the same test battery in the below listed order (see Table 3.3). Th ECliPS, CCC-2 and the locally compiled background questionnaire were completed by the caregiver during the testing day. The session started with a standard pure-tone audiogram and otoscopy to ensure that detection thresholds fulfil the study criteria and that there ~~are~~ no abnormalities in the ear canal and the eardrum. Next, the switching task was conducted. Since performance in the task was one of the main focuses in the study, and because little is known about any possible learning effect in the task, presentation of the two speech materials (ASL and CCRM) was counterbalanced within each group, where about half of the children started with the ASL and the other half with the CCRM speech material. In between the two ST versions, each child completed the CELF-RS and the SSN task, whereby again, the order of presentation was counterbalanced within each group. Since both CELF-RS and SSN

Table 3.3: Experimental design and measurements order.

Order	Group A	Group B	Group C	Group D
1	Otoscopoy	Otoscopoy	Otoscopoy	Otoscopoy
2	Standard audiometry	Standard audiometry	Standard audiometry	Standard audiometry
3	ST-ASL	ST-ASL	ST-CCRM	ST-CCRM
4	CELF-RS	SSN	CELF-RS	SSN
5	ST-CCRM	ST-CCRM	ST-ASL	ST-ASL
6	SSN	CELF-RS	SSN	CELF-RS
7	EHF audiometry	EHF audiometry	EHF audiometry	EHF audiometry
8	ENVASA	ENVASA	ENVASA	ENVASA
9	LiSNS-UK	LiSNS-UK	LiSNS-UK	LiSNS-UK

test duration are relatively short, they served as a short informal break between the ST test versions and kept the child engaged. Next, about half-way through the session, with a fixed order, all the participants were presented with the EHF audiometry, and the ENVASA task. The session was concluded with the LiSNS-UK, in-line with typical clinical assessment where the test is often presented last.

3.2.4 Data Analysis

All the data extraction, management and analysis in the present study was computed in R environment (Version 4.0.3; R Core Team, 2020b) using RStudio (Version 1.4.938; RStudio Team, 2019).

Age scaled scores

Age-independent scores were estimated using a linear regression model. The model was fitted per condition separately for each measure (ST-ASL, ST-CCRM, LiSNS-UK, SSN, & ENVASA) and was based on the control group data only with the respective test raw scores (e.g., SRdT, SRT or %-correct) as a dependent variable and age as a predictor. A two-steps model comparison was performed to test the assumption that performance displays a monotonic linear relationship with age versus a non-monotonic (segmented) linear relationship. Extreme outliers were initially trimmed from the TD group to reduce noise in the data and to improve the models fit. In the first step, both models were computed and the best model was selected based on F-statistic model comparison based on analysis of variance ANOVA, using *anova()* function. Standard residuals were next calculated for each TD listener, based on the selected model prediction. Since age was included in the model, the standardised residuals are age-independent and are comparable to z-scores for data with normal distribution, with a mean and SD of approximately 0 and 1. Since the main goal of the study was to find a measure that is able to well separate between the APD group and the typically develop~~ed~~ control group, individual differences and group differences were explored using a deviance analysis procedure proposed by Ramus et al. (2003). Abnormal scores were defined by a two-tailed deviance cut-off of ± 1.96 SD from the TD group mean. Thus, circa 95% of the normal population residuals are expected to be within the deviance range of ± 1.96. Occasional occurrence of abnormal scores in the normal population is not unusual in behavioural measures. Therefore, since the prediction of the residuals is based on the control data, such outliers may skew the TD group true mean or SD and thus may introduce an error in the model prediction. Therefore, in the second step, additional TD outliers (with standardised residuals below/above TD mean ± 1.96) were trimmed from the data and the two models were refitted and compared again. Finally, the model with the best fit was selected and was used to calculate the standardised residuals for all the listeners, including the trimmed TD observations and the APD group.

Statistical analyses

Residual analysis was performed separately for each measure to determine whether the data fulfils parametric methods assumptions of normal distribution using Shapiro-Wilk test (*shapiro.test()*, R Core Team, 2020b) and homogeneity of variance using Levene's test (*leveneTest()*; Fox & Weisberg, 2019). Consequently, statistical analyses for factorial design data that met these requirement was performed using linear mixed-effects regression models (LMEMs). LMEM was fitted using the *lmer()* function (lme4 package; Bates et al., 2015). Backward model selection procedure was applied to find the model that gives the best fit using a likelihood ratio test

(χ^2). Main effects and interaction terms were tested by comparing predictions of the full model to a reduced model where each fixed term was separately removed, starting with the interaction terms. When applicable, post-hoc paired comparison t-test ~~was~~ performed on the fitted model and included adjusted least-squared-mean for the random intercepts (subjects) using the *lsmeans()* function from the emmeans R package (Lenth, 2020). In addition, group differences for a single parametric measure such as in the CELF-RS and the CCC-2 total score were examined using one-way analysis of variance, ~~ANOVA test~~ [*anova()* function]. Post-hoc pairwise comparison t-tests with Bonferroni correction ~~was~~ computed using *pairwise_t_test()* function (rstatix package; Kassambara, 2021).

Nonparamertic data ~~was~~ analysed using *nparLD()* function (nparLD package; Noguchi et al., 2012) which is a robust rank-based method for analysis of skewed data or for data with outliers or from a small sample size (see Feys, 2016, for a good introduction ~~on~~ robust nonparametric techniques). Th~~e~~ function enables different types of nonparametric tests for factorial design data with repeated measures with variable between-/within-subjects factors. The results reported in the present study were based on the ANOVA-type statistic test (ATS) output. Inspection of the ENVASA task age-independent z-scores revealed that the assumption of sphericity (Mauchly's test) was violated. Therefore, analysis was performed using *npIntFactRep* package (Feys, 2015), which is another robust aligned rank technique that enables sphericity correction (Greenhouse-Geisser). When applicable, post-hoc pairwise comparisons were computed using Wilcoxon rank-sum test which is a t-test equivalent for non-parametric data using *wilcox_test()* function either from the rstatix package (Kassambara, 2021) or the coin package which also enables permutation (Hothorn et al., 2006). Group differences for the ECLiPS total score were examined using a robust one-way ANOVA with trimming means (20%) and bootstrapping (N = 2000) using *t1waybt()* function from the WRS2 package (Mair & Wilcox, 2020)~~. F~~ollowed with a corresponding post-hoc test with the same trimming and bootstrapping using *mcppb20()* function from the same package.

Perhaps the way I use *wilcox_test()* needs a further examination.

- I can't get *coin::wilcox_test()* function to run for groups with 3 levels, so used *rstatix::wilcox_test()* instead.
- On the other hand, only *coin::wilcox_test()* worked for 2-way interaction.

3.3 Results

3.3.1 Standard audiometry

The listeners' detection thresholds for the left and the right ear are plotted in Figure 3.2. The shaded grey area represents the TD group thresholds ~~range~~ and the white line represents the group mean at each frequency. The black lines mark~~s~~ the individual thresholds in the APD group and the group mean is marked by the bold black line. The dashed line indicates the maximal threshold~~s~~ criteria of $\leq$ 25 dB HL for participation in the study.

Boxplots of listeners pure-tone~~s~~ detection thresholds measured at six octave frequency~~ band~~s between 0.25 to 8 kHz and their corresponding pure-tone-average (PTA) are shown in Figure 3.3 A-B. Individual~~s~~ PTAs were calculated by averaging thresholds at the frequency bands 0.5, 1, 2 and 4 kHz separately for the right and left ear (PTA_{Right}, PTA_{Left}) and by taking the grand mean for thresholds in both ears (denoted as PTA), whereas the listeners' PTA at the better-ear is denoted as BE. Threshold~~s~~ descriptives by frequency ~~bands~~ and ear split by the two groups ~~is~~ given in Table 3.4, as well as Table 3.6 for PTAs and BE with additional statistics.

Differences between groups (APD, and TD children with/without an APD sibling) for detection thresholds across frequency bands and ears were statistically

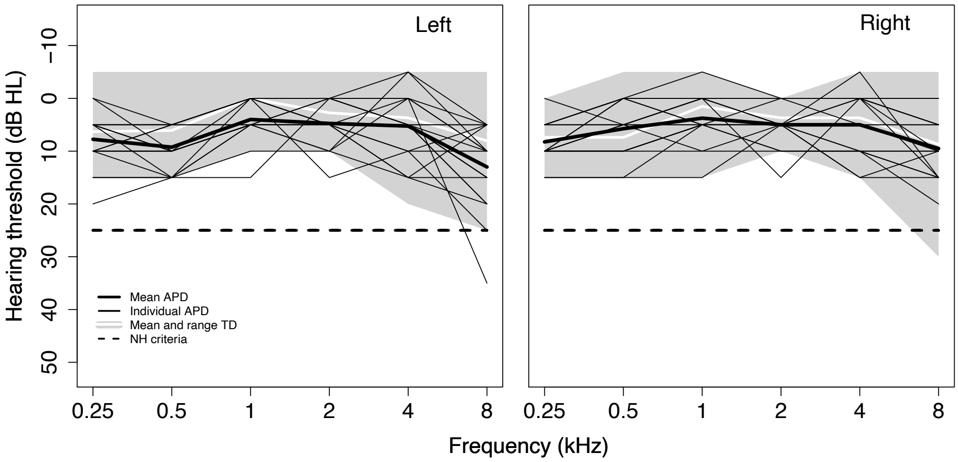

Figure 3.2: Standard audiometry: APD participants pure-tone detection thresholds plotted seperately for the left and the right ear (black lines). The shaded grey area represents the TD group thresholds range and the white line represents the TD group mean at each frequency. The dashed line represents the threshold criteria of hearing level $\leq$ 25 dB HL.

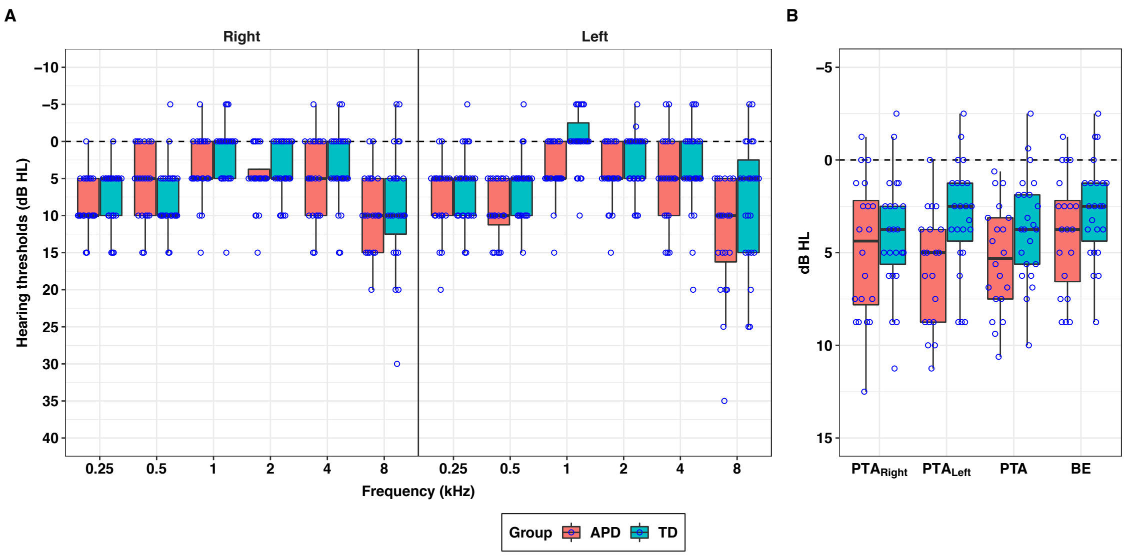

Figure 3.3: Standard audiometry: Pure-tone detection thresholds by frequency ~~bands~~ between 0.25 to 8 kHz (A), and averaged thresholds (B). Individual scores are indicated by circles. The boxes show the data interquartile range (25th-75th percentile) and the horizontal line indicate the median (i.e., 50th percentile). Values that fall within 1.5 times the interquartile range are indicated by the wiskers.

Table 3.4: Standard audiometry: Descriptives for pure-tone detection thresholds (dB HL) by frequency ~~bands~~ (kHz) and ear split by the two groups.

		APD					TD				
Frequency	Ear	N	median	sd	min	max	N	median	sd	min	max
0.25	R	20	10	3.73	0	15	23	5	3.95	0	15
0.5	R	20	5	4.94	0	15	23	10	4.49	-5	15
1	R	20	5	4.55	-5	15	23	0	4.38	-5	15
2	R	20	5	3.97	0	15	23	5	3.76	0	10
4	R	20	5	5.62	-5	15	23	5	5.27	-5	15
8	R	20	10	5.36	0	20	23	10	8.29	-5	30
0.25	L	20	5	4.99	0	20	23	5	5.05	-5	15
0.5	L	20	10	4.06	5	15	23	5	4.25	-5	15
1	L	20	5	3.84	0	15	23	0	3.99	-5	10
2	L	20	5	4.13	0	15	23	0	4.03	-5	10
4	L	20	5	5.95	-5	15	23	5	6.07	-5	20
8	L	20	10	8.01	5	35	23	5	8.49	-5	25

tested with a three-way 6 x 2 x 3 factorial design with repeated measures. Inspection of the data for ~~a~~ linear model residuals revealed that the assumption of normality and homoscedasticity were violated. Therefore, a non-parametric approach was adopted, using a~~n~~ rank-based ANOVA-type statistic test (ATS) with the *nparLD()* function (nparLD package; Noguchi et al., 2012). The ATS test results are given in Table 3.5. There was no significant three-way or two-way interaction between the three predictors, nor a significant main effect of Ear or Group (all p's > 0.05), whereas there was a highly significant main effect of Frequency ($p < 0.001$).

Group differences for PTAs and BE were examined using a 4 x 2 LMEM model (parametric model assumptions were met). Detection measures (PTA_{Right}, PTA_{Right}, PTA and BE) and Group (APD/TD) were set as fixed factors (reference levels: PTA_{Right}, APD group) and detection threshold (in dB HL) as dependent variable, as well as random intercepts for subjects. Note that the TD children ~~was~~ considered as a single group, since there was no significant difference in thresholds across the TD children with or without APD sibling. A model with interaction term was found to give the best fit, showing a significant interaction between the calculated detection

Table 3.5: Standard audiometry: Statistical analysis for the effects of Frequency (0.25 - 8 kHz), Ear (left/right) and Group (APD, and TD with/without an APD sibling) and their interaction (6x2x3 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f2 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and f2 refers to two within-subjects factors (Frequency and Ear).

	Statistic	df	p-value
Group	2.126	1.836	*0.124*
Frequency	18.505	2.861	***< 0.001***
Ear	0.855	1.000	*0.355*
Group:Frequency	0.555	3.900	*0.691*
Ear:Frequency	0.400	3.767	*0.798*
Group:Ear	1.747	1.759	*0.179*
Group:Frequency:Ear	1.659	5.855	*0.128*

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.6: Post-hoc paired comparison t-test for PTA x Group. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercept (subjects) using lsmeans package (Lenth, 2020).

	APD					TD					post-hoc paired t-test					
	N	median	sd	min	max	N	median	sd	min	max	Estimate	SE	Df	t-value	p-value	95%-CI
PTA_{Right}	20	4.38	3.78	-1.25	12.50	23	3.75	3.16	-2.5	11.25	0.799	0.942	59.762	0.848	*0.4*	-1.09 - 2.68
PTA_{Left}	20	5.00	3.04	0.00	11.25	23	2.50	3.01	-2.5	8.75	2.682	0.942	59.762	2.846	***0.006***	0.8 - 4.57
PTA	20	5.31	2.92	0.62	10.62	23	3.75	2.87	-2.5	10.00	1.740	0.942	59.762	1.846	*0.07*	-0.15 - 3.62
BE	20	3.75	3.17	-1.25	8.75	23	2.50	2.68	-2.5	8.75	1.242	0.942	59.762	1.318	*0.193*	-0.64 - 3.13

* significant p-values ($p < 0.05$) are shown in bold.
PTA: average detection threshold (dB HL) at 0.5, 1, 2, & 4 kHz.
BE: PTA at the better ear.

measures and Group [$\chi^2(3) = 12.27$, $p < 0.05$]. Post-hoc paired comparison t-test based on the fitted model ~~was~~ computed using lsmeans function (*emmeans* package, Lenth (2020); see Table 3.6) which revealed a significant difference between the groups for PTA measured in the left ear ($p < 0.05$). However, a group~~s~~ difference of 2.5 dB is rather small and clinically negligible, and is likely to occur due to sampling error. No significant difference was found in the remaining measures (all p's > 0.05).

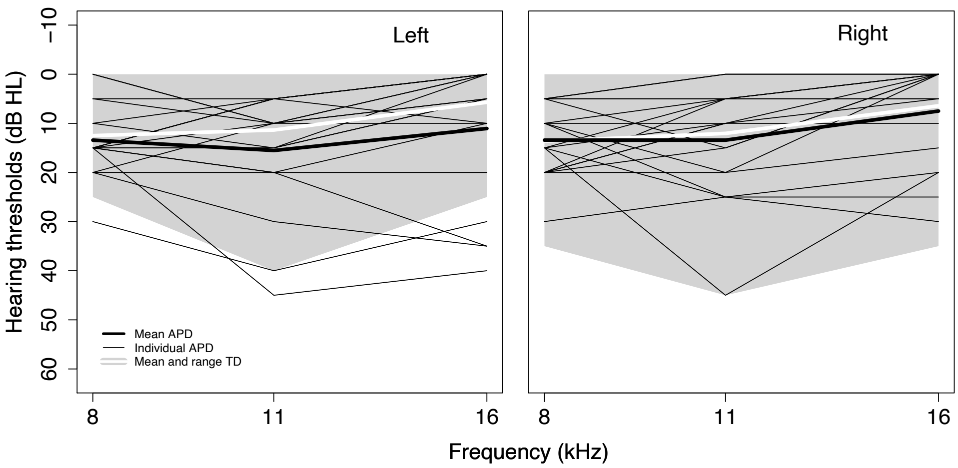

Figure 3.4: EHF audiometry: Pure-tone detection thresholds for extended high-frequenc~~y~~ ~~bands~~ measured in the left and the right ear. The thin black lines represent~~s~~ the individual thresholds in the APD group and the group mean is marked by the bold black line. The shaded grey area represents the TD group threshold range and the white line represents the TD group mean at each frequency.

3.3.2 EHF audiometry

The listeners pure-tone detection thresholds measured at the ~~octave~~ frequenc~~y bands~~ 8, 11 and 16 kHz are plotted in Figure 3.4 separately for the left and the right ear. Again, the thin black lines represents individuals' thresholds in the APD group and the group mean is marked by the bold black line. The shaded grey area represents the TD group thresholds range and the white line represents their mean at each frequency. In many cases it was not possible to recor~~e~~d a reliable response for thresholds measured at 20 kHz, resulting in a large portion of missing data points in both groups. Therefore, thresholds measured at 20 kHz were not included in the analysis. A comparison of the group means reveals relatively small differences in thresholds between the groups, with a relatively larger difference in the left ear, where APD thresholds at 11 and 16 kHz were on average 5 dB higher (i.e., poorer). Boxplots of the listeners thresholds by frequency and ear as well as their calculated PTA~~'~~s and BE are shown in Figure 3.5 A-B. Descriptives of the groups detection thresholds ~~is~~ given in Table 3.7.

Difference in thresholds across group (APD, and TD children with/without

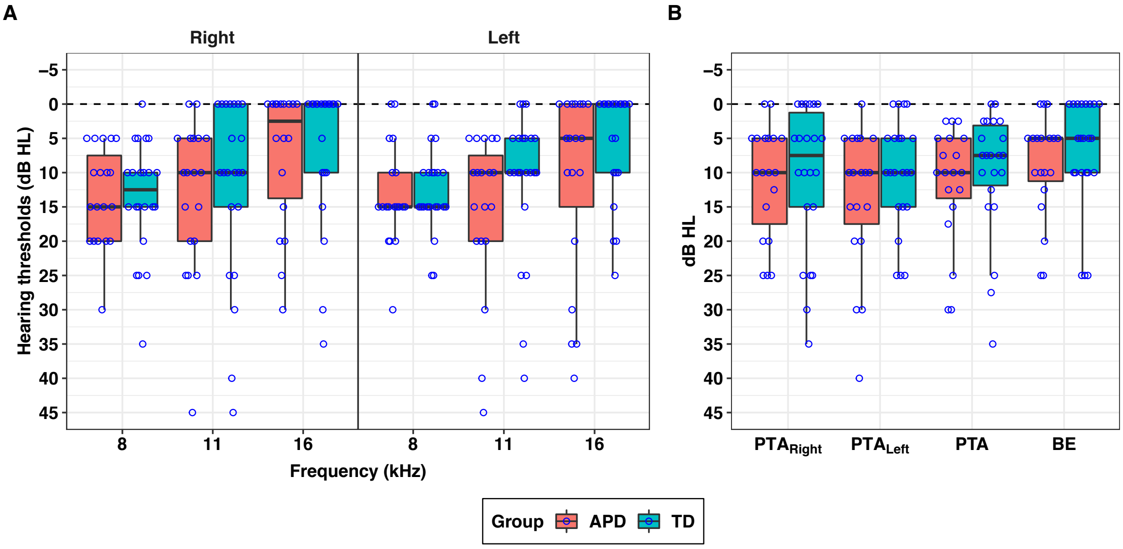

Figure 3.5: EHF audiometry: Boxplots for pure-tone detection thresholds measured at the extended high-frequency bands split by ear and groups (A). Boxplots of the groups averaged PTAs and better-ear BE thresholds are depicted in figure B. Individual scores are indicated by circles.

Table 3.7: EHF audiometry: Descriptive for pure-tone detection thresholds (dB HL) by extended-high frequency ~~bands~~ (kHz) split by ear and group.

		APD					TD				
	Ear	N	median	sd	min	max	N	median	sd	min	max
Octave frequency bands											
8	R	19	15.0	7.08	5.0	30	22	12.5	8.34	0	35
11	R	19	10.0	11.19	0.0	45	22	10.0	13.24	0	45
16	R	19	2.5	10.04	0.0	30	22	0.0	10.51	0	35
8	L	19	15.0	7.27	0.0	30	22	15.0	6.50	0	25
11	L	19	10.0	11.65	5.0	45	22	10.0	10.71	0	40
16	L	19	5.0	13.80	0.0	40	22	0.0	8.11	0	25
PTAs and better-ear											
PTA_{Right}	R	19	10.0	8.27	0.0	25	22	7.5	10.83	0	35
PTA_{Left}	L	19	10.0	10.39	0.0	40	22	10.0	8.09	0	25
PTA		19	10.0	8.59	2.5	30	22	7.5	9.05	0	35
BE		19	5.0	7.60	0.0	25	22	5.0	8.27	0	25

PTA: average detection threshold at 8, 11, & 16 kHz.
BE: PTA at the better ear.

Table 3.8: EHF audiometry: statistical analysis for the effects of Frequency (8 - 16 kHz), Ear (left/right) and Group (APD, and TD with/without an APD sibling) as well as their interaction (3x2x3 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f2 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and f2 refers to two within-subjects factors (Frequency and Ear).

	Statistic	df	p-value
Group	1.124	1.911	*0.323*
Frequency	29.793	1.992	***< 0.001***
Ear	0.226	1.000	*0.635*
Group:Frequency	1.924	3.564	*0.112*
Ear:Frequency	0.150	1.940	*0.855*
Group:Ear	0.167	1.998	*0.846*
Group:Frequency:Ear	0.716	3.638	*0.568*

* significant p-values ($p < 0.05$) are shown in bold.

APD sibling), frequencies (8, 11, & 16 kHz) and ears (left/right) were examined for a 3 x 2 x 3 repeated measures factorial design. Inspection of parametric model assumptions revealed that the assumptions of normality and homoscedasticity were violated. Therefore, the ~~exact~~ same nonparametric procedure as used for standard audiometry was performed using nparLD package. The ATS ANOVA-type test (given in Table 3.8) found no significant three-way nor two way interaction between the different predictors. There was however a highly significant difference in thresholds between the three frequency bands ($p < 0.001$), whereas no significant main effect for Group or Ear was found.

Similarly, additional nonparametric 4 x 2 factorial design model was used to examine the difference between the two groups (APD/TD) for the four combined threshold measures (PTA_{Right}, PTA_{Left}, PTA, and BE). As before, the TD group was treated as a single group since no significant difference was found between TD children with or without an APD sibling. The nparLD ATS test found no significant two-way interaction between Group and Measure nor a main effect of groups, while there was a significant main effect of measure (see Table 3.9).

Table 3.9: EHF audiometry: Statistical analysis for the ~~effects of the listeners~~ calculated ~~measures~~ (PTA_{Right}, PTA_{Left}, PTA, and BE) and Group (APD/TD) as well as their interaction (4x2 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f1 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and a single within-subjects factor (Measure).

	Statistic	df	p-value
Group	0.907	1.000	*0.341*
Measure	7.695	1.389	***0.002***
Group:Measure	0.154	1.389	*0.777*

* significant p-values ($p < 0.05$) are shown in bold.

3.3.3 ST

Outliers & missing data

As a first step, the listeners adaptive track and psychometric functions were manually inspected for abnormalities. The proportion of correct keywords within the final test trials (LevsPC) was calculated as a measure describing the success of the adaptive procedure. Since the adaptive procedure was set to yield 50%-correct of key words in sentences, a successful procedure is expected to have a LevsPC at approximately 50% ~~range.~~ A binomial statistical test was applied to identify observations that significantly differ from 50%. Observations with LevsPC $\leq$ 35% were flagged as possible outliers and were further inspected (see Figure 3.6). Interestingly, most of the flagged outliers belonged to the CCRM material with 29 observations out of 258 (6 conditions x 43 listeners), whereas only 3 observations out of 215 (5 conditions x 43 listeners) were flagged for data measured with the ASL speech material.

As expected, most of the identified cases in both materials were for observations measured with the more demanding conditions with speech distractors. In five cases (2 ASL; 3 CCRM) we were able to confidently determine that the listener's true score was near to ceiling, and thus these observations were set to the maximal DC in the task (0.97). In other cases it was not possible to confidently determine the true

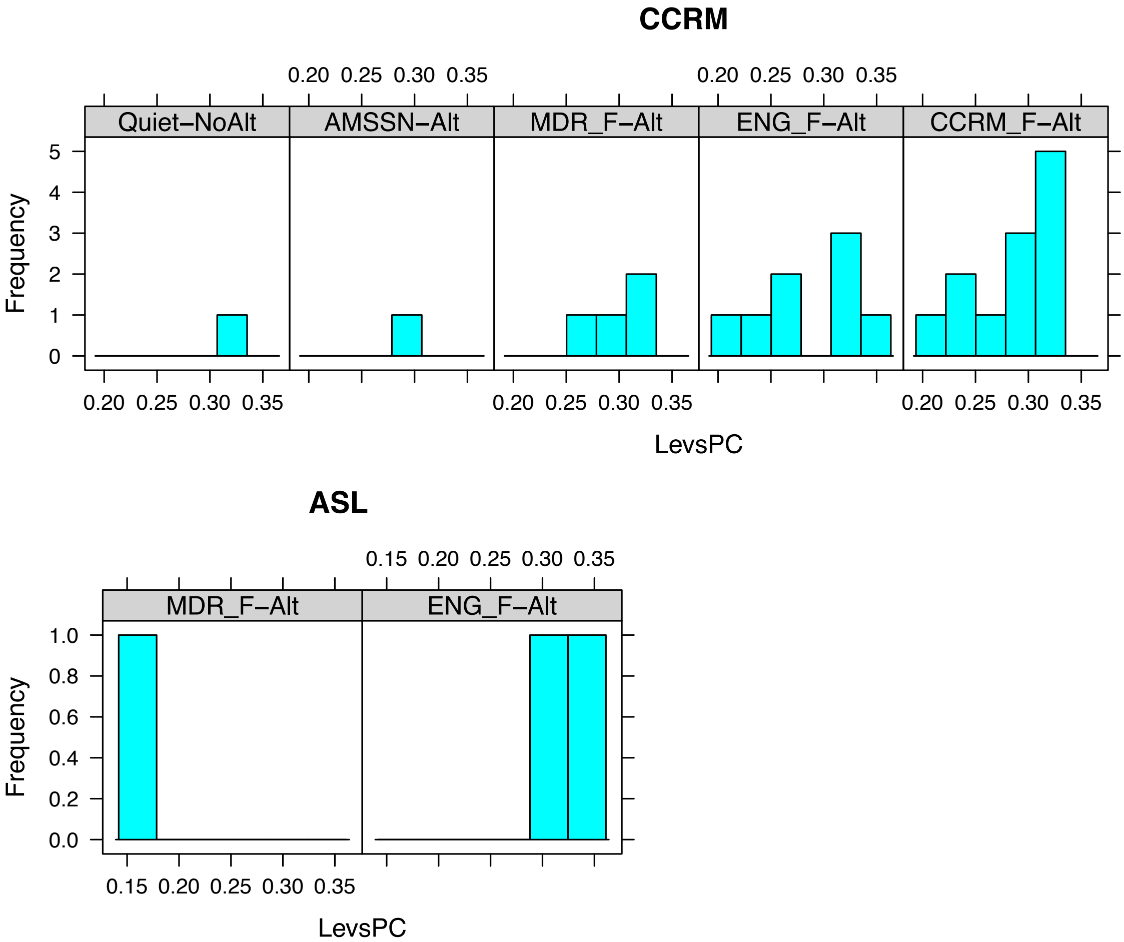

Figure 3.6: ST raw data: Frequency of potential outliers with LevsPC $\leq$ 35%. LevsPC denotes the proportion of correct keywords within the final test trials.

SRdT, either because the procedure ended after reaching the maximum number of trials before a minimum number of test reversals was obtained (x1 CCRM, x2 ASL), or due to aberrant adaptive tracks (x5 CCRM). Since all these cases belonged to more challenging test conditions with speech distractors, it is very likely that the children's true score is at or beyond the upper DC limit ~~(i.e., at ceiling)~~. Thus, to account for that, rather than removing these observations, which will consequently reduce the statistical power and may not represent the true performance in the group, they were set to a DC of 1, which is above the task's upper DC limit of 0.97.

SRdTs by age

Since the present study sample comprised ~~of~~ young children ~~from~~ different age ~~groups~~ from circa 7 to 13 years, developmental age effect was expected, whereby performance was expected to improve with ~~an~~ increasing age ~~due to different maturity effects~~. This is illustrated by the scatterplots and linear regression lines plotted in Figure 3.7 A-B split by groups for the listeners SRdTs obtained across the different test conditions and speech material (ASL / CCRM) as a function of age. Note that smaller SRdT indicate better performance. Age effect was tested against the TD group alone because this group is more ~~hetero~~geneous and thus expected do display smaller variability than the APD group. Nonetheless, despite the larger spread in the APD group, the group showed similar trend in performance, albeit shifted towards higher SRdTs (i.e., poorer performance). The TD regression lines were determined based on a model comparison and outlier~~s~~ trimming procedure ~~described in Section 3.2.4~~ to improve model prediction. ~~Regular~~ regression lines were found to be the most suitable in describing the relationship between the TD children performance and age in all test conditions but the MDR_F condition for the ASL material, where a segmented line was found to give the best fit. MDR_F segmented line indicated that DC improved with age by circa 0.1 per year until reaching a plateau at the age of 9.5 years.

Looking at Figure 3.7 A-B, it is noticeable that children in both groups showed a larger decrement in performance when presented with speech distractors. The regression lines indicate~~s~~ that the improvement in performance by age was more prominent for speech distractors, with relatively steeper slopes (at least twice as steep) than for the non-speech distractor (AMSSN) or for conditions without a distractor. Furthermore, as expected, CCRM sentences were more intelligible, with performance shifted towards lower DC ~~range~~ relative to performance for the ASL speech material. The lower DC meant that the children were able to understand 50% of the sentences with larger portions of the speech information missing.

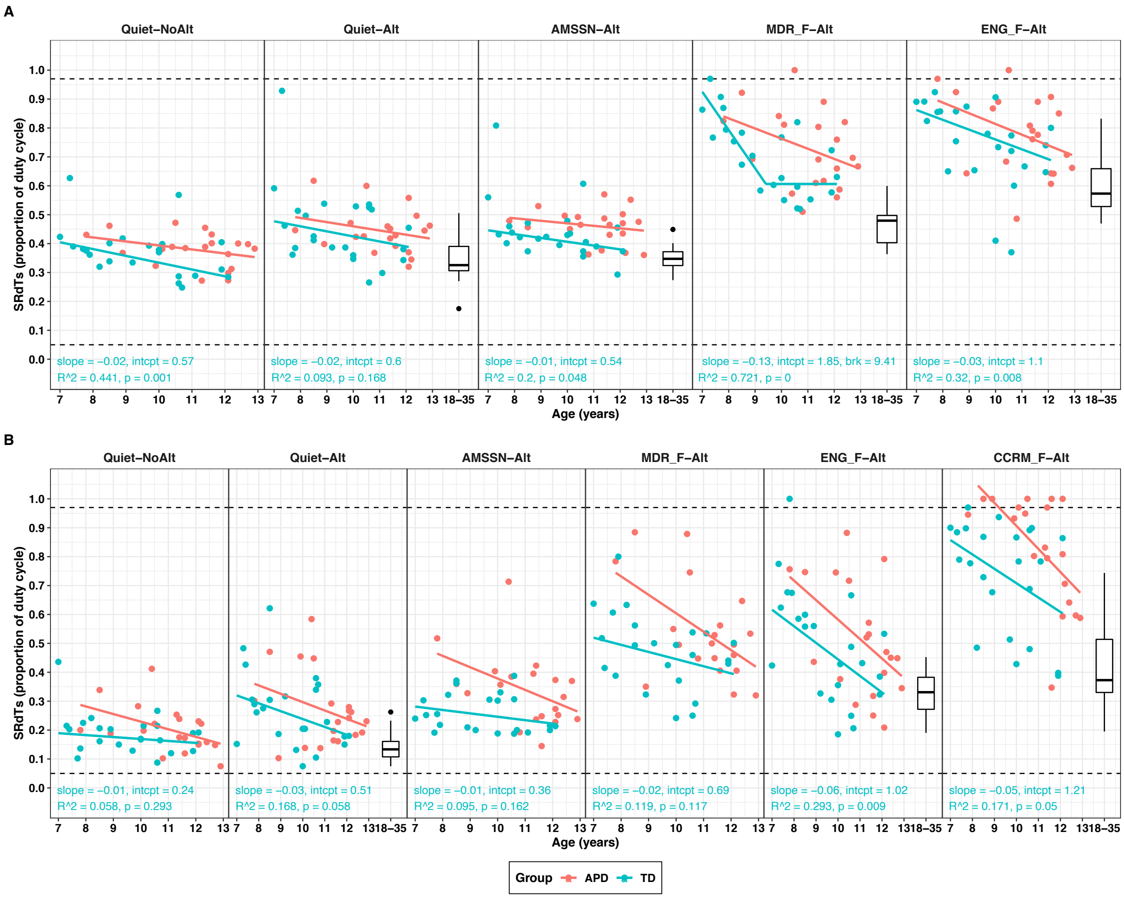

Figure 3.7: ST: Scatterplot and linear regression lines for the listeners SRdTs measured with the ASL (A) and CCRM speech material (B) as a function of age. Corresponding regression coefficients and statistics ~~is~~ provided for TD group only. Red indicates data from the APD group and cyan indicates data from the TD control group. Data for normal hearing adults taken from Chapter 2 is shown in the boxplots as a reference.

A closer look at the ~~linear~~ lines shows several interesting trends. The non-speech AMSSN distractor ~~showed to have~~ little-to-no effect on performance, at least in the TD group, where performance was fairly similar to performance in the Quiet conditions. Introducing alternations (as in Quiet-Alt vs. Quiet-NoAlt), seems to hinder intelligibility in both groups, however the effect is relatively small and may not be significant due to the large spread in the APD group. Furthermore, when comparing the regression lines, there appears to be a relatively larger separation between the groups for ~~data~~ measured with the CCRM material, especially for AMSSN, but also for the speech distractors. However, it is possible that the APD regression lines do not reflect the true population due to the large spread in performance and the small sample size and thus any interpretation should be taken with caution. Another interesting observation is that the children showed little-to-no *masking-release* for speech spoken in an unfamiliar language (MDR_F) when compared with a distractor spoken in English (ENG_F). This is in agreement with findings in the adults' study in Chapter 2. Lastly, it is apparent from the figure that performance for CCRM_F distractor was near-to-ceiling for some children, mostly among the APD group.

Belongs to discussion? I think that is probably right.

An exploratory comparison between the children's data measured in the present study with data measured across young NH adults collected in Chapter 2 further highlight the strong developmental trend, with SRdTs still not entirely "adult-like" even at the age of 13 years, especially for speech distractors (see boxplots in Figure 3.7 A-B). The children in both groups seems to be markedly susceptible to competing CCRM sentences and for familiar- or unfamiliar-speech presented with ASL sentences, with performance at the age of 12 years still largely differing from those obtained by the adults. On the other hand, by the age of 12 years, the TD children reached near to "adult-like" performance when CCRM target sentences were presented with ENG_F speech distractor or when ASL sentences were presented with AMSSN distractor.

Next, age effect was tested using LMEM model, with Condition (Quiet-NoAlt, Quiet-Alt, AMSSN, MDR_F, & ENG_F), Material (ASL / CCRM), Age, and Group (TD children with/without an APD sibling, 1/0) as fixed factors, SRdT as dependent variable and random intercepts for subjects (reference levels: Condition = Quiet-NoAlt; Material = ASL, Group = 0). Note that data for CCRM_F was excluded from the model since it was only measured for the CCRM material[2]. The final LMEM model that gave the best fit and main effects are given in Table 3.10.

[2] A separate model for the CCRM data with CCRM_F-Alt condition showed similar results, with a strong significant Condition x Age interaction ($p < 0.001$) and no main effect of Group ($p > 0.05$).

Table 3.10: ST: Age effect analysis using LMEM for SRdTs measured across condition, speech material, age and group (TD children with/without an APD sibling, 1/0) as fixed factors and random intercepts for subjects. Reference levels: Condition = Quiet-NoAlt, Material = ASL, Group = none sibling (0). Note: only data for the control group (TD) following outliers trimming was included.

SRdT ~ Condition + Material + Age + Group + Condition:Material + Condition:Age + Material:Age + Condition: Group + (1 | Subjects)

Main effects	Df	χ^2	p
Condition:Material	4	10.073	***0.039***
Condition:Age	4	15.948	***0.003***
Material:Age	1	2.073	*0.150*
Condition:Group	4	2.724	*0.605*
Material:Group	1	4.927	***0.026***

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.11: ST: Age-effect - post-hoc paired comparison t-test for Condition x Material interaction. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercepts (subjects) using lsmeans package (emmeans package; Lenth, 2020).

ASL - CCRM	Estimate	SE	Df	t-value	p-value	95%-CI
Quiet-NoAlt	0.19	0.03	222.42	7.26	$<$ ***0.001***	0.14 - 0.24
Quiet-Alt	0.19	0.03	222.16	7.56	$<$ ***0.001***	0.14 - 0.24
AMSSN-Alt	0.18	0.03	222.24	6.97	$<$ ***0.001***	0.13 - 0.23
MDR_F-Alt	0.23	0.03	222.16	9.26	$<$ ***0.001***	0.18 - 0.28
ENG_F-Alt	0.27	0.03	222.12	10.57	$<$ ***0.001***	0.22 - 0.32

* significant p-values ($p < 0.05$) are shown in bold.

Inspection of parametric assumptions based on the model's residuals confirmed that both the assumption of normal distribution and homogeneity of variance were met. Model comparison revealed a significant two-way interaction between Condition x Age, between Condition x Material, and between Material x Group (all p's < 0.05).

The significant Material x Age interaction indicates that the developmental trend is different between the two speech materials, with a larger age effect (i.e., steeper slopes) for the ASL sentences, with an average improvement of 0.01 DC per 1 year, which is approximately 6% higher than for the CCRM sentences across the

Table 3.12: ST: Age-effect - post-hoc paired comparison t-test for Material (ASL/CCRM) x Group (0/1) interaction. The test was performed on the fitted LMEM model and included adjusted least-squared-mean for the random intercepts (subjects) using lsmeans package (emmeans package; Lenth, 2020).

contrast	Estimate	SE	Df	t-value	p-value	95%-CI
ASL 0 - CCRM 0	0.24	0.01	222.02	17.06	$<$ ***0.001***	0.21 - 0.26
ASL 0 - ASL 1	0.01	0.03	35.16	0.21	*0.84*	-0.06 - 0.07
ASLN 0 - CCRM 1	0.19	0.03	36.33	5.84	$<$ ***0.001***	0.13 - 0.26
CCRM 0 - ASL 1	-0.23	0.03	34.87	-7.08	$<$ ***0.001***	-0.3 - -0.16
CCRM 0 - CCRM 1	-0.04	0.03	36.93	-1.36	*0.18*	-0.11 - 0.02
ASL 1 - CCRM 1	0.19	0.02	222.47	9.44	$<$ ***0.001***	0.15 - 0.22

* significant p-values ($p < 0.05$) are shown in bold.

age span. Furthermore, the significant Condition x Material interaction implies that performance in the different test conditions differed between the two speech materials. A post-hoc t-test comparison based on the fitted model given in Table 3.11, revealed a highly significant difference in performance between the speech materials across all five test conditions (all p's < 0.001). The estimated mean difference between the contrast pairs ranged between +0.18 to +0.27; hence, the CCRM speech material was significantly more intelligible than the ASL material, across all test conditions.

The significant Condition x Age interaction supports the observation in Figure 3.7 A-B, that the effect of age was different across the test conditions. These findings raises the following questions – do all the conditions show a significant age effect? Moreover, since the effect of age is not the same across the test conditions, which conditions showed the largest age effect? One possible way to tackle these questions is to compare the separate regression models using F-statistics. Nonetheless, due to the small sample-size and the large number of paired comparisons, such test lacks a statistical power and the results may not reflect the true effect in a larger sample. The TD group regression model's R^2 and p-values are given at the bottom part of Figure A and B. The ASL models p-values indicated a highly significant age effect for ENG_F, MDR_F and Quiet-NoAlt condition as well as a marginal effect for AMSSN ($p = 0.048$), whereas no significant age effect was found for Quiet-Alt (p = 0.168). As for the CCRM material, there was a highly significant age effect for ENG_F and a marginal effect for the Quiet-Alt condition (p = 0.058) and for CCRM_F condition (p = 0.05) which was not included in the LMEM model, whilst there was no significant age effect found for Quiet-NoAlt, AMSSN and MDR_F conditions. Furthermore, age was found to be a better predictor (i.e., accounting for larger variance in SRdT) for conditions with speech distractors, with R^2 ranging between 32% to 72% for the ASL material and about 12% to 29% for the CCRM. A comparison between the test conditions regression line slopes split by test material is depicted in Figure 3.8. A possible pattern emerges from the figure, where slopes for the quiet and non-speech conditions are fairly similar across the two speech material (indicated by their proximity to the diagonal line), while, differences between the slopes are relatively larger for speech distractors, in particularly for MDR_F where the slope for the ASL material (-0.13) is about six times steeper than the slope for the CCRM material (-0.02).

Lastly, the significant interaction between test Material and Group was explored using two separate LMEM models for ASL and CCRM data. Both models found no significant effect of Group. Comparison of the average performance by material between the two control groups revealed an opposite direction of performance. While the performance of the TD children with an APD sibling was on average 0.04 better than their TD peers for the ASL material, their performance was 0.02 poorer for the CCRM material. These results are in contradiction to our expectation ~~in case~~ differences ~~between the two TD groups were to occure~~ In such case, we predicted a larger ~~in~~crement in performance (i.e., poorer score) for the ASL sentences which are more linguistically challenging than for the CCRM sentences. This dis~~s~~agreement and the very small estimated mean difference for the interaction between Material and Group in the full model (0.052) suggests that the differences picked up by the model are due to sampling error.

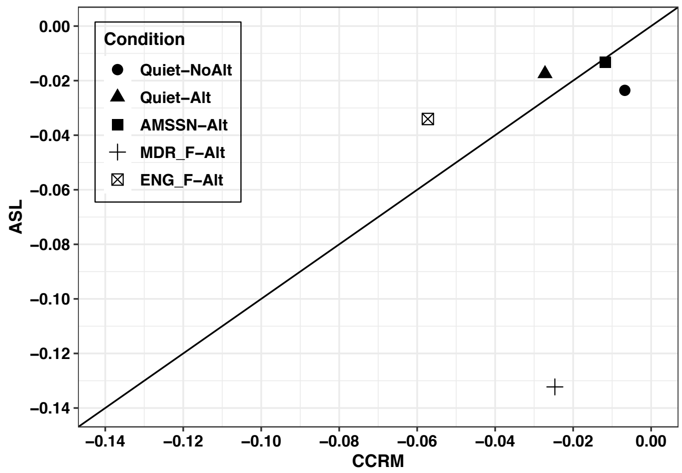

Figure 3.8: ST: Age effect - a comparison beteween the regression line~~s~~ slopes fitted for the CCRM (x-axis) and ASL speech material (y-axis). Test conditions are represented by the different symbols. The diagonal line represents an optimal agreement between the speech materials. Observations falling below the line indicate a steeper slope for the ASL material than for the CCRM material.

Where should it best go?

Simple correlation for the listeners SRdTs between conditions is given in appendices, separately for the ASL (Figure C.1) and CCRM material (Figure C.2).

Age-independent z-scores

Age-independent standardised residuals (z-scores) were calculated based on a model prediction for the TD group data using a multiple-case study approach [Ramus et al. (2003); or see section 3.2.4 for more details]. Descriptive statistics for the listeners z-scores ~~is~~ given in Table 3.13. Additional boxplots are shown in Figure 3.9 A-B, for the ASL and CCRM speech material respectively. Scores were calculated separately for each test condition, with better performance indicated by lower z-score. The grey area marks the two-tailed 1.96 deviance cut-off for abnormal score from the theoretical control group mean (z = 0), where only about 5% of the normal population is expected to score below and above it. Overall, APD children performance ~~was~~ noticeably poorer ~~in~~ both test material, with higher median z-scores ~~than~~ compared with the TD children. The next paragraphs will cover inspection of the data and statistical analysis of group differences separately

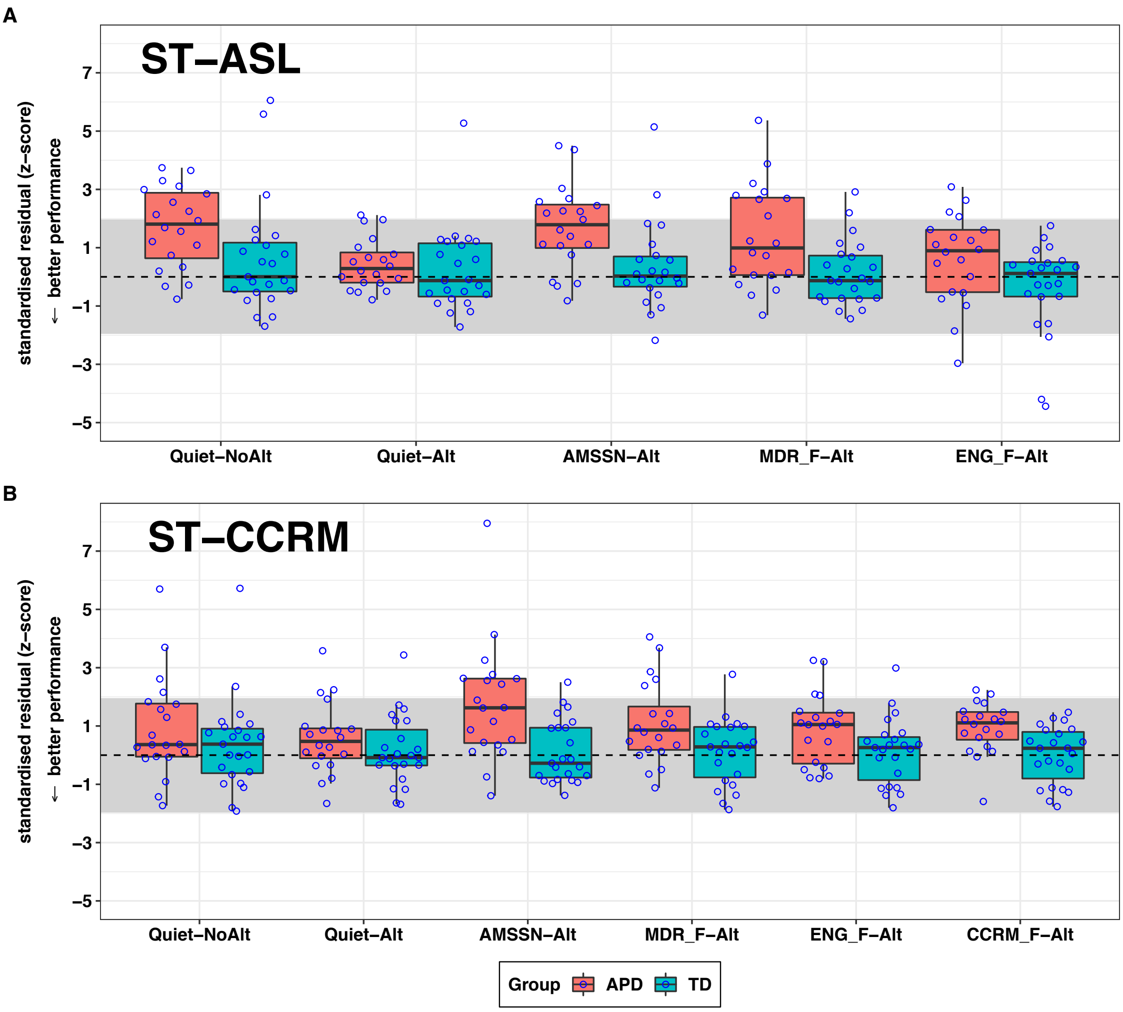

Figure 3.9: ST: Boxplots of the listeners age-independent standardised residuals for data measured with the ASL (A) and the CCRM speech material (B). Residuals were calculated separately for each condition and are based on a model prediction for TD group only. The grey area represents the deviance cut-off for abnormal score (SD ± 1.96 below and above the TD mean), where about 95% of the normal population is expected to lay within. The dashed line represents the theoretical TD group mean (z = 0). Individual scores are indicated by circles.

for each speech material.

Where should it best go?

Again, simple correlation for the listeners z-scores between conditions is given in appendices, separately for the ASL (Figure C.3) and CCRM material (Figure C.4).

Table 3.13: ST: Descriptives for standardised residuals (z-scores) calculated for data measured with the ASL and CCRM speech material.

	APD						TD					
	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal
ASL												
Quiet-NoAlt	20	1.81	1.39	-0.76	3.74	45.00%	23	0.00	1.96	-1.69	6.05	13.04%
Quiet-Alt	20	0.29	0.87	-0.79	2.12	10.00%	23	-0.13	1.46	-1.72	5.27	4.35%
AMSSN-Alt	20	1.79	1.45	-0.82	4.50	50.00%	23	0.10	2.35	-2.18	9.04	13.04%
MDR_F-Alt	20	0.99	1.75	-1.31	5.37	40.00%	23	-0.13	1.11	-1.44	2.91	8.70%
ENG_F-Alt	20	0.90	1.53	-2.96	3.09	20.00%	23	0.12	1.55	-4.44	1.75	0.00%
CCRM												
Quiet-NoAlt	20	0.36	1.75	-1.73	5.70	20.00%	23	0.38	1.57	-1.92	5.72	8.70%
Quiet-Alt	20	0.47	1.23	-1.66	3.58	15.00%	23	-0.08	1.19	-1.68	3.44	4.35%
AMSSN-Alt	20	1.62	2.03	-1.39	7.95	40.00%	23	-0.28	1.09	-1.38	2.50	4.35%
MDR_F-Alt	20	0.86	1.40	-1.12	4.06	25.00%	23	0.28	1.11	-1.87	2.77	4.35%
ENG_F-Alt	20	1.05	1.22	-0.80	3.25	20.00%	23	0.26	1.14	-1.80	2.99	4.35%
CCRM_F-Alt	20	1.11	0.89	-1.59	2.24	10.00%	23	0.24	0.98	-1.76	1.47	0.00%

abnormal: defined as the percentage of abnormal z-score > 1.96.

ASL speech material

Surprisingly, a comparison of the groups averaged z-score reveals that the non-switched quiet condition (Quiet-NoAlt) and the switched condition with the nonspeech distractor (AMSSN) yielded the largest separation between the groups, with APD median z-score of 1.81 and 1.79, respectively, laying just within the norms upper limit. Performance of the APD children was also noticeably poorer for conditions with speech distractors (MDR_F and ENG_F), each with a median z-score of circa 1, whereas performance for Quiet-Alt condition was fairly similar between the groups.

Within the APD group AMSSN, Quiet-NoAlt and MDR_F resulted in the highest proportion of abnormal scores[3]. Surprisingly, AMSSN distractor yielded the highest proportion of abnormal scores, where half of the APD children fell outside the norm (20/10, 50%). Followed by the non-switched condition Quiet-NoAlt, where paradoxically and against our expectation 45% of the APD group (9/20) had abnormally poor score, whereas only 10% (2/20) had abnormal score in

[3] with the aim to develop a clinically applicable test that exhibits good sensitivity and specificity ~~rates,~~ we were only interested in identifying children with clinically poor performance. Thus, abnormal score was defined as a one-tailed deviance cut-off of z-score > 1.96, within which circa 97.5% of the normal population is expected to lay.

the switched condition Quiet-Alt. Another interesting finding was that the APD children did not benefit from a release from masking for a speech distractor spoken in an unfamiliar language (MDR_F) as opposed to a familiar speech spoken in English (ENG_F), with median scores very similar in both conditions. This sits well with our previous findings with adults where adults showed no benefit for MDR_F speech masker (see Chapter 1). Moreover, while the overall performance was similar in the two conditions, the percentage of abnormal score was twice as large for MDR_F condition (8/20, 40%) than for ENG_F condition (4/20, 20%). The proportion of abnormal scores amongst the TD group ranged between 0% to 13% (M = 7.8%), which is relatively higher than expected in the normal population.

CCRM speech material

Figure 3.9 B reveals a similar trend for the CCRM sentences, nonetheless with more modest differences between the two groups. Again, AMSSN yielded the largest separation between the groups, where 40% (8/20) of the APD children obtained an abnormal score and with a median score of 1.62, which is relatively close to the +1.96 upper deviance cut-off. In comparison, only 4.3% of the TD children (1/23) had abnormal performance for AMSSN condition. The APD group median score for the speech distractors was approximately 1 (range: 0.86-1.11), however the proportion of abnormal APD children was noticeably smaller than seen for AMSSN, with 25% (5/20) for MDR_F, 20% (4/20) for ENG_F, and only 10% (2/20) for CCRM_F distractor. Lastly, in contrast to the ASL material, performance for the CCRM sentences presented in quiet were relatively better without switching (NoAlt) than with switching (Alt). Nonetheless, the spread in performance for the non-switched condition was larger. The percentage of abnormal scores in the TD group were relatively low, ranging between 0 to 8.7% (M = 4.3%).

A three-way 3 x 2 x 5 factorial design model with repeated measures was used to test the main effects of Group (APD, and TD with/without APD sibling), Material

(ASL / CCRM) and Condition as well as their interaction on performance in the task with z-scores as a dependent variable. Note that the model did not include the CCRM test condition with CCRM-type sentences as distractor (CCRM_F) since there was no comparable condition in the ASL speech material. Inspection of parametric methods assumptions for the residuals of a linear model revealed that the assumption of normal distribution was rejected, whereas the assumption of homogeneity of the variance was met. Since there are several obvious outliers in the data and due to the incomplete fulfilment of parametric assumptions a non-parametric approach was adopted. This was tested with a rank-based ANOVA-type statistic test (ATS) using the *nparLD()* function (nparLD package; Noguchi et al., 2012). The analysis was based on a f2-ld-f1 design ATS test, whereby f2 refers to an experimental design with two between-subjects factors (Group & Material) and f1 refers to a single within-subjects factor (Condition). The test results are given in Table 3.14. No significant three-way or two-way interaction were found, except for a Group x Material interaction ($p < 0.05$). In addition, there was no significant main effect of Condition.

To further examine the significant Group x Material interaction, we tested two separate f1.ld.f1 models per test material with Group and Condition as predictors. Both models found a significant effect of Group [ASL: Statistic = 4.099, df=1.720, $p < 0.05$; CCRM: Statistic = 3.699, df = 1.605, $p < 0.05$], while there was no significant interaction or main effect of condition (all p's > 0.05). A post-hoc pairwise comparison using Wilcoxon rank-sum test was performed to examine the main effect of Group (see Table 3.15. There was a highly significant difference in performance between the APD and the TD children without an APD sibling in both test materials. Furthermore, the performance of APD children for the ASL material was (highly) significantly poorer than the performance of the TD children with an APD sibling. However, there was no significant difference in performance between the two groups when measured with the CCRM material. Furthermore, while there was no significant difference between the TD groups when measured

Table 3.14: ST: Statistical analysis for the effects of Group, Material, and Condition as well as their interaction (3x2x5 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f2-ld-f1 design ANOVA-type statistic (ATS) test, whereby f2 refers to an experimental design with two between-subjects factors (Group and Material) and f1 refers to a single within-subjects factor (Condition).

	Statistic	df	p-value
Group	4.009	1.531	***0.028***
Material	0.047	1.000	*0.828*
Condition	1.394	3.251	*0.24*
Group:Material	3.767	1.952	***0.024***
Condition:Material	0.669	2.682	*0.554*
Group:Condition	1.594	5.261	*0.154*
Group:Material:Condition	0.660	4.294	*0.631*

* significant p-values ($p < 0.05$) are shown in bold.

Table 3.15: ST: Post-hoc paired comparison (Wilcoxon rank-sum test) for Group differences in z-score split by material.

	ASL model							CCRM model						
contrast	estimate	n1	n2	95%-CI	p	r	magnitude	estimate	n1	n2	95%-CI	p	r	magnitude
APD - TD_{none}	0.89	100	75	0.46 - 1.34	< **0.001**	0.30	moderate	0.96	100	75	0.58 - 1.33	< **0.001**	0.37	*moderate*
APD - $TD_{sibling}$	1.15	100	40	0.6 - 1.68	< **0.001**	0.33	moderate	0.40	100	40	-0.1 - 0.94	0.12	0.13	*small*
TD_{none} - $TD_{sibling}$	0.24	75	40	-0.23 - 0.7	0.33	0.09	small	-0.57	75	40	-1.03 - -0.1	**0.02**	0.22	*small*

* significant p-values ($p < 0.05$) are shown in bold.

with the ASL material, TD children with an APD sibling performed significantly poorer than their TD peers when presented with the CCRM material.

An additional 3 x 6 model was computed for the full CCRM data, including the test condition with the CCRM-type distractor (CCRM_F-Alt). The model included Group and Condition as between- and within-subjects predictors, respectively, with z-scores as the dependent variable using nparLD ATS test (f1.ld.f1 design). The ATS test results were similar to those of the full model, with a significant main effect of group (Statistic = 4.922, df = 1.597, $p < 0.012$). However, there was no significant main effect for Condition nor a significant Group x Condition interaction (both p's > 0.05). A post-hoc pairwise comparison for groups (Wilcox rank-sum tests) found a significant difference between all three groups.

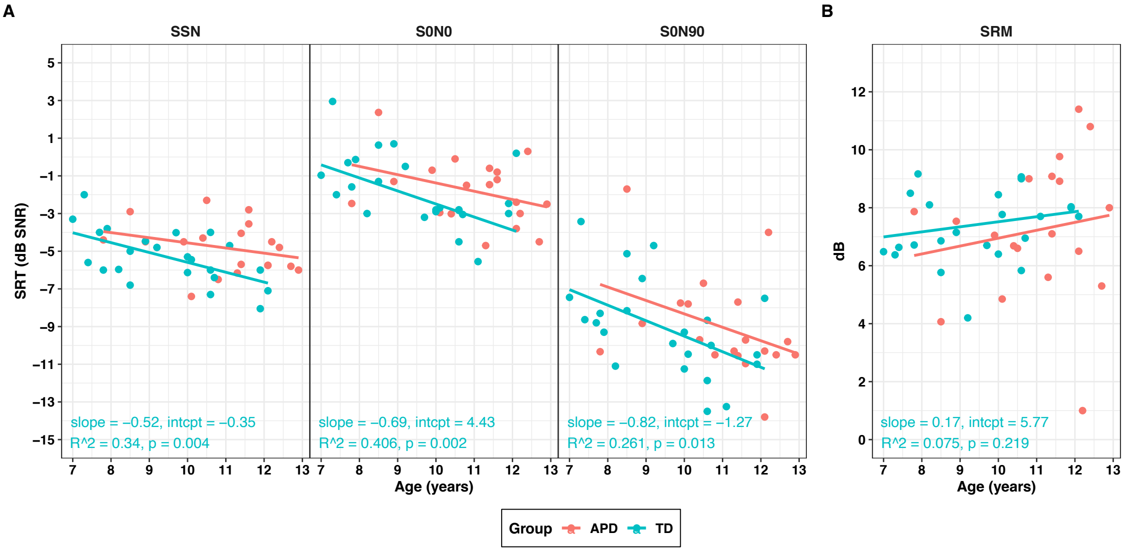

Figure 3.10: LiSNS-UK: Age-effect - scatterplot and linear regression lines for SRTs obtained for SSN and the spatialised conditions S0N0 (collocated) and S0N90 (separated) (A) and the derived measure SRM (B) as a function of the listeners age. Corresponding regression coefficients and statistics is provided for TD group only. Red indicates data from the APD group and cyan indicates data from the TD control group.

3.3.4 LiSNS-UK

SRTs by age

~~The distribution of the~~ listeners SRTs and their corresponding regression lines split by group ~~is~~ shown in Figure 3.10 A for the spatially- collocated (S0N0) and separated condition (S0N90), as well as for the non-spatialised condition where the ASL sentences were presented with a speech-shaped-noise (SSN). The listeners binaural advantage, calculated as the difference between the collocated and separated spatial conditions (SRM = S0N0 - S0N90) is shown in Figure 3.10 B. As in the switching task, age effect was tested ~~against~~ the TD group only, where the regression lines for the TD group were estimated based on a model comparison and outliers-trimming procedure to improve the model's fit (model coefficients and statistic are given at the bottom of the figures).

As previously reported by other researchers that used similar test paradigm in children from a similar age group (e.g., Cameron & Dillon, 2007; Murphy et al., 2019), the scatterplots shows a clear developmental trend, with an overall improvement in performance with an increase in age. The test conditions S0N90 and S0N0 showed the largest age effect, with near to 1 dB improvement in performance per 1 year increase (TD slope: -0.82 & -0.69, respectively). The ~~regression lines~~ slope for SSN ~~conditions~~ was shallower, with roughly half a dB improvement in performance per 1 year increase, with a TD slope of -0.52. Difference in performance with age for the SRM ~~was~~ negligible, with a predicted improvement of circa 1 dB between the age of 7 to 13 years. There was a significant effect of age in all three test conditions ('moderate' effect size), with the largest effect for S0N0, accounting for circa 40% of variability in performance, followed by SSN with 34% and about 26% for S0N90. The linear regression ~~fit~~ for SRM showed no significant age effect ~~for SRM~~ ($R^2 = 0.075$, $p = 0.219$).

A factorial design model with repeated measures was used to test the main effects for Condition (SSN, SON0, S0N90, & SRM), Age, and Group (TD children with/without an APD sibling, 1/0) with TD group SRTs as a dependent variable. Interaction terms were included as well as random intercepts for subjects. Note that also here the model included only data for the control group. Assumptions of normal distribution and homogeneity were met, and thus a parametric approach was applied using LMEM (reference levels: Condition = SSN; Group = TD without an APD sibling, 0). The model with the best fit and main effects are given in Table 3.16. There was a significant interaction between Condition x Age ($p < 0.05$), thus indicating that age affected performance differently across the different test conditions. Moreover, there was no significant main effect of Group ($p > 0.05$), thus suggesting that performance of TD children with and without an APD sibling was fairly similar. Maybe worth saying you're not going to think about that any more in this section?

Age-independent z-scores

Boxplots of the listeners age-independent standardised residuals z-scores (blue circles) collapsed across the different test conditions are shown in Figure 3.11, separately for the APD group (red) and TD group (cyan). The z-scores were

Table 3.16: LiSNS-UK: Age effect - LMEM model for SRT with condition, age and group (TD children with/without an APD sibling, 1/0) as fixed factors and random intercepts for subjects (reference level: SSN, Group = TD without APD sibling, 0). Note: only data measured with the control group (TD) following outliers trimming was included.

SRT ~ Condition + Age + Group + Condition:Age + (1 | Subjects)			
Main effects	Df	χ^2	p
Condition:Age	3	14.292	***0.003***
Group	1	0.051	*0.822*

* significant p-values ($p < 0.05$) are shown in bold.

calculated in the ~~exact~~ same way as for ST. Again, the dashed line indicates the theoretical TD group mean of zero, and the grey area indicates the lower and upper limit of the normal population (TD mean ± 1.96). Descriptive statistics collapsed by group and test conditions are given in Table 3.17. Overall, when compared with the control group, the APD children exhibited poorer performance across all three test conditions (i.e., higher z-score) as well as for the derived SRM measure (i.e., lower z-score).

S0N0 and SRM yielded the largest separation between the groups~~, h~~owever the spread in scores was relatively large and the percentage of abnormal performance in the APD group was rather small, with only circa 26% (5/19) in each condition. ~~Whereas~~ only about 16% (3/19) and 10% (2/19) of the APD children had abnormal score for SSN and S0N90, respectively. No abnormal performance was obtained in the TD group for SSN and S0N90, while two TD children (~9%) had abnormal score for S0N0 ~~condition~~ and one child for SRM. Nonetheless, when excluding the TD outliers that were trimmed during the z-score calculation procedure, all the TD observations were within the norms.

Group differences between APD and TD children for the test conditions SSN, the spatialised conditions S0N0 and S0N90, and SRM were tested with a rank-based ANOVA-type ATS test using nparLD package (f1.ld.f1 design) since the assumption of homogeneity of variance was not met. The test results are given in Table 3.18.

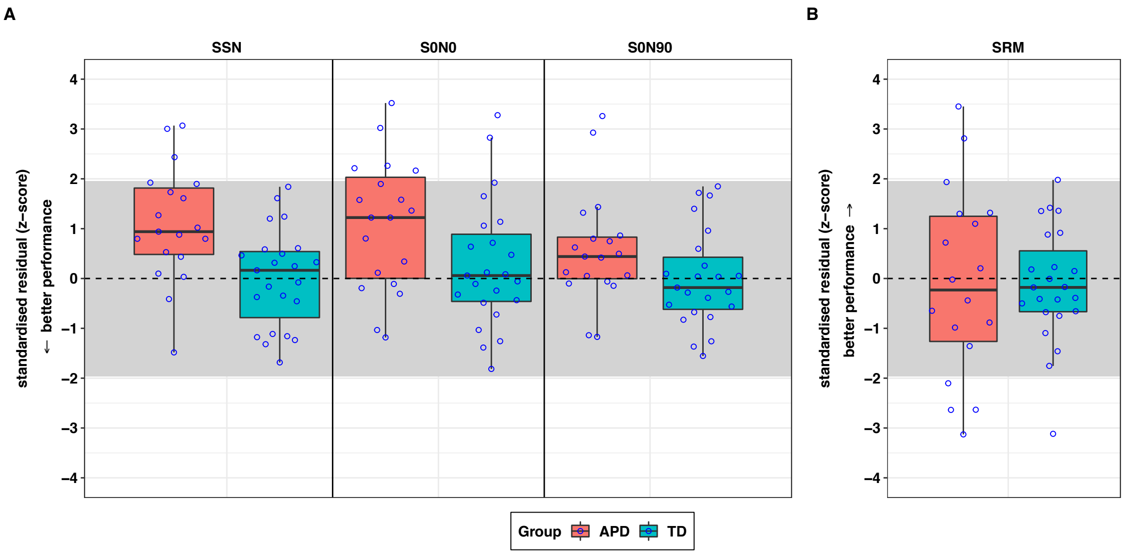

Figure 3.11: LiSNS-UK: Boxplots of the listeners age-independent standardised residuals (open circles) for data measured with LiSNS-UK task (A) and the derived measure SRM (B). Residuals were calculated separately for each condition and are based on a model prediction for TD group only. The grey area represents the deviance cut-off for abnormal score (SD ± 1.96 below and above the TD mean), where about 95% of the normal population is expected to lay within. The dashed line represents the theoretical TD group mean (z = 0).

Table 3.17: LiSNS-UK standard residuals (z-scores) descriptives by group. abnormal: defined as the percentage of ~~abnormal~~ z-score > 1.96 (SSN, S0N0, & S0N90) and z-score < 1.96 (SRM).

	APD						TD					
	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal
SSN	19	0.94	1.14	-1.48	3.07	15.79%	23	0.16	0.98	-1.68	1.84	0.00%
S0N0	19	1.22	1.31	-1.18	3.52	26.32%	23	0.06	1.28	-1.81	3.28	8.70%
S0N90	19	0.44	1.11	-1.17	3.26	10.53%	23	-0.18	0.98	-1.55	1.85	0.00%
SRM	19	-0.44	2.39	-6.77	3.45	26.32%	23	-0.18	1.15	-3.12	1.98	4.35%

There was no significant interaction between Group and Condition ($p > 0.05$), while there was a significant main effect of both Group and Condition ($p< 0.05$).

Does it make sense to run a post hoc test? z-score goes to the opposite direction (larger -> better). Or better, perhaps I should not include SRM in the full model?

Table 3.18: LiSNS: Statistical analysis for the effects of Group and Condition as well as their interaction (3x4 factorial design with repeated measures) tested with a robust rank-based method for analysis of nonparametric data using nparLD package (Noguchi et al., 2012). Analysis was based on a f1-ld-f1 design ANOVA-type statistic (ATS) test, whereby f1 refers to an experimental design with a single between-subjects factor (Group) and a within-subjects factor (Condition).

	Statistic	df	p-value
Group	8.857	1.000	***0.003***
Condition	3.405	1.897	***0.036***
Group:Condition	1.679	1.897	*0.188*

* significant p-values ($p < 0.05$) are shown in bold.

3.3.5 ENVASA

Due to technical problems, observations for six listeners are missing (x2 TD; x4 APD), resulting in a total sample-size of 21 and 17 for the TD and the APD group, respectively. Initial inspection ~~of the individuals performance~~ was performed to ensure that the task instructions were followed and well understood. Performance for the reference condition (single incongruent background at a high SNR), which is expected to least impact performance, was compared with a cut-off criterion of 56%, calculated as 2 SD from the TD group mean (84% ± 14%). Individuals with performance below the cut-off criterion were excluded from the analysis. One TD listener aged 7 years old scored 45 % and was thus excluded, resulting in a total of 20 listeners in the TD group.

%-correct by age

The ENVASA measurements followed the same factorial design as used by Leech et al. (2009), with 2 background types (single/dual) x 4 SNRs (low: -6, -3 dB; high: 0 +3 dB), resulting in a total of 92 responses (%-correct, PC) per listener or between 10 to 11 test items per background-SNR combination. Because of the small number of test items per condition, responses were averaged into three measures: 1. *single background*, 2. *dual backgrounds*, and 3. *combined background* which reflects the overall performance across the two background types.

The relationship between performance and age was inspected in the same way as carried out for the other auditory tasks, with the listeners average ~~response~~ plotted as a function age, with linear regression lines and model coefficients for the trimmed TD group (see Figure 3.12). The regression lines revealed a noticeable developmental trend in all three measures, where performance improved with increasing age. A single linear regression line with a monotonic increase in performance by age was found to best fit performance for a single background, with an increase of circa 3.5% in PC per year. Performance for dual backgrounds and the combined score on the other hand were best described using segmented linear regression models, with an increase of PC by circa 12% per year until the age of 9 years, where PC plateaued thereafter.

The effect of age was statistically tested using an LMEM model with PC as dependent variable, and ~~the three background measures~~, the listeners' age, and group (TD chidren with/without an APD sibling, 1/0) as fixed factors as well as random intercepts for subjects (reference levels: Background = single-background, Group = 0). A model without an interaction term was found to give the best fit (see Table 3.19). Model comparison revealed a highly significant main effect of Age and Background ($p < 0.001$). This is in agreement with ~~the~~ Krishnan et al. (2013) ~~study~~ where they found a strong developmental effect across normal-hearing typically developing children in a similar age range to those measured in the present study. The main effect of Group was marginal ($p = 0.05$) with an estimated mean difference of 5.15 (SE = 2.50, CI = 0.24 - 10.05) and a 'small' effect-size (Cohen's d).

Age-independent z-scores

For further analysis, age was controlled for using the same multiple-case approach method described in Section 3.2.4. Boxplots of the age-independent z-scores for the three ENVASA measures are shown in Figure 3.13, with larger z-score indicating better performance. The grey area indicates the upper and lower cut-off $\pm$ 1.96 for

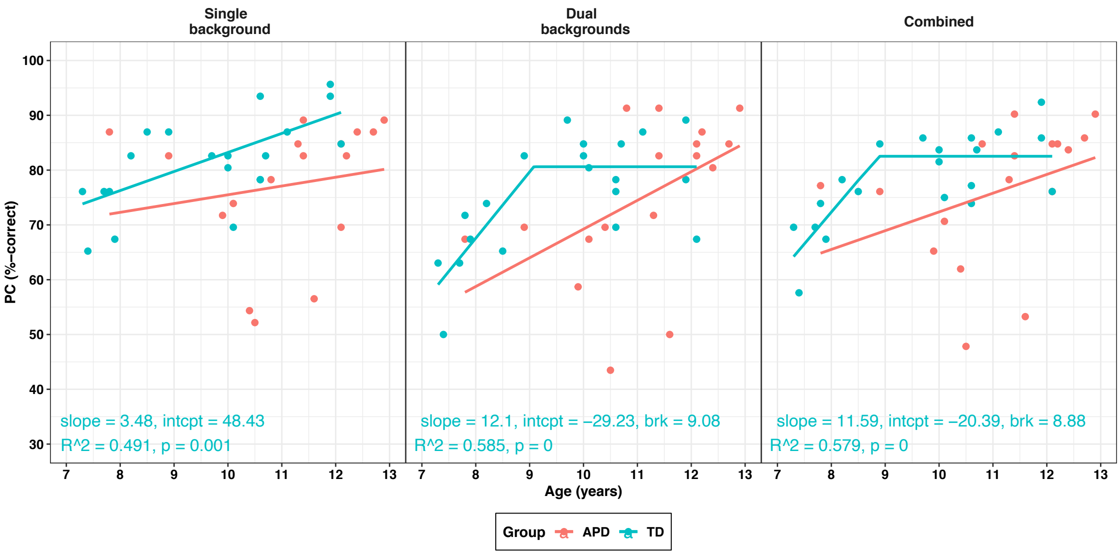

Figure 3.12: ENVASA: Scatterplot and linear regression lines for the listeners' PC (%-correct) as a function of age for single background, dual backgrounds and the combined measure. Red indicates data from the APD group and cyan indicates data from the TD control group.

Table 3.19: ENVASA: Age effect - LMEM model for PC (%-correct) in the three background measures single, dual, & combind background/s, group (TD children with/without an APD sibling, 1/0), and age as fixed factors and random intercepts for subjects (reference levels: Background = single-background, Group = non-APD sibling, 0). Note: only data measured with the control group following outliers trimming was included.

PC ~ Background + Age + Group + (1 | Subjects)			
Main effects	Df	χ^2	p
Background	2	21.173	***<0.001***
Age	1	17.516	***<0.001***
Group	1	3.834	*0.050*

* significant p-values (p < 0.05) are shown in bold.

normal score, where scores of about 95% of the normal population are expected to lay within. Surprisingly, the less demanding condition with single competing background yielded the largest separation between the group with a median z-score of roughly -1, while the median performance for dual backgrounds and the combined score was relatively similar to those in the control group, albeit with larger spread. The percentage of abnormal APD scores was relatively low, with circa 29% (5/17) for the combined score, 24% (4/17) for single background and 18% (3/17) for dual

Table 3.20: ENVASA: Descriptive and statistics of the listeners age-independent standard residuals (z-scores) split by groups and test measures.

	APD						TD						Wilcoxon rank-sum test			
background	N	median	sd	min	max	abnormal	N	median	sd	min	max	abnormal	95%-CI	p	r	magnitude
Single	17	-0.97	2.11	-5.56	1.93	23.53%	20	0.03	1.08	-2.37	1.52	5.00%	-2.27 - -0.34	***0.02***	0.39	moderate
Dual	17	0.29	2.07	-5.34	1.54	17.65%	20	0.22	0.95	-1.90	1.22	0.00%	-1.56 - 0.62	*0.66*	0.08	small
Combined	17	0.02	2.39	-6.42	1.42	29.41%	20	0.22	0.95	-1.59	1.83	0.00%	-1.81 - 0.4	*0.29*	0.18	small

* significant p-values ($p < 0.05$) are shown in bold.

backgrounds condition. There was only one case of abnormal score in the TD group for single background (5%, 1/20) when trimmed TD outliers ~~are~~ included.

A two-way interaction between Group (APD/TD) and Condition (2 x 3 factorial design data with repeated measures) was tested with a non-parametric robust aligned rank test using *npIntFactRep* package (Feys, 2015). Mauchly's test indicated that the assumption of sphericity for the two-way interaction term had been violated ($p < 0.001$); therefore the degrees of freedom was corrected using Greenhouse-Geisser estimate of sphericity ($\varepsilon = 0.55$). The test showed a significant two-way interaction between Group and Condition [$F(1.64,57.57) = 10.82$, $p < 0.001$]. Difference between the groups were examined using unpaired two samples Wilcoxon rank-sum test with permutation (N=999999) which is a t-test equivalent for non-parametric data (*coin::wilcox_test()*; Hothorn et al., 2006). Groups descriptives collapsed by the three test measures as well as p statistics and effect-size r are given in Table 3.20. Performance of the APD children was significantly poorer than of the TD children in the single background condition ($p < 0.05$, 'moderate' effect), whereas there was no significant difference between the groups in the dual backgrounds or the combined background measure (both p's > 0.05).

3.3.6 CELF-RS

The children's raw scores were converted into age-corrected scaled scores using the CELF-5 UK Recalling Sentences subtest standardised norms ($M = 10$, $SD = 3$). Boxplots of the children's scaled scores split by groups are given in Figure 3.14. The grey area indicates the upper and lower limit among the normal population

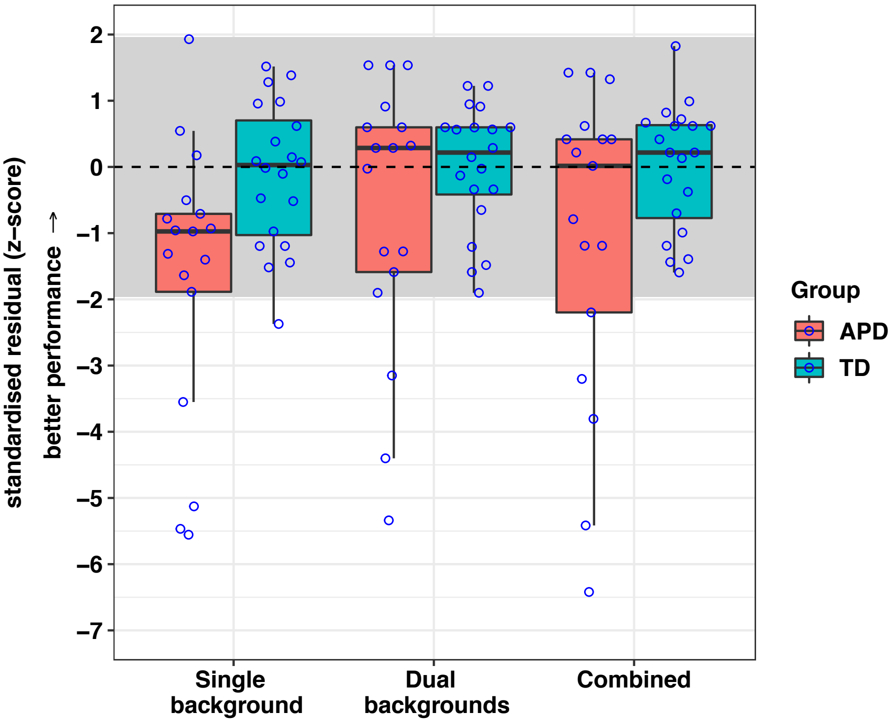

Figure 3.13: ENVASA: Listeners' age-independent standardised residuals for single background, dual backgrounds & the combined measure. Residuals were calculated seperately for each condition and are based on a model predicton for TD group only. The grey area represents the deviance cut-off for abnormal score (SD ± 1.96 below and above the TD mean), where about 95% of the normal population is expected to lay within. The dashed line represents the theorethical TD group mean (z = 0).

I hope your theorising is always ethical!

(± 1 SD). On average, performance was within the norms range in both the APD group (Mdn = 9) and the TD group, albeit laying within the upper limit (Mdn = 13). Thus, although the majority of the APD children expressive language skills were within the norms, the figure shows a clear difference in performance between the group, ~~where~~ the TD children ~~expressive language skills are~~ noticeably better. Almost half of the TD children obtained a scaled score above the average and none exhibited abnormal scores. On the other hand, only three APD children performed above the average and performance of two children was considered abnormal (scaled score < 7). A one-way ANOVA was computed using *anova()* function to compare the listeners scaled scores in the three groups (APD, and TD children with/without an APD sibling). Parametric assumption of homoscedasticity was met while the assumption of normal distribution was marginally significant (Shapiro-Wilk test; p = 0.041). However, since nonparametric methods gave similar results, it was

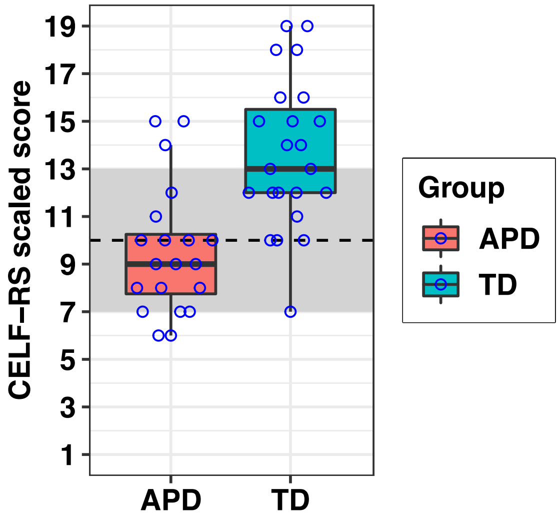

Figure 3.14: CELF-RS: Boxplots for CELF-5 UK Recall Sentences subtest scaled scores by groups. The dashed line represents the norms mean and the grey area indicates the upper and lower limit average performance in the normal population (± 1 SD).

decided to report here only the outcomes of the parametric method. There was a highly significant difference in scaled scores between the groups [$F(2,40) = 14.476$, $p < 0.001$]. A post-hoc pairwise comparison t-tests with Bonferroni correction using *pairwise_t_test()* function (rstatix package; Kassambara, 2021) found a highly significant difference between the APD group (Mdn = 9.0, SD = 2.7) and the TD group without an APD sibling [Mdn = 15, SD = 2.4, $t(31.9) = -5.84$, $p < 0.001$], whereas there was no significant difference between APD and TD children with an APD sibling [Mdn = 11, SD = 3.5, $t(10.6) = -1.50$, $p = 0.486$] or between the two TD groups [$t(10.7) = 2.19$, $p = 0.155$].

3.3.7 Questionnaires

CCC-2

Data for one TD listener was flagged as inconsistent using the test scorer and was thus removed from the analysis. The groups descriptives for the parental reports in the different sub-scales as well as the GCC and SIDC composites are given in Table 3.21. GCC stands for general communication composite, calculated by taking the sum for scaled scores A to H. It is used to clinically identify abnormal communication skills, defined by a GCC < 55 (10^{th} percentile). The SIDC stands for social-interaction deviance composite [sum(E+H+I+J)-sum(A+B+C+D)], where in combination with abnormal GCC score, the SIDC can be used to identify the child's primary difficulty, whereby, a positive SIDC is indicative of a predominantly structural language deficit (referred here as DLD), and a negative SIDC reflects social communication problems and is indicative of autistic spectrum disorder (ASD) traits (Bishop, 2003; Norbury, 2014).

Boxplots of the groups scaled scores in the ten sub-scales and a scatterplot depicting the relationship between GCC and SIDC are shown in Figure 3.15 A-B, respectively. A striking 90% of the APD children (18/20) obtained a scaled score below the 5th percentile two or more times, which has been found to indicate clinically significant communication problems (Bishop, 2003), whereas, only one such case (out of 22) was found in the TD group. The single-value GCC composite showed the exact same proportion of abnormal scores in both groups when a cut-off value of 55 was used, where only one TD child had abnormal communication skills (see Figure 3.15 B). Half of the APD children with abnormal GCC score (45%, 9/20) exhibited a score pattern that is indicative of DLD, whereas the other half exhibited a negative SIDC, indicating social communication deficits as the primary difficulty. Interestingly, out of the nine APD children who fell within the later category, three were reported by their parents to have HF-ASD diagnosis, and an additional two children were undergoing an ASD assessment at the time of testing (see scores marked with open circles and rectangles in Figure 3.15 B). Differences in GCC between the three groups (APD, and TD with/without an APD sibling) were tested using a one-way ANOVA test using *anova()* function. Parametric assumption of normal distribution and homoscedasticity were met. There was a highly significant difference between the groups [$f(2,39) = 43.712$, $p < 0.001$]. A post-hoc pairwise comparison t-tests with Bonferroni correction (*pairwise_t_test()*, rstatix package) revealed that performance of the APD group (Mdn = 42.0, SD =

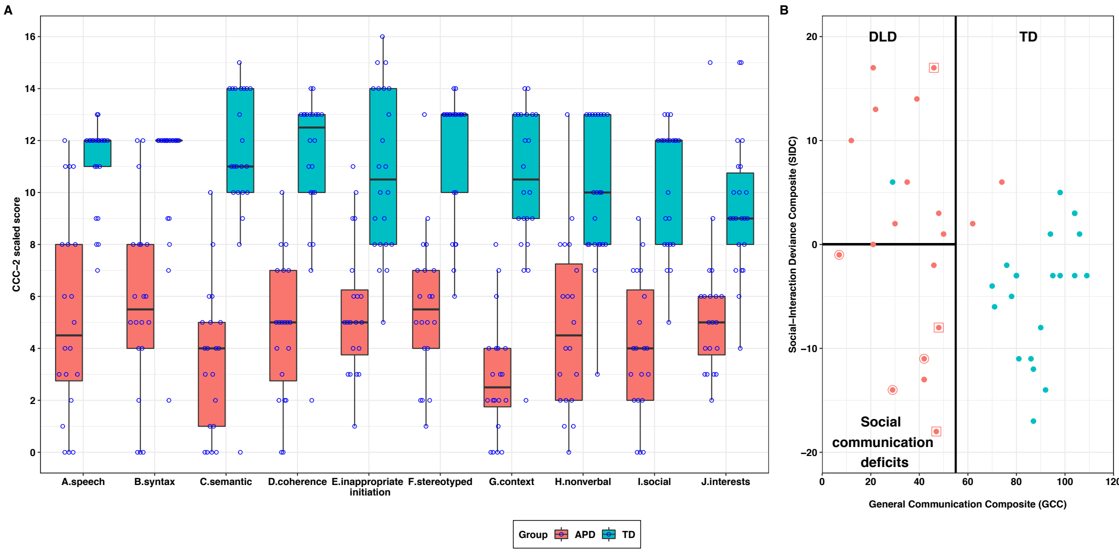

Figure 3.15: CCC-2 parental reports for the APD (red) and TD group (cyan). (A) Boxplots for scaled scores in the ten sub-scales. (B) Scatterplot for General Communication Composite (GCC) as a function of Social-Interaction Deviance Composite, (SIDC). APD children with diagnosed high-functioning Autism (HF-ASD) are denoted with open circles. APD children with undergoing ASD assessment on the day of testing are marked with open rectangles. The lines indicates the GCC cut-off criteria for typically developing children (TD) SIDC scores indicative of predominantly structural developmental language disorder (DLD) and more social communication deficits (cf. Norbury, 2013).

16.4) was significantly poorer than of the TD group with [Mdn = 90.5, SD = 25.5, t(9.4) = -4.7, $p < 0.01$] or without an APD sibling [Mdn = 88.5, SD = 11.3, t(32.0) = -10.7, $p < 0.001$]. Furthermore, there was no significant difference found between the two control groups [t(8.61) = 0.53, p = 1.00].

ECLIPS

Descriptives of the ECLiPS parental report scaled scores for the different subscales and composite measures split by groups ~~is~~ given in Table 3.22 and depicted in Figure 3.16. A score below the 10^{th} percentile (corresponding to a scale score of circa 6) is generally considered ~~as~~ clinically significant listening and processing difficulties (Barry & Moore, 2014). Overall, the ECLiPS was able to well separate between the two groups across all the different sub-scales. All APD children exhibited abnormal Total score, whereas only two TD children (out of 22) ~~obtained abnormal Total score~~.

Table 3.21: CCC-2 subscales descriptives split by groups.

	APD					TD				
Measure	N	median	sd	min	max	N	median	sd	min	max
A.speech	20	4.5	3.96	0	12	22	12.0	1.72	7	13
B.syntax	20	5.5	3.61	0	12	22	12.0	2.49	2	12
C.semantic	20	4.0	2.78	0	10	22	11.0	3.23	0	15
D.coherence	20	5.0	2.68	0	10	22	12.5	2.87	2	14
E.inappropriate.initiation	20	5.0	2.61	1	11	22	10.5	3.17	5	16
F.stereotyped	20	5.5	2.82	1	13	22	13.0	2.52	6	14
G.use.of.context	20	2.5	2.28	0	8	22	10.5	2.97	2	14
H.nonverbal	20	4.5	3.31	0	13	22	10.0	2.75	3	13
I.social	20	4.0	2.68	0	9	22	12.0	2.41	5	13
J.interests	20	5.0	2.84	2	15	22	9.0	2.63	4	15
GCC	20	42.0	16.38	7	74	22	88.5	17.38	29	109
SIDC	20	1.5	10.70	-18	17	22	-3.0	6.14	-17	6

GCC, General Communication Composite sum(A+B+C+D+E+F+G+H);
SIDC, Social Interaction Deviance Composite sum(E+H+I+J) - sum(A+B+C+D)

A closer look at the boxplots in Figure 3.16 reveals a clear difference in the distribution of the scaled scores across the two groups, with relatively larger spread for the TD group. Inspection of the Total score by groups revealed that the APD group did not follow a normal distribution and that the assumption of homoscedasticity was violated ($p < 0.05$). Thus group~~s~~ difference for the listeners' Total score was examined using a robust one-way ANOVA ~~test~~ with trimm~~ing~~ means (20%) and bootstrapping (N=2000) using *t1waybt()* function (WRS2 package; Mair & Wilcox, 2020). The test found a highly significant difference between the groups ($F = 99.35$, $p < 0.001$). A post-hoc pairwise comparison of groups with bootstrapping (N=2000) was computed using *mcppb20()* function from the same package, whereby $\hat{\psi}$ denotes the pairwise trimmed difference (Mair & Wilcox, 2020). There was a highly significant difference ($p < 0.001$) between the APD group (Mdn = 1.0 , SD = 0.91) and both TD groups with (Mdn = 12.0, SD = 3.45, $\hat{\psi}$ = -11.08, 95%-CI = -13.08 - -7.42) or without an APD sibling (Mdn = 9.0, SD = 2.47, $\hat{\psi}$ = -8.47, 95%-CI = -10.03 - -6.86), whereas no significant difference was found between the TD groups ($\hat{\psi}$ = -2.61, 95%-CI = -5.05 - 1.28, p = 0.106).

Table 3.22: ECLiPS descriptives split by groups and sub-scales.

	APD					TD				
Measure	N	median	sd	min	max	N	median	sd	min	max
SAP	20	1	0.93	0	3	23	10	2.77	4	14
L/L/L	20	2	1.55	0	5	23	10	2.78	3	15
M&A	20	3	1.10	2	6	23	11	2.69	7	16
PSS	20	4	1.53	2	8	23	10	3.37	3	15
EAS	20	3	1.64	1	8	23	9	3.52	4	14
Listening	20	2	0.93	1	4	23	11	2.69	5	15
Language	20	2	1.28	0	4	23	11	2.48	5	15
Social	20	3	1.52	1	7	23	10	3.15	5	15
Total	20	1	0.91	0	3	23	10	2.92	4	15

SAP = Speech & Auditory Processing; L/L/L = Language, Literacy & Laterality; M&A = Memory & Attention; PSS = Pragmatic & Social skills; EAS = Environmental & Auditory sensitivity; Listening = (SAP + PSS) / 2; Language = (L/L/L + M&A) / 2; Social = (PSS + EAS) / 2; Total = mean of all sub-scales

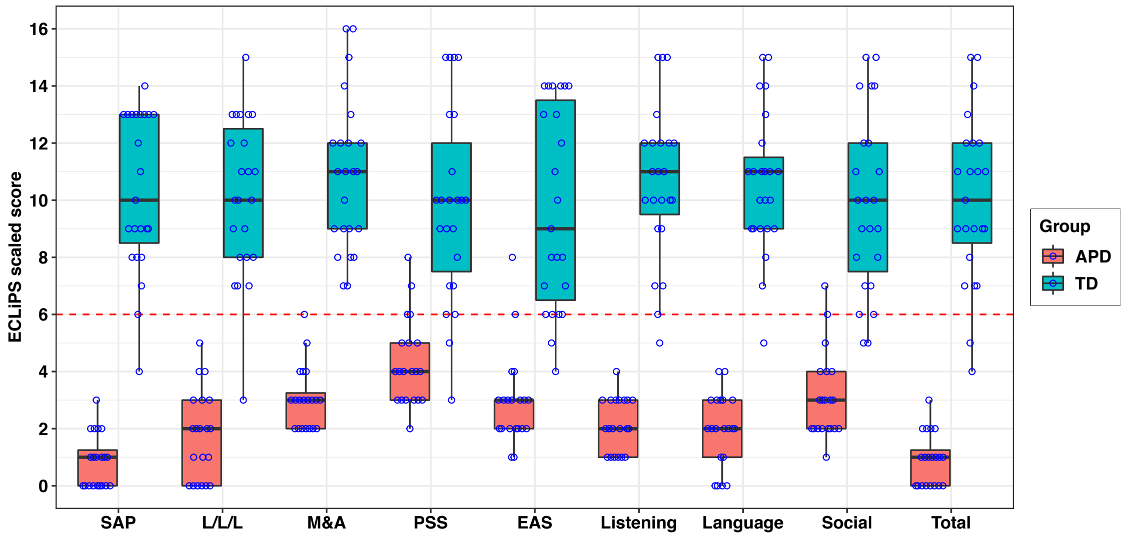

Figure 3.16: ECLiPS parental report scaled scores split by groups and sub-scales.

3.4 Overall performance

An overview of the children’~~s~~ performance split by group is given in Figure 3.17, provid~~ing~~ ~~an overlook at individuals that performed outside the norm in one or more tasks~~ (filled black cells). Abnormally poor performance for the listeners age-independent scores was defined using standardised norms for the CELF-RS, ECLiPS and the CCC-2 data or was defined as a one-tailed cut-off of ± 1.96 (where circa 97.5% of the normal population is expected to lay within) for the rest of the tasks.

Note that DLD and PLI were composed as a way to discriminate children with more structural versus pragmatic language deficit and were based on the CCC-2 data as a combination of abnormal GCC score (< 55) and the SIDC score. DLD score (developmental language disorder) denotes a combination of abnormal GCC and a positive SIDC (≥ 0) which is expected to capture severe deficit in structural language in conjunction ~~to~~ only ~~a~~ mild pragmatic difficulties. PLI score (pragmatic language impairment) on the other hand denotes a combination of abnormal GCC and a negative SIDC (≤ 0) which is expected to be a strong indicator for social communication problems with only mild structural language difficulties.

As seen in the figure, the proportion of abnormal scores across the APD group is substantially higher than in the TD group. The majority of the APD children (80%, 16/20) performed abnormally in at least two test conditions either in the ST or LiSNS-UK task, whereas there were only three cases (13%, 3/23) in the TD children. Another interesting observation is that apart from one TD child, ~~which~~ experienced difficulties in various measures including the CCC-2, none of the other TD children experienced language difficulties. This is in contrast to the APD group where 90% (18/20) of the children experienced some kind of language deficit. The CELF-RS has been reported to be a good marker for children with DLD, nevertheless, the results of the present study suggests otherwise. While performance in the APD group was noticeably poorer than in the ~~APD~~ group, only two APD children obtained abnormally poor CELF-RS score, whereas nearly half of the APD children (45%, 9/20) exhibited a CCC-2 score indicative of DLD, and about the remaining half (40% 8/20) obtained a CCC-2 score indicative of pragmatic language and social communication deficit (PLI).

Potential experimental bias of reporters when recruited due to an informed group affiliation? email Courtenay!

The proportion of abnormal scores by measure or task split by group is shown

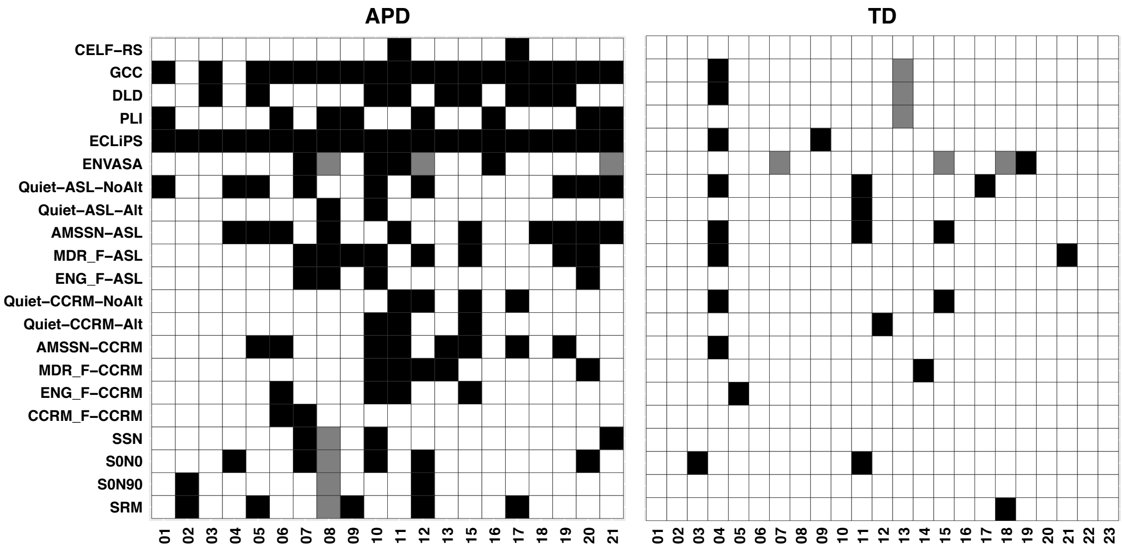

Figure 3.17: Overall perfromance: Abnormal (black cells) and normal (empty cells) performance in the ~~present study~~ test battery of individuals from the APD group (n=20) and the TD group (n=23). Missing data is marked by the grey cells.

in Figure 3.18. Both the ECLiPS total score and the CCC-2 GCC sum score resulted in the largest separation between the groups. Out of the auditory tasks, the tests conditions that resulted in the highest proportion of abnormal scores in the APD group were AMSSN (ASL: 50%, CCRM: 40/%), Quiet-ASL-NoAlt (45%) and MDR_F-ASL (40%), whereas only 26% of the APD children had abnormal SRM score.

3.4.1 Switching task: effect-size

This section is a provisional draft, trying to find the test material/condition that gave the largest separation between the groups.

Idea: calculation of d' per condition.

Effect-size [Cohen's d; *rstatix::cohens_d()*] was calculated for pairwise group comparisons by material & condition (see Table 3.23). Three test conditions resulted in a 'large' effect size: CCRM_F-Alt-CCRM (d = 1.01), AMSSN-Alt-CCRM (d = 1.00) and MDR_F-ASL-Alt (d = 0.85). Four other conditions resulted

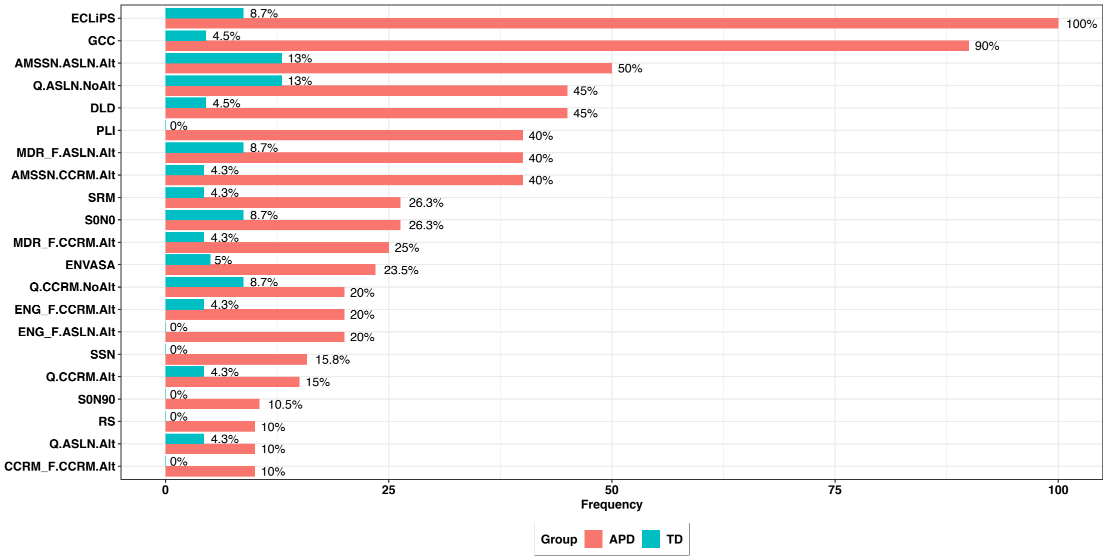

Figure 3.18: Overall perfromance: Proportion of abnormal score per measure or task split by groups.

Table 3.23: Cohen's d by condition and material.

.y.	group1	group2	effsize	material	CondCode	n1	n2	conf.low	conf.high	magnitude
z	APD	TD	0.63	ASLN	Q-ASLN-NoAlt	20	23	0.02	1.72	moderate
z	APD	TD	0.18	ASLN	Q-ASLN-Alt	20	23	-0.39	0.98	negligible
z	APD	TD	0.49	ASLN	AMSSN-ASLN-Alt	20	23	-0.15	1.57	small
z	APD	TD	0.85	ASLN	MDR_F-ASLN-Alt	20	23	0.20	1.54	large
z	APD	TD	0.64	ASLN	ENG_F-ASLN-Alt	20	23	0.05	1.31	moderate
z	APD	TD	0.34	CCRM	Q-CCRM-NoAlt	20	23	-0.26	0.97	small
z	APD	TD	0.35	CCRM	Q-CCRM-Alt	20	23	-0.27	1.07	small
z	APD	TD	1.01	CCRM	AMSSN-CCRM-Alt	20	23	0.49	1.79	large
z	APD	TD	0.76	CCRM	MDR_F-CCRM-Alt	20	23	0.20	1.38	moderate
z	APD	TD	0.64	CCRM	ENG_F-CCRM-Alt	20	23	-0.01	1.36	moderate
z	APD	TD	1.01	CCRM	CCRM_F-CCRM-Alt	20	23	0.43	1.86	large

in a 'moderate' effect-size, with d ranging between 0.63 to 0.75. These conditions comprised of speech distractors (ENG or MDR) from either material and Quiet-NoAlt for the ASL material. The test material which resulted in the largest effect-size was estimated by averaging d across conditions for each material. Both materials had a 'moderate' average effect-size, whereby the CCRM material had the largest effect-size of 0.69, following with 0.56 for the ASL material.

3.4.2 Interaction between measures

The present study involved a large number of test conditions and various measures assessing different skills. For example, the ST data alone comprises ~~of~~ 11 different conditions (x5 ASL, x6 CCRM speech material). Another set of measures consisting of the CELF-RS, ECLiPS and the CCC-2 taps into language and communication related skills, whereby the latter two consists of a sum of 15 different sub-scales ~~and~~ have been shown to strongly correlate with one another (Barry & Moore, 2014). Examining the extent to which ~~the groups~~ performance is explained by such a large number of measures will result in a very conservative significance level in order to minimise Type-I error (false positive), and could increase Type II error rate (false negative) (McDonald, 2014). Since the measures within the ST and within the language dataset are expected to strongly correlate, it was decided to use an exploratory data analysis technique ~~using~~ Principal Components Analysis (PCA). PCA is a technique used to reduce a large number of correlated parameters into a smaller set of components that together explain a considerable amount of the variability in the large dataset. ~~Whereby,~~ each of the PCA components is composed of a linear combination of the input parameters (James et al., 2013). PCA was performed separately for the ST and language data set using FactoMineR package (Lê et al., 2008) with scaled units and will be discussed separately below.

ST

PCA for the ST z-scores comprised ~~of~~ 11 input variables and a sample size of 43. Sample size adequacy for PCA was verified using Kaiser-Meyer-Olkin test (psych::KMO; Revelle, 2020), with an overall KMO of 0.76 (‘good’; Field et al., 2012), and a KMO range between 0.66 to 0.85 across the conditions. Bartlett’s sphericity test was significant [$\chi^2(55) = 190.36$, $p < 0.001$], indicating that the correlations between the different items were large enough for a PCA. Table 3.24 shows the variables loadings (no rotation was applied), their eigenvalues and percentage of variance explained. Loadings are indicators of substantive importance of a given variable to a given component (Field et al., 2012). The first three components were used, yielding eigenvalues > 1 (Kaiser's criterion), explaining together circa 67% of the variance in the data. The first component (PC.ST) accounted for the largest portion of spread in the data of 40.6% and was interpreted as an overall measure for performance in the switching task with relatively high loadings across all ~~input variables.~~ The remaining components explained each circa 16% and 11% of the variance (ascending order). Figure 3.19 illustrates the different dimensions in the data captured by the three PCA components. Clustering in the second component (PC2.Material) reflected differences in performance across the two speech materials (ASL & CCRM). The third component (PC3.Nz) reflected the degree of distractability introduced by speech distractors (MDR_F, ENG_F, CCRM_F) irrespective of the speech material used, resulting in decrement in performance when compared with non-speech distractors or target-only conditions (Quiet and AMSSN). Boxplots of the listeners weighted scores for the PCA components split by group ~~is~~ shown in Figure 3.20. PC1.ST shows ~~to separate very well~~ between the two groups, with very little overlap in scores between the TD group and the majority of the APD children. ~~Whereas~~ separation between the two groups in the remaining components ~~are~~ noticeably smaller.

Figure 3.21 illustrates the relationship between the listeners weighted scores based on the three PCA components (PC1.ST, PC2.Material and PC3.Nz) and three calculated composites composed from the listeners z-scores based on the interpretation stated above; where *ST* denoted the listeners' aggregated overall score across all ST conditions, and the two calculated discrepancy composites denoted as *Material* and *Nz*. The Material composite was calculated by subtracting the mean score of all CCRM conditions ($\overline{CCRM}$) from the mean score of all ASL conditions ($\overline{ASL}$), i.e., Material = $\overline{ASL}$ - $\overline{CCRM}$. The remaining composite, Nz, was calculated by subtracting the listeners performance averaged across conditions with speech distractors ($\overline{Spch}$) from the average performance taken across the nonspeech and quiet conditions ($\overline{NoSpch}$), i.e., Nz = $\overline{NoSpch}$ - $\overline{Spch}$. As can be seen in the figure, the PCA components highly correlated with the respective calculated composites

Table 3.24: Switching task PCA: Input variables loading.

Item	PC1.ST	PC2.Material	PC3.Nz
Q-ASLN-NoAlt	**0.59**	**0.60**	0.08
Q-ASLN-Alt	**0.61**	**0.42**	**0.43**
AMSSN-ASLN.Alt	**0.61**	**0.50**	**0.36**
MDR__F-ASLN-Alt	**0.68**	**0.36**	**-0.41**
ENG__F-ASLN-Alt	**0.69**	0.22	**-0.40**
Q-CCRM-NoAlt	**0.52**	**-0.35**	**0.56**
Q-CCRM-Alt	**0.59**	**-0.42**	0.09
AMSSN-CCRM-Alt	**0.67**	**-0.49**	0.17
MDR__F-CCRM-Alt	**0.72**	**-0.34**	-0.11
ENG__F-CCRM-Alt	**0.72**	**-0.34**	-0.16
CCRM__F-CCRM-Alt	**0.58**	-0.12	**-0.41**
eigenvalue	4.46	1.73	1.21
variance (%)	40.52	15.72	10.98
cumulative variance (%)	40.52	56.24	67.22

|loading| >0.3 are highlighted in bold.

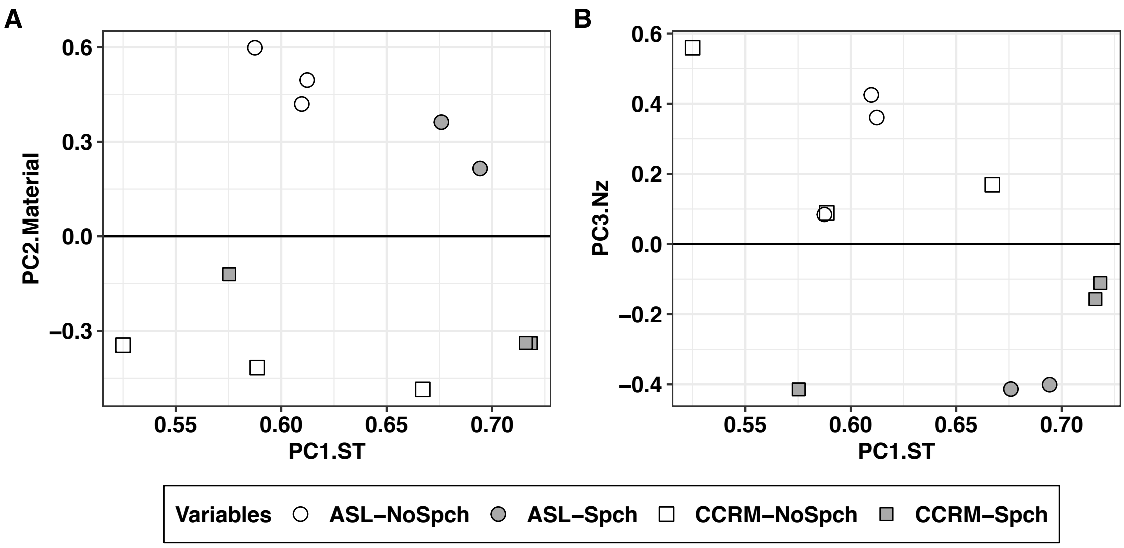

Figure 3.19: Switching task PCA: Scatterplot for the input variables as a function of PCA components: PC1.ST vs. PC2.Material (A), PC1.ST vs. PC3.Nz (B). Loadings for ASL conditions are indicated by circles and loadings for CCRM conditions are indicated by rectangles. Filled shapes denotes conditions with speech distractors (Spch) and non-filled shapes denote nonspeech conditions (No-Spch).

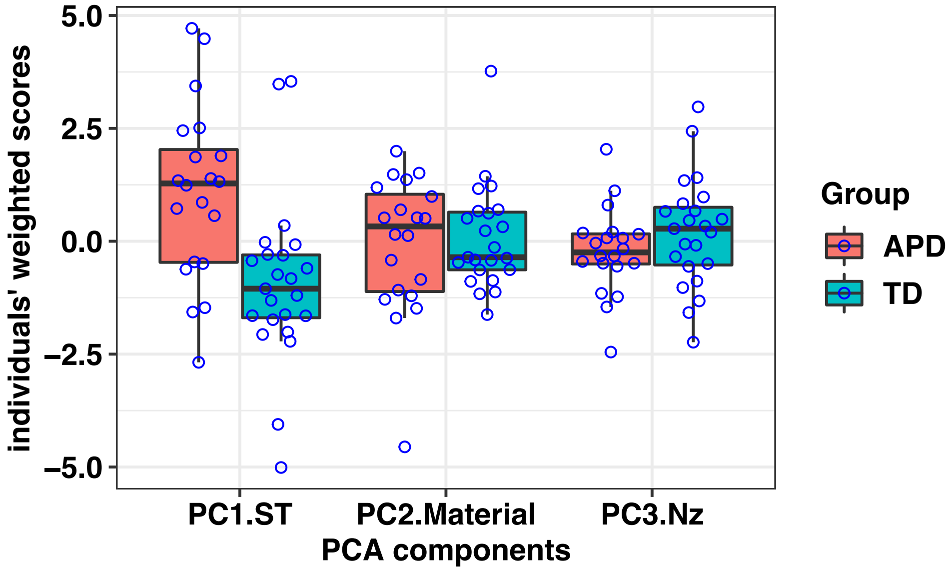

Figure 3.20: Switching task PCA: Listeners weighted scores split by components and group.

(PC1.ST - ST, PC2.Material - Material, PC3.Nz - Nz), whereas none correlated with another composite, thus indicating that the components are independent from one another and that each describe different dimensions within the data.

Language measures

PCA with three components was computed for the listeners scaled scores obtained in the different language measures, comprising of 16 input variables (x1 CELF-RS, x5 ECLiPS, x10 CCC-2) with a sample size of 42. Data for one TD child was excluded from the analysis due to inconsistent CCC-2 responses. Kaiser-Meyer-Olkin test for sample-size adequacy was 'superb' (Field et al., 2012) with an overall KMO of 0.93 (range: 0.86 - 0.97) and the assumption of sphericity was verified using Bartlett's sphericity test [$\chi^2(120) = 787.52$, $p < 0.0001$]. The PCA variables loadings, eigenvalues and percentage of variance explained split by components is given in Table 3.25. The first component (PC1.Lang) yielded eigenvalue > 1, explaining circa 73% of the variance, reflecting an overall performance averaged across all the language measures. The remaining components had eigenvalue of just under 1 (0.95 & 0.85, respectively), each explaining circa 6% and 5% of the variance. The second

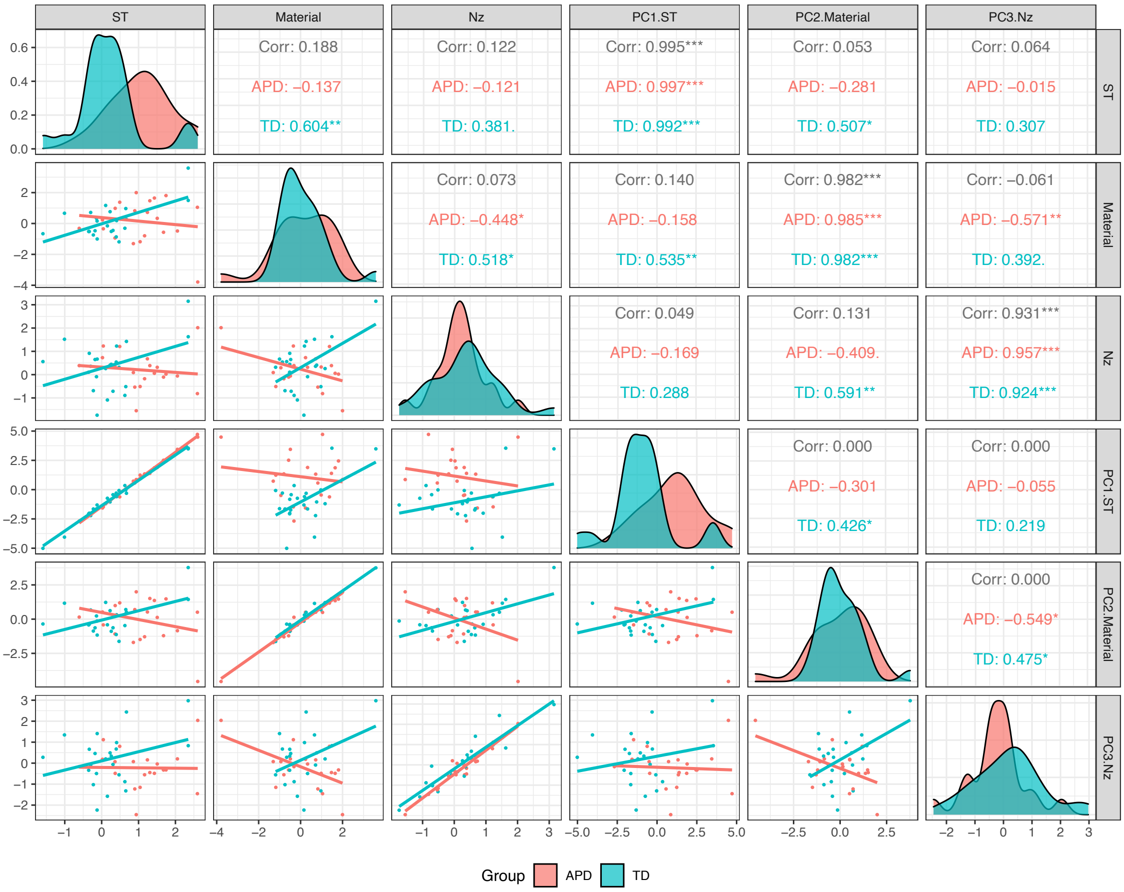

Figure 3.21: Switching task PCA: Comparison between PCA weighted scores and calculated measures: (1) ST = mean score across all ST data, (2) Material = $\overline{ASL}$ - $\overline{CCRM}$, (3) Nz = $\overline{NoSpch}$ - $\overline{Spch}$.

component (PC2.Lang) reflected discrepancy between expressive language skills, measured by the CELF-RS and listening and communication skills measured by the ECLiPS subscales. Interestingly, the third component (PC3.Lang) reflected once again a discrepancy, clustering together variables that taps onto pragmatic language and social interaction skills such as the ECLiPS subscale PSS (pragmatic & social skills) and the CCC-2 subscales E, H, I & J, separating them from other variables that assess more structural language skills such as the CELF-RS and the CCC-2 subscales speech (A) and Syntax (B). Boxplots of the listeners weighted scores for the PCA components split by group is shown in Figure 3.22. As seen in the ST data, the first component (PC1.Lang) best separated ~~between~~ the two groups, whereas separation between the two groups in the remaining components ~~were~~ noticeably smaller.

Table 3.25: Language measures PCA: Input variables loading.

Item	PC1.Lang	PC2.Lang	PC3.Lang
CELF-RS	**0.69**	**0.40**	**0.37**
ECLIPS.SAP	**0.91**	**-0.32**	0.14
ECLIPS.LLL	**0.92**	-0.14	0.11
ECLIPS.M.A	**0.88**	**-0.30**	0.14
ECLIPS.PSS	**0.83**	**-0.36**	-0.17
ECLIPS.EAS	**0.79**	**-0.52**	0.07
CCC2.A.speech	**0.78**	0.08	**0.35**
CCC2.B.syntax	**0.82**	0.19	0.27
CCC2.C.semantic	**0.92**	0.14	0.05
CCC2.D.coherence	**0.92**	0.09	0.05
CCC2.E.inappropriate.initiation	**0.82**	0.13	**-0.41**
CCC2.F.stereotyped	**0.89**	0.22	-0.03
CCC2.G.use.of.context	**0.93**	0.04	-0.08
CCC2.H.nonverbal	**0.84**	0.13	-0.26
CCC2.I.social	**0.88**	0.15	-0.16
CCC2.J.interests	**0.80**	0.11	**-0.40**
eigenvalue	11.67	0.95	0.85
variance (%)	72.96	5.95	5.3
cumulative variance (%)	72.96	78.91	84.21

|loading| >0.3 are highlighted in bold.

Despite the small proportion of variance explained by the later two principal components, they ~~yet~~ capture other aspects of language and communication skills that may be relevant in explaining the individual and group differences in the auditory tasks and were therefore included in the analysis. Nevertheless, interpretation of the relationship between these components with performance in the auditory tasks should be viewed with caution. Inspection of the individuals' scaled scores split by groups for loadings in PC1.Lang as a function of loadings in PC2.Lang and PC3.Lang shown in Figure 3.23 A-B revealed a linear relationship between PC1.Lang and PC2.Lang (APD group) and between PC1.Lang and PC3.Lang (TD group), thus indicating that they are not entirely independent from one another. The partial lack of independence may be in part explained by the large ~~polarity~~ in scores between the groups across the different input variables.

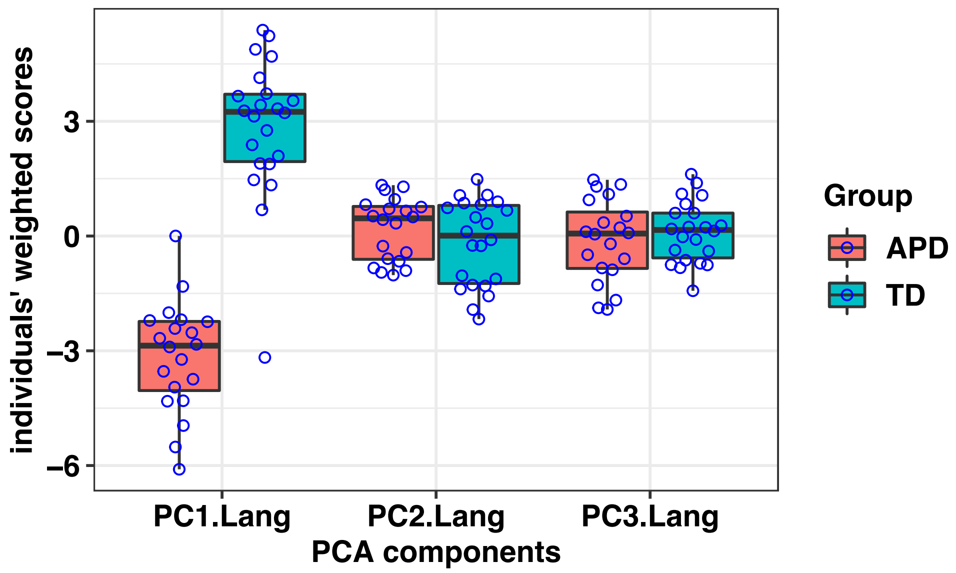

Figure 3.22: Language measures PCA: Listeners weighted scores split by components and group

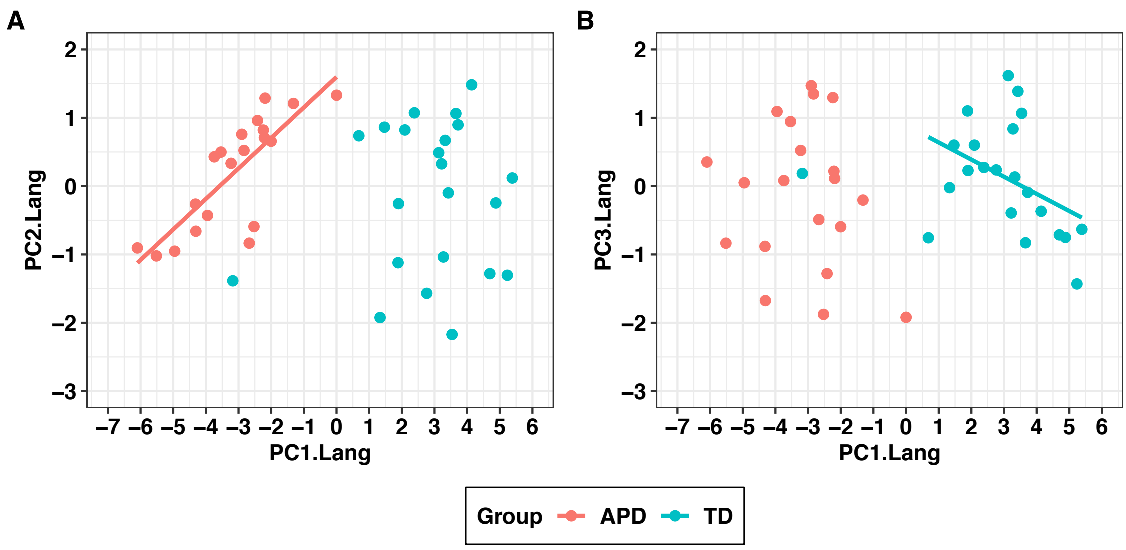

Figure 3.23: Language measures PCA: Individual scores split by groups for loadings in PC1.Lang as a function of scores for PC2.Lang (A), and PC3.Lang (B).

Again, the relationship between the PCA components (PC1.Lang, PC2.Lang and PC3.Lang) and the three calculated composites that reflects the components interpretations is illustrated in Figure 3.24. The calculated components were based on the listeners scaled scores, where *Lang1* represents the overall performance aggregated across all the language scores, *Lang2* represents discrepancy between

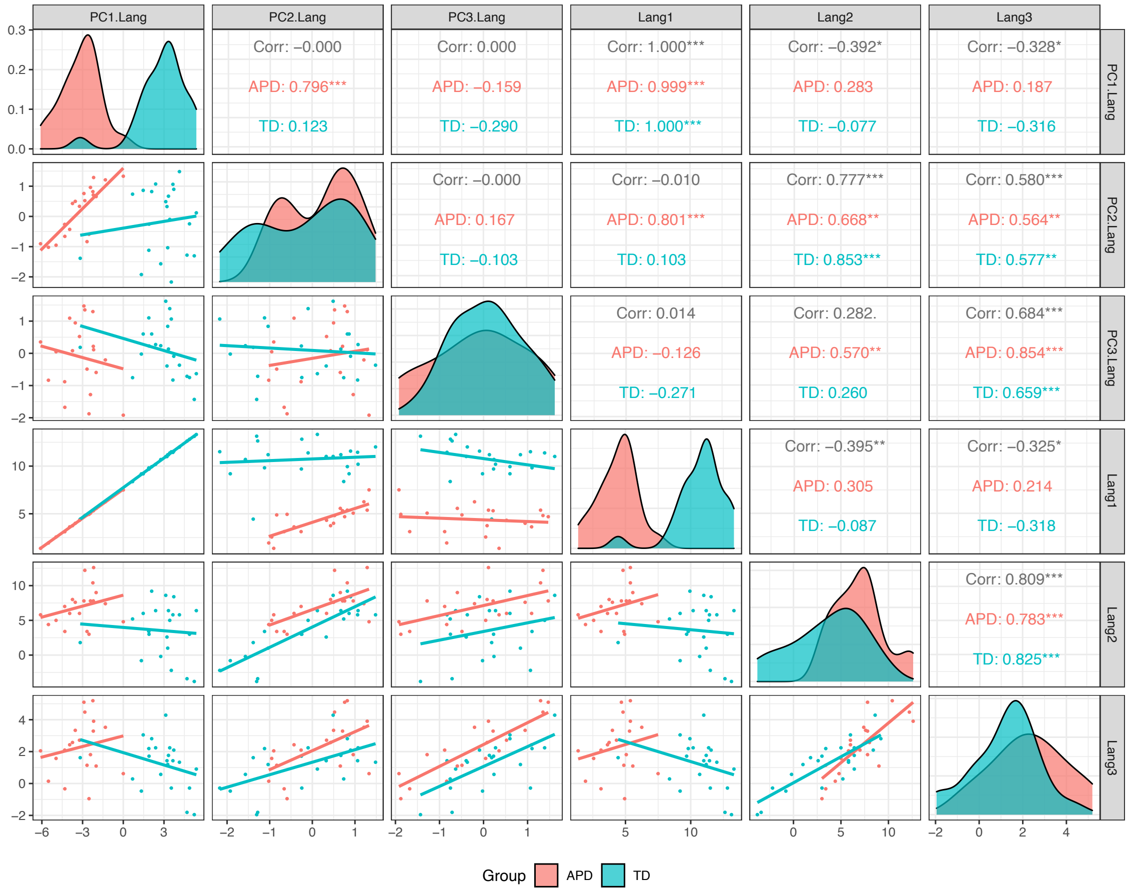

Figure 3.24: Language measures PCA: Comparison between the listeners weighted scores by components, PC1.Lang - PC3.Lang (A), and calculated measures, Lang1 - Lang3 (B).

expressive language skills (CELF-RS) and listening and communication skills (all ECLiPS subscales), and lastly, *Lang3* stands for discrepancy between structural and pragmatic & social skills. As seen in the figure, correlations were high between each PCA component and the corresponding calculated composites (range: 1 to 0.58).

Discussion(?): Which measures were best described by the PCA?

All the measures showed strong correlation with PC1.Lang, whereas the CCC-2 GCC score showed the largest correlation (ρ = 0.98, p <.0001). This was true not only for the data aggregated across groups, but also when correlations were examined separately in each group. Therefore, taking into account the short administration time and simplicity, the CCC-2 alone provides a good screening tool for children's language and communication skills with high levels of sensitivity and specificity. Nonetheless, children in the present study knowingly consented to take part in the study either as part of the clinical APD group or the control group, which may introduced bias in the reporters response, and may resulted in a larger separation between the two groups than one would expect across the true population.

Correlations

Next, the extent to which individual differences in speech perception could be explained by other measures was examined for the aggregated data across the two groups with multiple Spearman's rho correlations using *rcorr* function (Hmisc R package; Harrell Jr, 2020) between SSN scores, LiSNS-UK scores for the spatialised conditions and the derived score for spatial release from masking (S0N0, S0N90 & SRM), the principal components for the switching task PC1.ST, PC2.Material and PC3.NZ, and for the language measures PC1.Lang, PC2.Lang and PC3.Lang, average PTA at standard audiometry frequency ~~bands~~ (0.5-4 kHz), average PTA at high-frequency ~~bands~~ (PTA_{EHF}, at 8, 11 and 16 kHz), and ENVASA total score as a measure for sustained and selective-attention skills. Age effect was accounted for either by using standardised norms when available or by a regression model based z-score transformation. The correlation matrix outcomes are given in Table 3.26.

Table 3.26: Correlation matrix (Spearman) between the study test measures for aggregated data across the two groups.

	PTA	PTA_{EHF}	ENVASA	SSN	S0N0	S0N90	SRM	PC1.ST	PC2.Material	PC3.Nz	PC1.Lang	PC2.Lang
PTA												
PTA_{EHF}	0.31											
ENVASA	-0.10	-0.13										
SSN	0.26	0.01	-0.40*									
S0N0	0.26	0.02	-0.19	0.39*								
S0N90	0.45**	0.34*	-0.23	0.30	0.64****							
SRM	-0.39*	-0.36*	0.12	-0.07	0.08	-0.67****						
PC1.ST	0.46**	0.09	-0.27	0.46**	0.34*	0.44**	-0.23					
PC2.Material	-0.17	-0.14	0.01	0.30	0.34*	0.20	0.12	0.06				
PC3.Nz	-0.03	0.03	0.05	0.09	-0.10	-0.10	-0.03	-0.11	0.01			
PC1.Lang	-0.16	-0.07	0.46**	-0.51***	-0.19	-0.15	-0.02	-0.55***	-0.03	0.16		
PC2.Lang	0.07	0.07	0.12	-0.02	0.21	0.23	-0.14	-0.01	0.16	-0.04	0.08	
PC3.Lang	-0.10	-0.26	0.00	0.09	-0.03	-0.12	0.08	-0.02	0.09	-0.14	-0.05	-0.02

significant p-values: **** p < .0001, *** p < .001, ** p < .01, * p < .05

There was a significant correlation between the listeners overall performance in the switching task (PC1.ST) and their language skills (PC1.Lang; ρ = -0.55, $p < 0.001$), PTA ($\rho = 0.46$, $p < 0.01$), speech perception in noise (SSN; $\rho = 0.46$, $p < 0.01$), and the spatialised LiSNS-UK test conditions S0N0 ($\rho = 0.35$, $p < 0.05$) and S0N90 ($\rho = 0.45$, $p < 0.01$). The second ST principal component, PC2.Material, significantly correlated with S0N0 ($\rho = 0.33$, $p < 0.05$) and SRM ($\rho = 0.33$, $p < 0.05$), whereas no relationship was found between the third PC3.Nz and any of the study measures.

Performance in the LiSNS-UK exhibited the highest correlation coefficients, with highly significant correlation between S0N0 and S0N90, where better performance in one condition was highly associated with better performance in the other ($\rho = 0.64$, $p < 0.0001$), and between S0N90 and SRM (ρ = -0.67, $p < 0.0001$), where better SRM was predicted by better performance for S0N90, whereas correlation between S0N0 and SRM was not significant ($\rho = 0.08$, $p = 0.62$). Note that lower z-score in the spatialised conditions denotes better performance, whereas the opposite holds for SRM, with higher z-scores marking better performance, ~~which~~ explain~~s~~ the negative correlation between SRM and S0N90. A separate group-wise analysis gave similar results for correlation between S0N90 and SRM, whereas correlations in the APD group between S0N0 and S0N90, and between S0N0 and SRM were smaller and not significant (ρ: 0.35 and 0.30, respectively). The non-significant correlation between SRM and S0N0 stands in contrast~~s~~ to our expectations, for a positive correlation, where listeners with poorer (i.e., higher) S0N0 score were expected to have a larger (i.e., better) SRM. The insignificant and reduced correlation in the APD group is likely due to sampling error and due to the small sample size in the present study (correlation between the LiSNS-UK condition for the listeners SRT and z-scores are given in appendices in Figures C.5 and C.6).

SSN score was found to be related to performance in the two spatialised LiSNS-UK test conditions with correlation coefficients of 0.30 (S0N90) and 0.39 (S0N0)~~.~~

however only correlation for S0N0 was significant ($p < 0.05$), while p-value for correlation with S0N90 was just above the significance level ($p = 0.055$). The listeners S0N90 score significantly correlated with hearing sensitivity thresholds measured at both standard (PTA; $\rho = 0.45$, $p < 0.01$) and extended frequency ~~bands~~ (PTA_{EHF}; $\rho = 0.34$, $p < 0.05$). Moreover, none of the LiSNS-UK measures significantly correlated with the language principl~~e~~ components PC1.Lang - PC3.Lang or the attention measure ENVASA. Additional significant correlations were found between PC1.Lang and SSN (ρ = -0.51, $p < 0.0001$) and between PC1.Lang and the ENVASA task ($\rho = 0.46$, $p < 0.001$). No p-value Bonferroni correction for multiple comparisons was applied.

Exploratory predictors – APD group

Association of potential predictors with performance in the APD group was examined in the following section. Nevertheless, it is important to emphasise that this is an exploratory examination across a small sample size and thus the outcomes may not be generalised in a larger sample. Predictors were selected based on the caregivers response in the background questionnaire, where the APD children were subdivided into the following pair of groups: 1. APD diagnosis (APD vs. LiD), 2. SPD diagnosis (SPD vs. non-SPD), 3. Regular use of FM-device (FM vs. No FM), 4. History of middle ear problem (MEHx vs. No MEHx), 5. Pressure equalisation tube history (PET vs. No PET), and 6. Auditory training (Training vs. No training). The listeners performance subdivided by predictors is shown in Figure 3.25 for data measured with the ST task (PCA1.ST), the language composite (PCA1.Lang), SRM, and thresholds for standard audiometry (PTA) and EHF audiometry (EHF PTA). Individual observations are marked in circles, whereby observations of children diagnosed with APD are filled in dark blue, and LiD observations are filled in light blue. Individual data for the TD group is marked in black. From the boxplots, PET and MEHx emerges as the best predictors, explaining the largest portion of the within group differences. History of PET showed the highest association with poorer EHF PTA thresholds, and to a relatively smaller extent with PC1.ST (higher score indicates poorer performance) and with the SRM score (higher score indicates better performance). Consequently, it is not surprising that a related predictor – history of middle ear problem (MEHx) was also highly related to poorer EHF PTA thresholds; nevertheless, association between MEHx and the other measures was weak. Interestingly, there was no association between SRM score and a diagnosis of SPD, with only a small difference between APD children with or without an SPD diagnosis.

MEHx: is a composite calculated based on the caregivers indication of history of **Ear infection** & **Glue ear** in the background questionnaire. MEHx is 1 if response was 'Yes' to at least one of these items, whereas MEHx is 0 if response was 'No' for both items.

Exploratory predictors – TD group

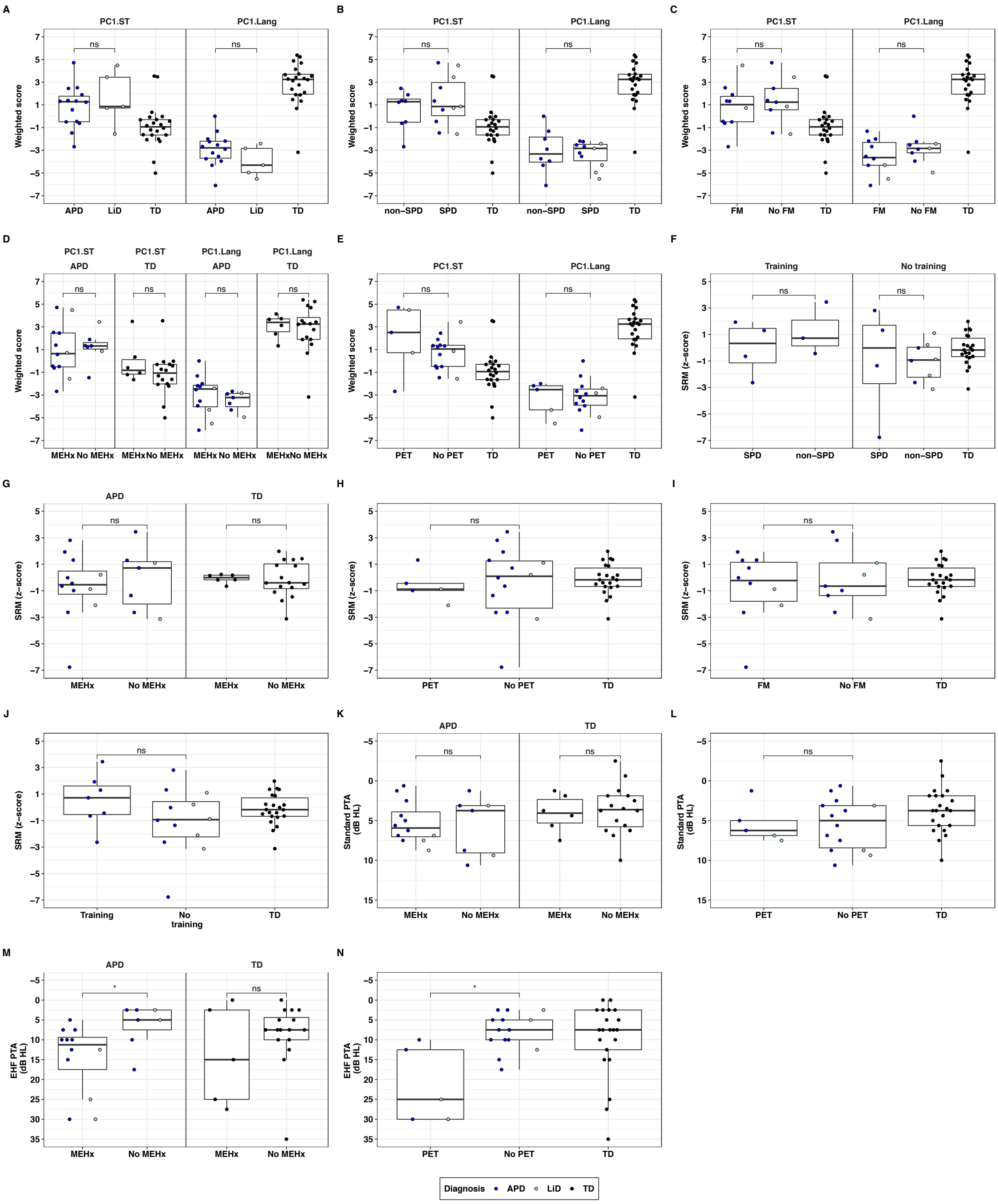

Figure 3.25: Association between predictors and performance in the APD group for the switching task composite (PC1.ST), language composite (PC1.Lang), SRM, standard and EHF PTA. Predictors included: 1. APD diagnosis (APD vs. LiD), 2. SPD diagnosis (SPD vs. non-SPD), 3. Regular use of FM-device (FM vs. No FM), 4. History of middle ear problem (MEHx vs. No MEHx), 5. Pressure equalisation tube history (PET vs. No PET), and 6. Auditory training (Training vs. No training). Individual observations are marked in circles. Observations of children diagnosed with APD are filled in dark blue, and LiD observations are filled in light blue. TD group observations are marked in black. Significant p-values for independent t-test comparison are marked with asterisk ($p < 0.05$).

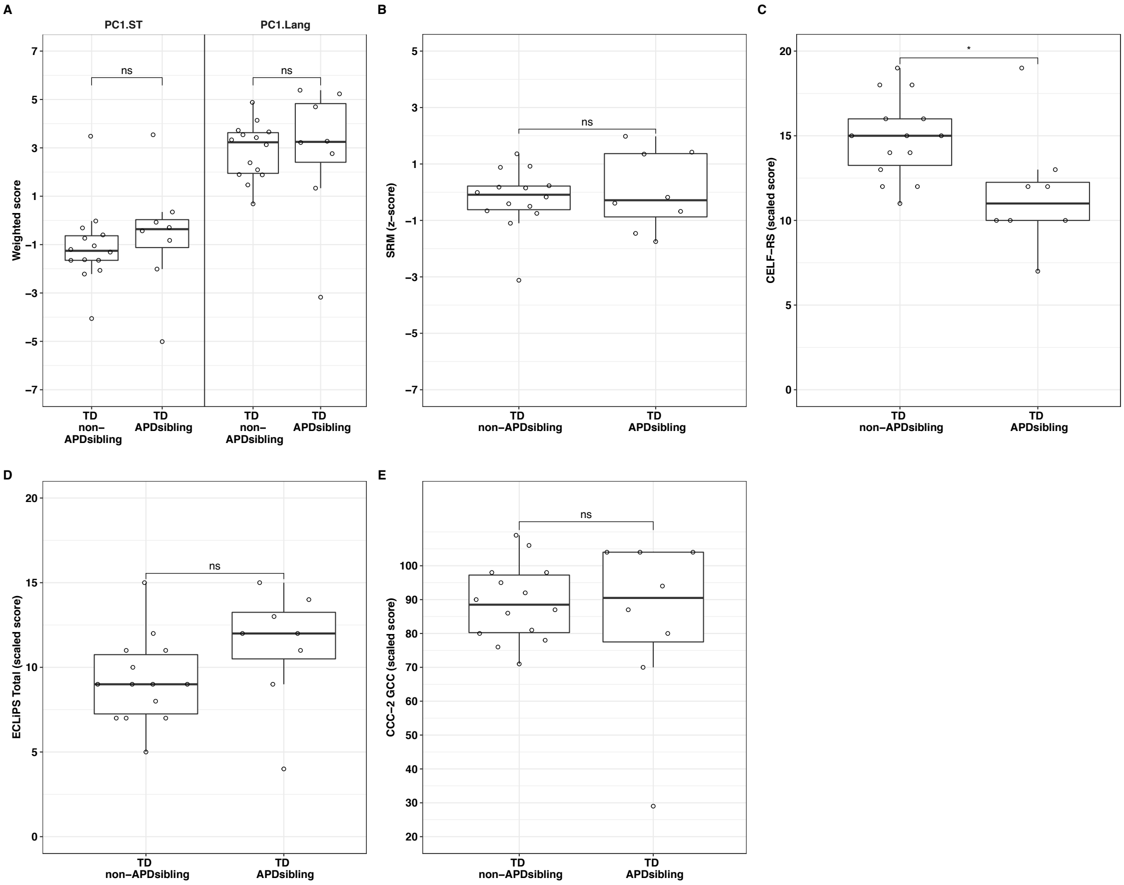

Figure 3.26: Add text here. Significant p-values for independent t-test comparison are marked with asterisk ($p < 0.05$).

3.5 Discussion

3.5.1 EHF

Lee's thresholds for 10-21 yrs group: 8=16.35 (1.46-29.33); 11=22.99, 16=48 (20.01-91.35); 20=93.07 (48.57-105.00) all dB SPL

EHF in children: Read Schechter et al., 1986:

- 6-10 yrs: 10k=23, 12k=20, 16k=39 dB SPL
- 11-15 yrs: 10k=21, 12k=22, 16k=51 dB SPL

3.5.2 ST

APDsibling While the causes of APD are not fully understood, amongst others, studies have shown a strong association between APD and history of hearing problems (e.g., due to chronic OME) causing auditory deprivation, and Developmental Language Deficit (DLD). Moreover, several studies demonstrated that these factors are genetically influenced (Pennington & Bishop, 2009; Bishop, Adams, & Norbury, 2006), while others suggested that some APD-linked aspects (tone sequencing) are more environmentally driven (cf. Moore 2007; Bishop, 2002). Studying twins or siblings is a usefull way to examin both genetic and environmental factors. Although the inclusion of control children with APD sibling(s) was not part of the study design and was not carefully balanced for the control group sample size, it is possible that the two control groups would preform differently in the study measures. Whether the influencing factors were mostly heritable or accquiered by the child's environment, we hypothesised that the control children with APD sibling will perform poorer than the non-sibling control children.

Why CCRM performance is better

The improved intelligibility in the CCRM material is amongst others due to the more simple speech material, the reduced confusion between the target sentences and the connected speech distractors as well as the restricted alternative responses of the CCRM matrix-based sentences.

z- scores by material: proportion of abnormal TD kids:

ASL: The proportion of abnormal scores amongst the TD group ranged between 0% to 13% (mean = 7.8%), which is relatively higher than expected in the normal population. Nonetheless, when taking into account TD observations that were trimmed during the z-score calculation procedure, the proportion of abnormal scores are smaller, ranging between 0% to 9.5% (mean = 3.8%), which corroborate fairly well with the theoretical probability of 2.5% (one-tailed).

CCRM: The percentage of abnormal scores in the TD group were relatively low ranging between 0 to 8.\7% (mean = 4.3%) and were at 0% across all conditions when TD observations that were trimmed as part of the z-score calculation procedure were accounted for.

Why there was no interaction between Group x Condition x Material? Discussion or here? The lack of significant interaction (Group x Condition or Group x Condition x Material), is somewhat surprising and do not reflect some of the differences seen in Figure 3.7 A-B between the two groups in some conditions or the overall difference in performance between the speech materials and may suggest that the model was under-powered to test these questions.

Points for age effect:

- Goldsworthy et al. 2018 found that age explained only a small portion of variability in speech perception performance (n.s.) for Quiet, SSN and 2-talker connected-speech distractors (children aged 5-17). See table 3.

Points for SSN:

- “Despite mature peripheral encoding, school-children have more difficulty understanding speech in noise compared with adults. For example, 5-7 year-old children require 3 to 6 dB more favourable SNR than adults to achieve comparable speech detection, word identification, or sentence recognition performance in a speech-shaped noise maker (e.g., Corbin et al., 2016)” [Leibold, Buss and Calandruccio, 2019, Acoustics today]. - “Speech recognition gradually improves until 9-10 years of age , after which mature performance is generally observed” [Leibold, Buss and Calandruccio, 2019, Acoustics today].

- SSN age effect in other studies are smaller

3.5.3 CCC-2

3.5.4 ECLiPS

Discussion: Correlation with CCC-2 sub-scales (Barry & Moore, 2014): Overall, all the EClIPS sub scales shows strong correlation with most of the CCRM 10 sub-scales. Interestingly, PSS strongly correlates with all 10 CCC-2 sub-scales, suggesting that both tests taps into similar abilities.

In the results: compare scores with scores obtained by: https://www.nature.com/articles/s41598-018-25316-9.pdf and Moore et al. 2020 (Listening Difficulties in Children: Behaviour and Brain Activation Produced by Dichotic Listening of CV Syllables)

Discussion: - Compare data with Ferguson et al. 2011

3.6 Conclusion

Alles Gescheite ist schon gedacht worden.
Man muss nur versuchen, es noch einmal zu denken.

All intelligent thoughts have already been thought;
what is necessary is only to try to think them again.

— Johann Wolfgang von Goethe (von Goethe, 1829)

General discussion

If we don't want Conclusion to have a chapter number next to it, we can add the `{-}` attribute.

More info

And here's some other random info: the first paragraph after a chapter title or section head *shouldn't be* indented, because indents are to tell the reader that you're starting a new paragraph. Since that's obvious after a chapter or section title, proper typesetting doesn't add an indent there.

Summary of main findings

Conclusion

Appendices

A
The First Appendix

This first appendix includes an R chunk that was hidden in the document (using `echo = FALSE`) to help with readibility:

In 02-rmd-basics-code.Rmd

```
library(tidyverse)
knitr::include_graphics("figures/chunk-parts.png")
```

And here's another one from the same chapter, i.e. Chapter ??:

B
The Second Appendix

C
The Third Appendix

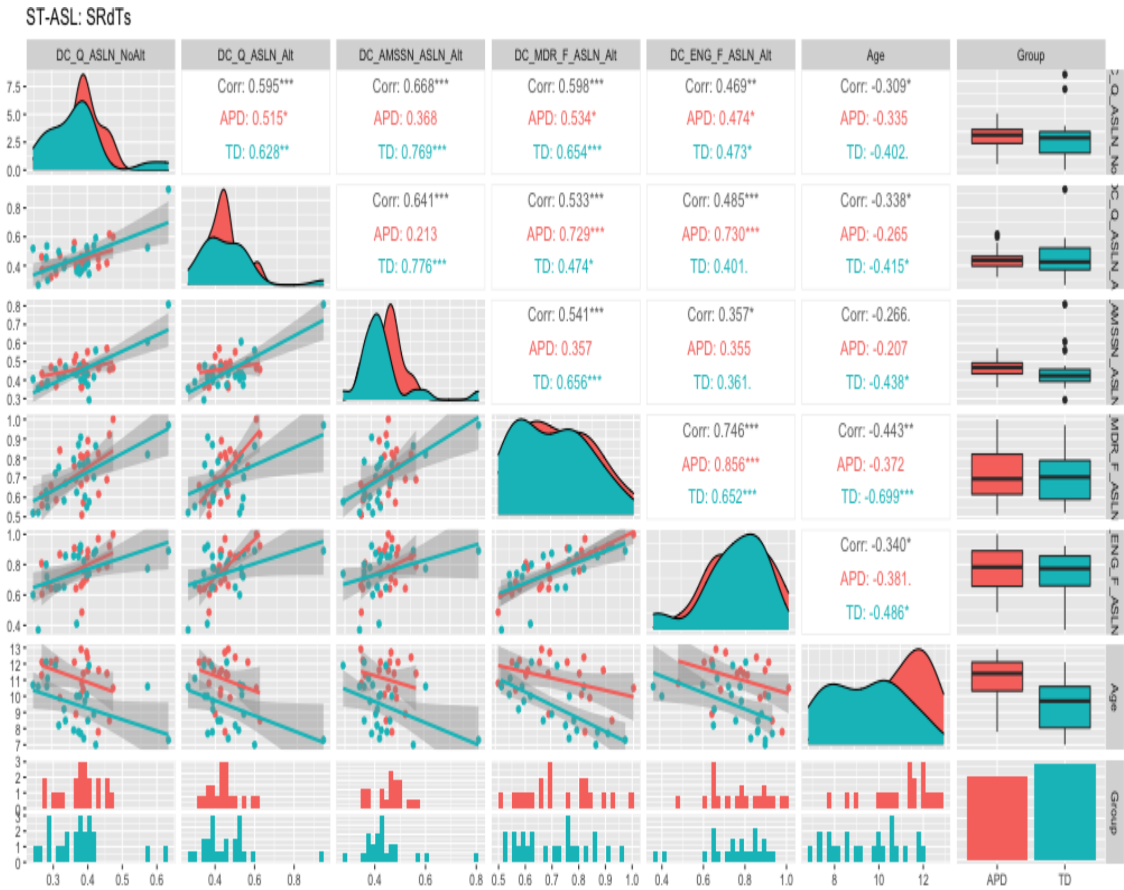

Figure C.1: Switching task: ASL speech material - correlations for listeners SRdTs (proportion of duty cycle).

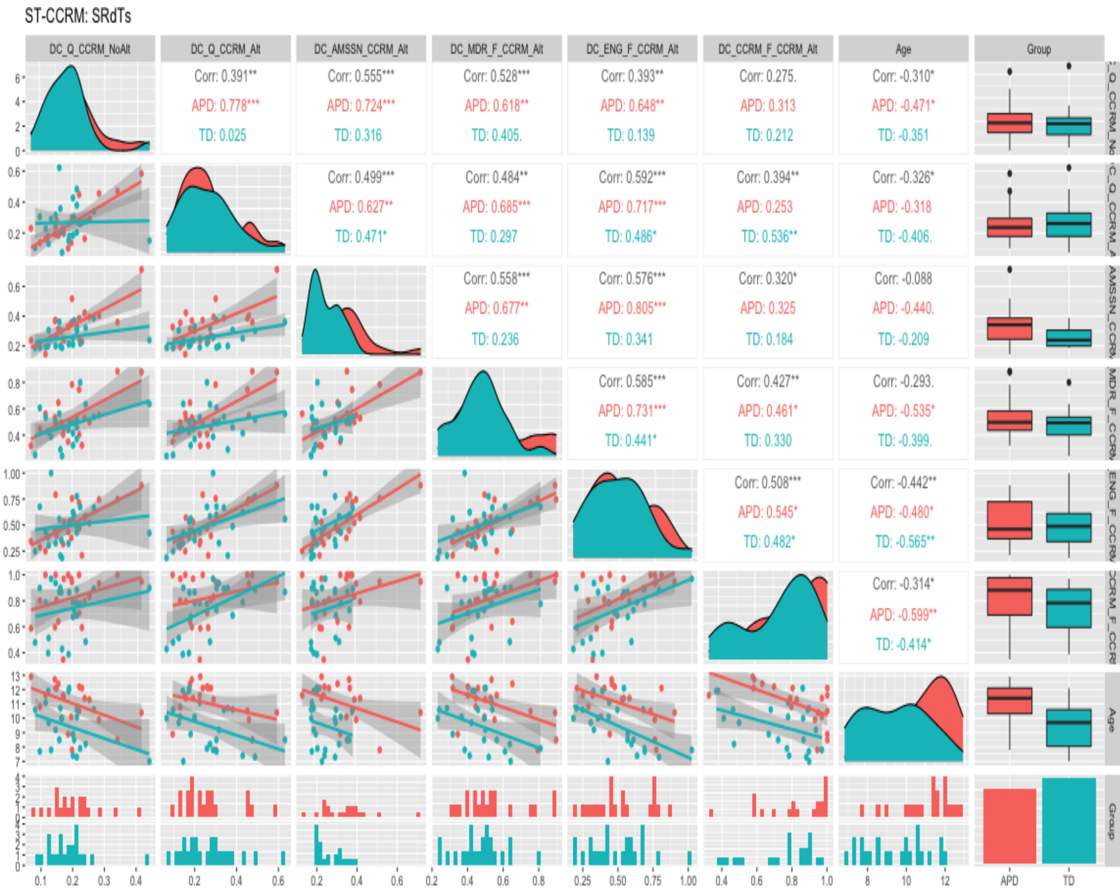

Figure C.2: Switching task: CCRM speech material - correlations for listeners SRdTs (proportion of duty cycle).

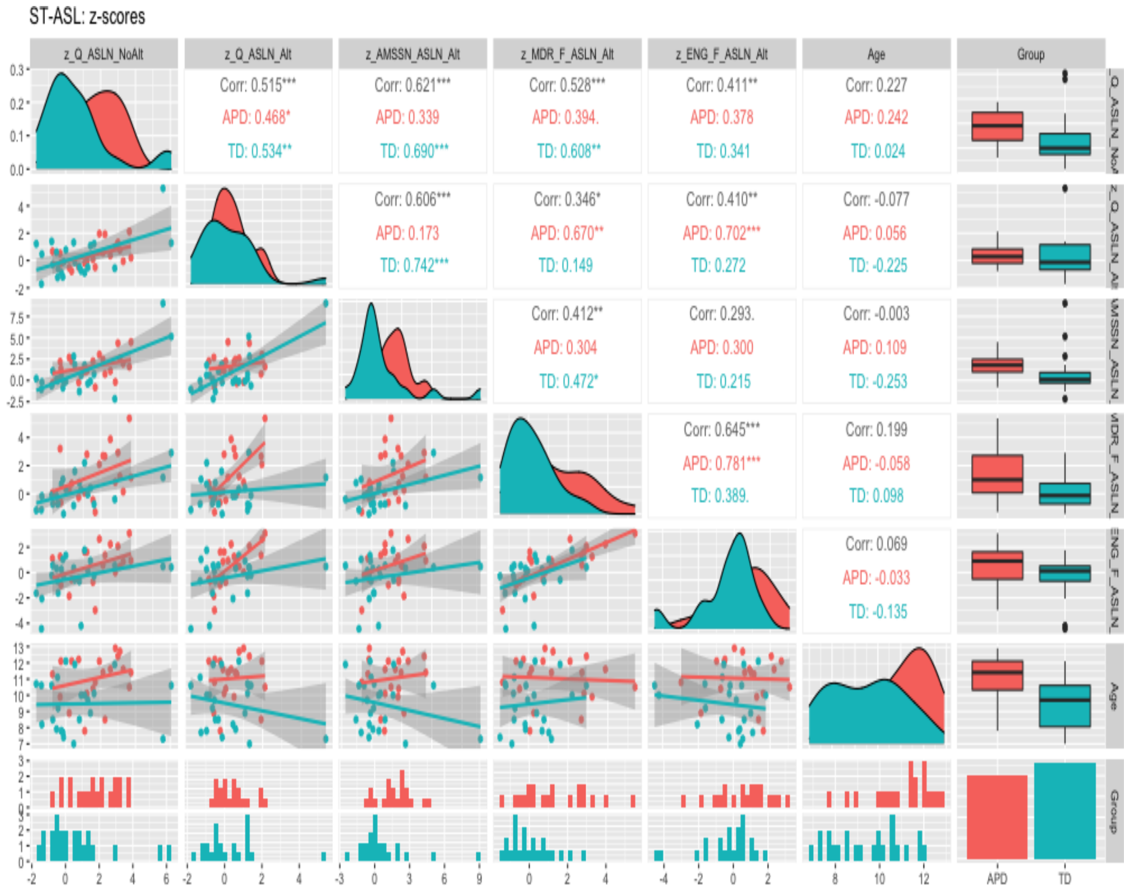

Figure C.3: Switching task: ASL speech material - correlations for listeners z-scores.

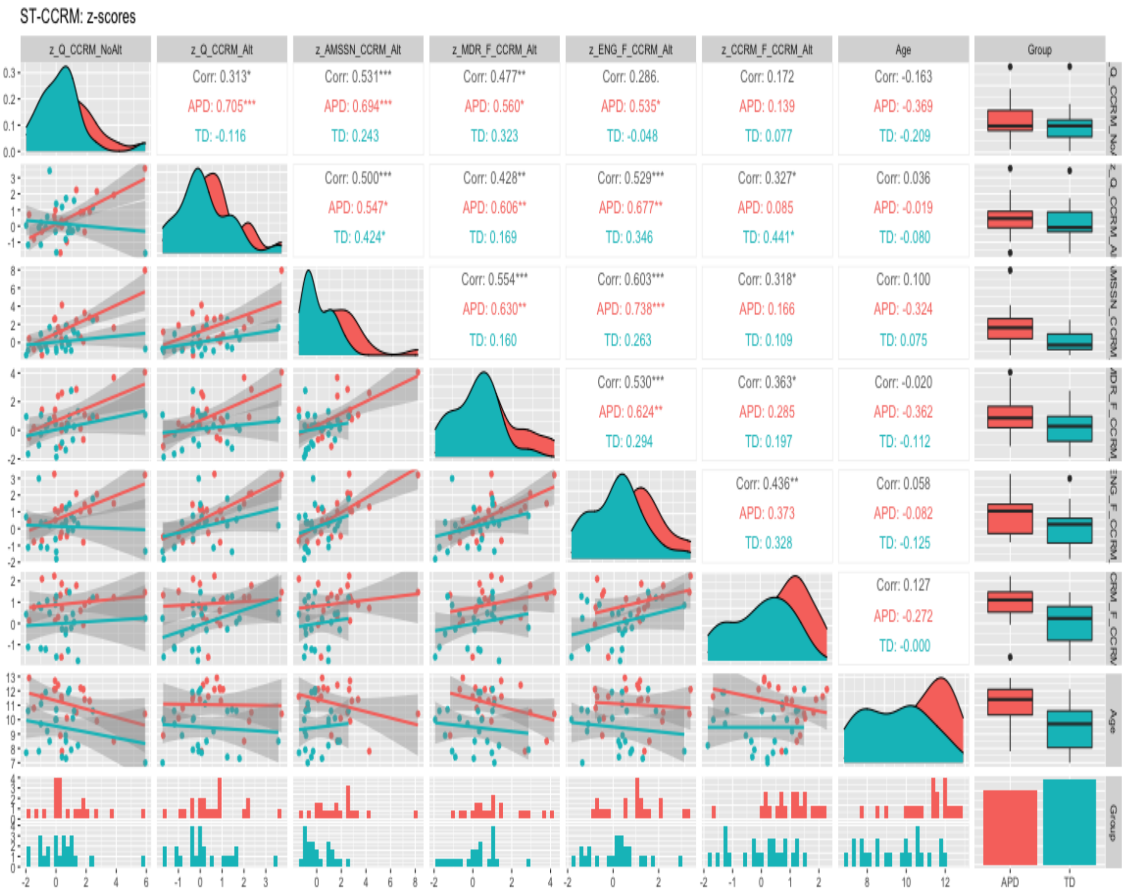

Figure C.4: Switching task: CCRM speech material - correlations for listeners z-scores.

Figure C.5: LiSNS-UK: Correlations for listeners SRTs (dB SNR).

Figure C.6: LiSNS-UK: Correlations for listeners age-independent z-scores.

References

Age changes in pure-tone hearing thresholds in a longitudinal study of normal human aging. (1990). *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.399731

Akeroyd, M. A. (2008). Are individual differences in speech reception related to individual differences in cognitive ability? A survey of twenty experimental studies with normal and hearing-impaired adults. *International Journal of Audiology*, *47*(SUPPL. 2).

Akinseye, G. (2015). *The perception of interrupted and speech in older and younger adults with normal hearing.* (unpublished BSc thesis). University College London, UCL.

Algazi, V. R., Duda, R. O., Thompson, D. M., & Avendano, C. (2001). The CIPIC HRTF database. *IEEE ASSP Workshop on Applications of Signal Processing to Audio and Acoustics*. https://doi.org/10.1109/aspaa.2001.969552

Arlinger, S., Lunner, T., Lyxell, B., & Kathleen Pichora-Fuller, M. (2009). The emergence of cognitive hearing science. *Scandinavian Journal of Psychology*, *50*(5), 371–384. https://doi.org/10.1111/j.1467-9450.2009.00753.x

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*. https://doi.org/10.1016/j.jml.2012.11.001

Barry, J. G., & Moore, D. R. (2014). *Evaluation of Children's Listening and Processing Skills (ECLiPS)* (tech. rep.). MRC-T. London, United Kingdom.

Bashford, J. A., Riener, K. R., & Warren, R. M. (1992). Increasing the intelligibility of speech through multiple phonemic restorations. *Perception & Psychophysics*, *51*(3), 211–217. https://doi.org/10.3758/BF03212247

Başkent, D., Clarke, J., Pals, C., Benard, M. R., Bhargava, P., Saija, J., Sarampalis, A., Wagner, A., & Gaudrain, E. (2016). Cognitive Compensation of Speech Perception With Hearing Impairment, Cochlear Implants, and Aging: How and to What Degree Can It Be Achieved? *Trends in Hearing*, *20*. https://doi.org/10.1177/2331216516670279

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). Fitting Linear Mixed-Effects Models using lme4. *67*(1). https://doi.org/10.18637/jss.v067.i01

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Bench, J., Kowal, Å., & Bamford, J. (1979). The Bkb (Bamford-Kowal-Bench) Sentence Lists for Partially-Hearing Children. *British Journal of Audiology*, *13*(3), 108–112. https://doi.org/10.3109/03005367909078884

Bergman, A. S. (1990). *Auditory scene analysis : the perceptual organization of sound.* Cambridge, Massachusetts : The MIT Press
Includes bibliographical references (pages 737-761) and index. Includes bibliographical references and index.

Bergman, Blumenfeld, Cascardo, Dash, Levitt, & Margulies. (1976). Age-Related Decrement in Hearing for Speech. *Journal of Gerontology*, *31*(5), 533–538.

Bergman, M. (1980). *Aging and the perception of speech*. University Park Press.

Best, V., Mason, C. R., & Kidd, G. (2011). Spatial release from masking in normally hearing and hearing-impaired listeners as a function of the temporal overlap of competing talkers. *The Journal of the Acoustical Society of America*, *129*(3), 1616–1625. https://doi.org/10.1121/1.3533733

Binns, C., & Culling, J. F. (2007). The role of fundamental frequency contours in the perception of speech against interfering speech. *The Journal of the Acoustical Society of America*, *122*(3), 1765–1776. https://doi.org/10.1121/1.2751394

Bishop, D. V. M. (2003). *The Children's Communication Checklist, Version 2 (CCC-2)* (tech. rep.). The Psycho- logical Corporation. London, United Kingdom.

Boersma, P. (2001). Praat, a system for doing phonetics by computer. *Glot International*, *5*(9/10), 341–345.

Bolia, R. S., Nelson, W. T., Ericson, M. A., & Simpson, B. D. (2000). A speech corpus for multitalker communications research. *The Journal of the Acoustical Society of America*, *107*(2), 1065–1066. https://doi.org/10.1121/1.428288

Brokx, J. P. L., & Nooteboom, S. G. (1982). Intonation and the perceptual separation of simultaneons voices. *Journal of Phonetics*, *10*, 23–36.

Bronkhorst, A. W. (2015). The cocktail-party problem revisited: early processing and selection of multi-talker speech. *Attention, Perception, and Psychophysics*, *77*, 1465–1487. https://doi.org/10.3758/s13414-015-0882-9

Brouwer, S., Van Engen, K. J., Calandruccio, L., & Bradlow, A. R. (2012). Linguistic contributions to speech-on-speech masking for native and non-native listeners: Language familiarity and semantic content. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.3675943

Brungart, D., Iyer, N., Thompson, E. R., Simpson, B. D., Gordon-Salant, S., Schurman, J., Vogel, C., & Grant, K. (2013). Interactions between listening effort and masker type on the energetic and informational masking of speech stimuli. *Proceedings of Meetings on Acoustics*, *19*(1), 60146. https://doi.org/10.1121/1.4800033
doi: 10.1121/1.4800033

Brungart, D. S., & Iyer, N. (2012). Better-ear glimpsing efficiency with symmetrically-placed interfering talkers. *The Journal of the Acoustical Society of America*, *132*(4), 2545–2556. https://doi.org/10.1121/1.4747005

Brungart, D. S., & Simpson, B. D. (2002). Within-ear and across-ear interference in a cocktail-party listening task. *The Journal of the Acoustical Society of America*, *112*(6), 2985–2995. https://doi.org/10.1121/1.1512703

Brungart, D. S., Simpson, B. D., Ericson, M. A., & Scott, K. R. (2001). Informational and energetic masking effects in the perception of multiple simultaneous talkers. *The Journal of the Acoustical Society of America*, *110*(5), 2527–2538. https://doi.org/10.1121/1.1408946

Buss, E., Whittle, L. N., Grose, J. H., & Hall, J. W. (2009). Masking release for words in amplitude-modulated noise as a function of modulation rate and task. *The Journal of the Acoustical Society of America*, *126*(1), 269–280. https://doi.org/10.1121/1.3129506

Calandruccio, L., Bradlow, A. R., & Dhar, S. (2014). Speech-on-speech masking with variable access to the linguistic content of the masker speech for native and nonnative English speakers. *Journal of the American Academy of Audiology.* https://doi.org/10.3766/jaaa.25.4.7

Calandruccio, L., Dhar, S., & Bradlow, A. R. (2010). Speech-on-speech masking with variable access to the linguistic content of the masker speech. *The Journal of the Acoustical Society of America.* https://doi.org/10.1121/1.3458857

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

Cameron, S., Glyde, H., & Dillon, H. (2011). Listening in Spatialized Noise—Sentences Test (LiSN-S): Normative and Retest Reliability Data for Adolescents and Adults up to 60 Years of Age. *Journal of the American Academy of Audiology*, *22*(10), 697–709. https://doi.org/10.3766/jaaa.22.10.7

Carlile, S., & Corkhill, C. (2015). Selective spatial attention modulates bottom-up informational masking of speech. *Scientific Reports.* https://doi.org/10.1038/srep08662

Chan, D., Fourcin, A., Gibbon, D., Grandstrom, B., Huckvale, M., Kokkinakis, G., Kvale, K., Lamel, L., Lindberg, B., Moreno, A., Mouropoulos, J., Senia, F., Trancoso, I., in'T Veld, C., & Zeiliger, J. (1995). EUROM - A spoken language resource for the EU. *European Conference on Speech Communication and Technology.*

Cherry, E. C. (1953). Some Experiments on the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *25*(5), 975–979. https://doi.org/10.1121/1.1907229

Cherry, E. C., & Taylor, W. K. (1954). Some Further Experiments upon the Recognition of Speech, with One and with Two Ears. *The Journal of the Acoustical Society of America*, *26*(4), 554–559. https://doi.org/10.1121/1.1907373

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

Cooke, M. (2006). A glimpsing model of speech perception in noise. *The Journal of the Acoustical Society of America*, *119*(3), 1562–1573. https://doi.org/10.1121/1.2166600

Darwin, C. J., Brungart, D. S., & Simpson, B. D. (2003). Effects of fundamental frequency and vocal-tract length changes on attention to one of two simultaneous talkers. *The Journal of the Acoustical Society of America*, *114*(5), 2913–2922. https://doi.org/10.1121/1.1616924

Drullman, R., & Bronkhorst, A. W. (2000). Multichannel speech intelligibility and talker recognition using monaural, binaural, and three-dimensional auditory presentation. *The Journal of the Acoustical Society of America*, *107*(4), 2224–2235. https://doi.org/10.1121/1.428503

Durlach, N. I., Mason, C. R., Shinn-Cunningham, B. G., Arbogast, T. L., Colburn, H. S., & Kidd, G. (2003). Informational masking: Counteracting the effects of stimulus uncertainty by decreasing target-masker similarity. *The Journal of the Acoustical Society of America*, *114*(1), 368–379. https://doi.org/10.1121/1.1577562

Festen, J. M., & Plomp, R. (1990). Effects of fluctuating noise and interfering speech on the speech-reception threshold for impaired and normal hearing. *The Journal of the Acoustical Society of America*, *88*(4), 1725–1736. https://doi.org/10.1121/1.400247

Feys, J. (2015). *Npintfactrep: Nonparametric interaction tests for factorial designs with repeated measures* [R package version 1.5]. https://CRAN.R-project.org/package=npIntFactRep

Feys, J. (2016). Nonparametric tests for the interaction in two-way factorial designs using R. *R Journal*. https://doi.org/10.32614/rj-2016-027

Field, A., Miles, J., & Field, Z. (2012). Discovering Statistics Using R - 17 Exploratory factor analysis. *Discovering statistics using r*.

Fox, J., & Weisberg, S. (2011). *An R companion to applied regression* (Second). Sage. http://socserv.socsci.mcmaster.ca/jfox/Books/Companion

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (Third). Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2001). Spatial release from informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *109*(5), 2112–2122. https://doi.org/10.1121/1.1354984

Freyman, R. L., Balakrishnan, U., & Helfer, K. S. (2004). Effect of number of masking talkers and auditory priming on informational masking in speech recognition. *The Journal of the Acoustical Society of America*, *115*(5), 2246–2256. https://doi.org/10.1121/1.1689343

Freyman, R. L., Helfer, K. S., McCall, D. D., & Clifton, R. K. (1999). The role of perceived spatial separation in the unmasking of speech. *The Journal of the Acoustical Society of America*, *106*(6), 3578–3588. https://doi.org/10.1121/1.428211

Gamer, M., Lemon, J., & <puspendra.pusp22@gmail.com>, I. F. P. S. (2019). *Irr: Various coefficients of interrater reliability and agreement* [R package version 0.84.1]. https://CRAN.R-project.org/package=irr

Goodman, A. S. (n.d.). Auditory research lab audio software (arlas). version 0.20.2, data 2017-04-11. [Accessed: 02-01-2021]. https://github.com/myKungFu/ARLas

Green, T., & Rosen, S. (2013). Phase effects on the masking of speech by harmonic complexes: Variations with level. *The Journal of the Acoustical Society of America*, *134*(4), 2876–2883. https://doi.org/10.1121/1.4820899

Grose, J. H., Porter, H. L., & Buss, E. (2016). Aging and Spectro-Temporal Integration of Speech. *Trends in Hearing*, *20*, 1–11. https://doi.org/10.1177/2331216516670388

Harrell Jr, F. E. (2020). *Hmisc: Harrell miscellaneous* [R package version 4.4-2]. https://CRAN.R-project.org/package=Hmisc

Hirsh, I. J. (1950). The Relation between Localization and Intelligibility. *The Journal of the Acoustical Society of America*, *22*(2), 196–200. https://doi.org/10.1121/1.1906588

Hoffman, I., & Levitt, H. (1978). A note on simultaneous and interleaved masking. https://doi.org/10.1016/0021-9924(78)90013-8

Hopkins, K., & Moore, B. C. J. (2010). The importance of temporal fine structure information in speech at different spectral regions for normal-hearing and hearing-impaired subjects. *The Journal of the Acoustical Society of America*, *127*(3), 1595–1608. https://doi.org/10.1121/1.3293003

Hothorn, T., Hornik, K., van de Wiel, M. A., & Zeileis, A. (2006). A Lego system for conditional inference. *The American Statistician*, *60*(3), 257–263. https://doi.org/10.1198/000313006X118430

Howard-Jones, P., & Rosen, S. (1993). The perception of speech in fluctuating noise. *Acta Acustica united with Acustica*, *78*(5), 258–272.

Huang, H. W. (2018). *The Effects of Different Types of Contralateral Distractors on Switching Attention for Speech in Elder & Younger Adults with Normal Hearing* (Master's thesis). University College London, UCL.

Huggins, A. W. F. (1964). Distortion of the Temporal Pattern of Speech: Interruption and Alternation. *The Journal of the Acoustical Society of America*, *36*(6), 1055–1064. https://doi.org/10.1121/1.1919151

Humes, L. E., & Dubno, J. R. (2010). Factors affecting speech understanding in older adults. In S. Gordon-Salant, R. Frisina, R. Fay, & A. Popper (Eds.), *The aging auditory system*. Springer-Verlag New York.

Humes, L. E., Kidd, G. R., & Lentz, J. J. (2013). Auditory and cognitive factors underlying individual differences in aided speech-understanding among older adults. *Frontiers in Systems Neuroscience*, *7*(October), 1–16. https://doi.org/10.3389/fnsys.2013.00055

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An introduction to statistical learning: With applications in r* (Vol. 103). Springer.

Kassambara, A. (2021). *Rstatix: Pipe-friendly framework for basic statistical tests* [R package version 0.6.0.999]. https://rpkgs.datanovia.com/rstatix/

Kidd, G. R., & Humes, L. E. (2012). Effects of age and hearing loss on the recognition of interrupted words in isolation and in sentences. *The Journal of the Acoustical Society of America*, *131*(2), 1434–1448. https://doi.org/10.1121/1.3675975

Kidd, G., Mason, C. R., & Arbogast, T. L. (2002). Similarity, uncertainty, and masking in the identification of nonspeech auditory patterns. *The Journal of the Acoustical Society of America*, *111*(3), 1367–1376. https://doi.org/10.1121/1.1448342

Kollmeier, B., Warzybok, A., Hochmuth, S., Zokoll, M. A., Uslar, V., Brand, T., & Wagener, K. C. (2015). The multilingual matrix test: Principles, applications, and comparison across languages: A review. https://doi.org/10.3109/14992027.2015.1020971

Koo, T. K., & Li, M. Y. (2016). A Guideline of Selecting and Reporting Intraclass Correlation Coefficients for Reliability Research (2016/03/31). *Journal of chiropractic medicine*, *15*(2), 155–163. https://doi.org/10.1016/j.jcm.2016.02.012

Krishnan, S., Leech, R., Aydelott, J., & Dick, F. (2013). School-age children's environmental object identification in natural auditory scenes: Effects of masking and contextual congruence. *Hearing Research*, *300*, 46–55. https://doi.org/10.1016/j.heares.2013.03.003

Laures, J. S., & Weismer, G. (1999). The effects of a flattened fundamental frequency on intelligibility at the sentence level. *Journal of speech, language, and hearing research : JSLHR*, *42*(5), 1148–1156. https://doi.org/10.1044/jslhr.4205.1148

Lê, S., Josse, J., & Husson, F. (2008). FactoMineR: A package for multivariate analysis. *Journal of Statistical Software*, *25*(1), 1–18. https://doi.org/10.18637/jss.v025.i01

Leclère, T., Lavandier, M., & Deroche, M. L. (2017). The intelligibility of speech in a harmonic masker varying in fundamental frequency contour, broadband temporal envelope, and spatial location. *Hearing Research*, *350*, 1–10. https://doi.org/10.1016/j.heares.2017.03.012

Lee, J., Dhar, S., Abel, R., Banakis, R., Grolley, E., Lee, J., Zecker, S., & Siegel, J. (2012). Behavioral Hearing Thresholds between 0.125 and 20 kHz Using Depth-Compensated Ear Simulator Calibration. *Ear and Hearing*. https://doi.org/10.1097/AUD.0b013e31823d7917

Leech, R., Gygi, B., Aydelott, J., & Dick, F. (2009). Informational factors in identifying environmental sounds in natural auditory scenes. *The Journal of the Acoustical Society of America*, *126*(6), 3147–3155. https://doi.org/10.1121/1.3238160

Leensen, M. C., & Dreschler, W. A. (2013). The applicability of a speech-in-noise screening test in occupational hearing conservation. *International Journal of Audiology*. https://doi.org/10.3109/14992027.2013.790565

Lenth, R. V. (2016). Least-Squares Means: The {R} Package {lsmeans}. *Journal of Statistical Software*, *69*, 1–33. https://doi.org/10.18637/jss.v069.i01

Lenth, R. V. (2020). *Emmeans: Estimated marginal means, aka least-squares means* [R package version 1.5.3]. https://CRAN.R-project.org/package=emmeans

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

MacLeod, A., & Summerfield, Q. (1990). A procedure for measuring auditory and audiovisual speech-reception thresholds for sentences in noise: Rationale, evaluation, and recommendations for use. *British Journal of Audiology*, *24*(1), 29–43. https://doi.org/10.3109/03005369009077840

Mair, K. R. (2013). *Speech Perception in Autism Spectrum Disorder: Susceptibility to Masking and Interference* (PhD dissertation March). University College London, UCL.

Mair, P., & Wilcox, R. (2020). Robust Statistical Methods in R Using the WRS2 Package. *Behavior Research Methods*, *52*, 464–488.

Matheson, G. J. (2019). We need to talk about reliability: Making better use of test-retest studies for study design and interpretation. *PeerJ*. https://doi.org/10.7717/peerj.6918

McDonald, J. (2014). Multiple comparisons. *Handbook of biological statistics* (3rd ed., pp. 254–260). Sparky House Publishing.

Miller, G. A., & Licklider, J. C. R. (1950). The Intelligibility of Interrupted Speech. *The Journal of the Acoustical Society of America*, *22*(2), 167–173. https://doi.org/10.1121/1.1906584

Miller, S. E., Schlauch, R. S., & Watson, P. J. (2010). The effects of fundamental frequency contour manipulations on speech intelligibility in background noise. *The Journal of the Acoustical Society of America*, *128*(1), 435–443. https://doi.org/10.1121/1.3397384

Moore, B. (2008). The role of temporal fine structure in normal and impaired hearing. *Auditory Signal Processing in Hearing-Impaired Listeners. 1st International Symposium on Auditory and Audiological Research (ISAAR 2007)*, (Isaar), 247–262.

Moore, B. C. J. (2012). *An introduction to the psychology of hearing* (6th ed.). Bingley : Emerald
Includes bibliographical references and index.

Moore, B. C. (2003). Temporal integration and context effects in hearing. *Journal of Phonetics*, *31*(3-4), 563–574. https://doi.org/10.1016/S0095-4470(03)00011-1

Moore, D. R., Ferguson, M. A., Edmondson-Jones, A. M., Ratib, S., & Riley, A. (2010). Nature of Auditory Processing Disorder in Children. *PEDIATRICS*, *126*(2), e382–e390.

Moray, N. (1959). Attention in Dichotic Listening: Affective Cues and the Influence of Instructions. *Quarterly Journal of Experimental Psychology*, *11*(1), 56–60.

Murphy, C. F., Hashim, E., Dillon, H., & Bamiou, D. E. (2019). British children's performance on the listening in spatialised noise-sentences test (LISN-S). *International Journal of Audiology*. https://doi.org/10.1080/14992027.2019.1627592

Nasreddine, Z., Phillips, N., Bedirian, V., Charbonneau, S., Whitehead, V., Collin, I., Cummings, J., & Chertkow, H. (2005). The Montreal Cognitive Assessment , MoCA : A Brief Screening. *Journal of the American Geriatric Society*, *53*, 695–699. https://doi.org/10.1111/j.1532-5415.2005.53221.x

Nelson, P. B., & Jin, S.-H. (2004). Factors affecting speech understanding in gated interference: Cochlear implant users and normal-hearing listeners. *The Journal of the Acoustical Society of America*, *115*(5), 2286–2294. https://doi.org/10.1121/1.1703538

Noguchi, K., Gel, Y. R., Brunner, E., & Konietschke, F. (2012). nparLD: An R software package for the nonparametric analysis of longitudinal data in factorial experiments. *Journal of Statistical Software*, *50*(12), 1–23. http://www.jstatsoft.org/v50/i12/

Norbury, C. F. (2014). Practitioner Review: Social (pragmatic) communication disorder conceptualization, evidence and clinical implications. *Journal of Child Psychology and Psychiatry and Allied Disciplines*. https://doi.org/10.1111/jcpp.12154

Norbury, C. F., & Bishop, D. V. M. (2005). Children ' s Communication Checklist - 2 : a validation study. *Publie dans Revue Tranel*, *42*, 53–63.

Pichora-Fuller, M. K., & Singh, G. (2006). Effects of Age on Auditory and Cognitive Processing: Implications for Hearing Aid Fitting and Audiologic Rehabilitation. *Trends in Amplification*, *10*(1), 29–59. https://doi.org/10.1177/108471380601000103

Pichora-Fuller, M. K., & Souza, P. E. (2003). Effects of aging on auditory processing of speech. *International Journal of Audiology*, *42*(sup2), 11–16. https://doi.org/10.3109/14992020309074638

Qin, S., Nelson, L., McLeod, L., Eremenco, S., & Coons, S. J. (2019). Assessing test–retest reliability of patient-reported outcome measures using intraclass correlation coefficients: recommendations for selecting and documenting the analytical formula. *Quality of Life Research*. https://doi.org/10.1007/s11136-018-2076-0

R Core Team. (2018). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020a). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

R Core Team. (2020b). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. https://www.R-project.org/

Ramus, F., Rosen, S., Dakin, S. C., Day, B. L., Castellote, J. M., White, S., & Frith, U. (2003). Theories of developmental dyslexia: Insights from a multiple case study of dyslexic adults. *Brain*, *126*(4), 841–865. https://doi.org/10.1093/brain/awg076

Revelle, W. (2020). *Psych: Procedures for psychological, psychometric, and personality research* [R package version 2.0.12]. Northwestern University. Evanston, Illinois. https://CRAN.R-project.org/package=psych

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2005). Release from informational masking by time reversal of native and non-native interfering speech. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.2000751

Rhebergen, K. S., Versfeld, N. J., & Dreschler, W. A. (2006). Extended speech intelligibility index for the prediction of the speech reception threshold in fluctuating noise. *The Journal of the Acoustical Society of America*, *120*(6), 3988–3997. https://doi.org/10.1121/1.2358008

Richmond, S. A., Kopun, J. G., Neely, S. T., Tan, H., & Gorga, M. P. (2011). Distribution of standing-wave errors in real-ear sound-level measurements. *The Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.3569726

Rosen, S., Souza, P., Ekelund, C., & Majeed, A. A. (2013). Listening to speech in a background of other talkers: Effects of talker number and noise vocoding. *The Journal of the Acoustical Society of America*, *133*(4), 2431–2443. https://doi.org/10.1121/1.4794379

RStudio Team. (2019). *Rstudio: Integrated development environment for r*. RStudio, Inc. Boston, MA. http://www.rstudio.com/

Saija, J. D., Akyürek, E. G., Andringa, T. C., & Başkent, D. (2014). Perceptual restoration of degraded speech is preserved with advancing age. *JARO - Journal of the Association for Research in Otolaryngology*, *15*(1), 139–148. https://doi.org/10.1007/s10162-013-0422-z

Scheffers, M. T. M. (1983). *Sifting vowels. Auditory pitch analysis and sound segregation* (PhD dissertation). University of Groningen.

Schubert, E. D., & Parker, C. D. (1955). Addition to Cherry's findings on switching speech between the two ears. *Journal of the Acoustical Society of America*, *27*, 792–794. https://doi.org/10.1121/1.1908042

Shafiro, V., Sheft, S., & Risley, R. (2011). Perception of interrupted speech: Effects of dual-rate gating on the intelligibility of words and sentences. *The Journal of the Acoustical Society of America*, *130*(4), 2076–2087. https://doi.org/10.1121/1.3631629

Shafiro, V., Sheft, S., Risley, R., & Gygi, B. (2015). Effects of age and hearing loss on the intelligibility of interrupted speech. *The Journal of the Acoustical Society of America*, *137*(2), 745–756. https://doi.org/10.1121/1.4906275

Shen, J., & Souza, P. E. (2017). The Effect of Dynamic Pitch on Speech Recognition in Temporally Modulated Noise. *Journal of Speech Language and Hearing Research*, *60*(September), 2725–2739. https://doi.org/10.1044/2017_JSLHR-H-16-0389

Shinn-Cunningham, B. G. (2008). Object-based auditory and visual attention. *Trends in Cognitive Sciences*, *12*(5), 182–186. https://doi.org/10.1016/j.tics.2008.02.003

Siegel, J. H. (1994). Ear-canal standing waves and high-frequency sound calibration using otoacoustic emission probes. *Journal of the Acoustical Society of America*. https://doi.org/10.1121/1.409829

Steinmetzger, K., & Rosen, S. (2015). The role of periodicity in perceiving speech in quiet and in background noise. *The Journal of the Acoustical Society of America*, *138*(6), 3586–3599. https://doi.org/10.1121/1.4936945

Stone, M. A., Füllgrabe, C., & Moore, B. C. J. (2012). Notionally steady background noise acts primarily as a modulation masker of speech. *The Journal of the Acoustical Society of America*, *132*(1), 317–326. https://doi.org/10.1121/1.4725766

Stone, M. A., & Moore, B. C. J. (2014). On the near non-existence of “pure” energetic masking release for speech. *The Journal of the Acoustical Society of America.* https://doi.org/10.1121/1.4868392

Stuart, A. (2008). Reception Thresholds for Sentences in Quiet, Continuous Noise, and Interrupted Noise in School-Age Children. *Journal of the American Academy of Audiology*, *19*(2), 135–146. https://doi.org/10.3766/jaaa.19.2.4

Summers, R. J., & Roberts, B. (2020). Informational masking of speech by acoustically similar intelligible and unintelligible interferers. *The Journal of the Acoustical Society of America*, *147*(2), 1113–1125. https://doi.org/10.1121/10.0000688

Surprenant, A. M., & Watson, C. S. (2001). Individual differences in the processing of speech and nonspeech sounds by normal-hearing listeners. *The Journal of the Acoustical Society of America*, *110*(4), 2085–2095. https://doi.org/10.1121/1.1404973

Torchiano, M. (2020). *Effsize: Efficient effect size computation* [R package version 0.8.1]. https://doi.org/10.5281/zenodo.1480624

Van Engen, K. J., & Bradlow, A. R. (2007). Sentence recognition in native- and foreign-language multi-talker background noise. *The Journal of the Acoustical Society of America.* https://doi.org/10.1121/1.2400666

van Esch, T. E., Kollmeier, B., Vormann, M., Lyzenga, J., Houtgast, T., Hällgren, M., Larsby, B., Athalye, S. P., Lutman, M. E., & Dreschler, W. A. (2013). Evaluation of the preliminary auditory profile test battery in an international multi-centre study. *International Journal of Audiology*, *52*(5), 305–321. https://doi.org/10.3109/14992027.2012.759665

von Goethe, J. W. (1829). *Wilhelm Meisters Wanderjahre oder die Entsagenden.* Cotta.

Warren, R. M. (1970). Perceptual Restoration of Missing Speech Sounds. *Science*, *167*(3917), 392 LP –393. http://science.sciencemag.org/content/167/3917/392.abstract

Watson, C. S. (1987). Uncertainty, informational masking and the capacity of immediate auditory memory. In W. A. Yost & C. S. Watson (Eds.), *Auditory processing of complex sounds* (pp. 267–277). Hillsdale, N.J. : L. Erlbaum Associates Includes bibliographies and indexes.

Wierstorf, H., Geier, M., Raake, A., & Spors, S. (2011). A Free Database of Head-Related Impulse Response Measurements in the Horizontal Plane with Multiple Distances. *AES130.*

Wiig, E., H, Semel, E., & Secord, W. (2017). *Clinical Evaluation of Language Fundamentals - Fifth Edition UK (CELF-5UK)* (tech. rep.). PsychCorp, Pearson Clinical Assessment.

World Health Organisation. (1998). *Occupational exposure to noise: evaluation, prevention and control* (tech. rep.).

Xu, Y. (2013). ProsodyPro — A Tool for Large-scale Systematic Prosody Analysis. *In Proceedings of Tools and Resources for the Analysis of Speech Prosody (TRASP 2013)*, 7–10.