Research in Autism Spectrum Disorders 62 (2019) 85–98

Contents lists available at ScienceDirect

Research in Autism Spectrum Disorders

journal homepage: www.elsevier.com/locate/rasd

Development of a gaze contingent method for auditory threshold evaluation in non-verbal ASD children

Brian Sullivan[a,c,1,*], C. Ellie Wilson[b,1], David Saldaña[b]

[a] *School of Psychological Science, University of Bristol, UK*
[b] *Individual Differences, Language and Cognition Lab, Department of Developmental and Educational Psychology, University of Seville, Spain*
[c] *Tobii AB, Danderyd, Sweden*

ARTICLE INFO

Number of reviews completed is 2

Keywords:
Children
Audition
Non-verbal
Eye tracking
Gaze contingent
Perceptual threshold

ABSTRACT

Background: Minimally verbal children with autistic spectrum disorder (ASD) make up an estimated third of the ASD population (Downs et al., 2005), but have been understudied due to difficulties in running experiments with such participants. We sought to develop an instrument to evaluate auditory perception, with the goal of testing both typically developing (TD) and ASD children, including minimally verbal ASD. Audio difference thresholds are typically measured by an audiologist using visual reinforcement audiometry (VRA) techniques, but this requires a trained clinician. Alternatively, mismatch negativity (MMN) via an electroencephalogram can provide an objective threshold measure and the participant can passively attend to stimuli. However, EEG equipment is expensive, and the procedure can be uncomfortable and difficult with anxious or touch sensitive participants.

Method: We developed a testing software for estimating auditory thresholds in children using a gaze contingent 'game'. Our open source software uses an eye tracker, Matlab and child-oriented stimuli to automate aspects of VRA. Initial results suggest that audio thresholds can be obtained using our affordable non-invasive system, operated with minimal training, but refinement is necessary.

Results: Our method can obtain thresholds for most typical children, but data collection in young ASD children proved more challenging, yielding poor results, and will require further development to make the game more accessible. While promising, these results need to be corroborated with an alternate measure of difference threshold.

Conclusion: We document our efforts to design an effective interactive game to assess auditory perception using gaze-contingent eye-tracking methods; and provide case level insights on the testing individual participants and the heterogeneous ability and performance levels within ASD. We discuss the challenges experienced in testing and eye tracking both typical and ASD children to inform clinical and research groups to advance this promising line of research.

[*] Corresponding author at: School of Psychological Science, 12A Priory Road, Bristol BS8 1TU, UK.
E-mail address: brian.sullivan@bristol.ac.uk (B. Sullivan).

[1] Authors contributed equally to the research and manuscript.

https://doi.org/10.1016/j.rasd.2019.02.006
Received 20 June 2018; Received in revised form 20 February 2019; Accepted 25 February 2019
Available online 21 March 2019

1. Introduction

1.1. Background and motivation

Linguistic ability within autism spectrum disorder (ASD) can range from fluent speech with average (or above) verbal intelligence, to ‘minimally verbal’, defined here as an individual with fewer than five productive words by the time they reach school age (Tager-Flusberg & Kasari, 2013). It is estimated that 30% of ASD children remain minimally verbal (Tager-Flusberg & Kasari) and it is this severely impaired end of the spectrum that is arguably most in need of attention. Yet the majority of research in ASD focuses on children with age-appropriate or moderately-impaired language, and it is questionable whether insights gained from these studies may be extrapolated to those who are minimally verbal. A primary reason for the lack of research into the ‘neglected end of the spectrum’ is that there are substantial challenges involved in completing tests with these individuals. As a result, a priority for ASD research today is to develop evaluation techniques using available technology that can effectively assess cognitive and behavioral traits of individuals across the autistic spectrum.

Assessing receptive and expressive language is important but targeting fundamental traits that might underlie linguistic difficulties is also critical. Atypical sensory behaviors – including hyper-sensitivities, hypo-sensitivities, or unusual sensory seeking behaviors – are a common characteristic of ASD with prevalence estimates between 75% (Klintwall et al., 2011) and 90% (Crane, Goddard, & Pring, 2009). Perceptual experiences are acquired early and form the basis of our interaction with the environment, therefore atypical perception of auditory information may have knock-on effects for the development of more advanced functional domains, such as speech and language (Stevenson et al., 2014; Watson, Roberts, Baranek, Mandulak, & Dalton, 2012). Rhythm perception, for example, has been linked to skill level in speech comprehension (Bertoncini, Nazzi, Cabrera, & Lorenzi, 2011; Drullman, Festen, & Plomp, 1994; Elliott & Theunissen, 2009; Slater et al., 2015), grammar (e.g. Gordon et al., 2015) and reading (e.g. Huss, Verney, Fosker, Mead, & Goswami, 2011; Carr, White-Schwoch, Tierney, Strait, & Kraus, 2014). Rhythm perception has also been a target for therapeutic intervention in speech and language disorders such as ASD, stuttering, aphasia and Parkinson's disease (Fujii & Wan, 2014). Pitch perception, on the other hand, has repeatedly been reported as atypical (often enhanced) in at least a subgroup of ASD individuals (Bonnel et al., 2003, 2010; Jones et al., 2009; Samson, Mottron, Jemel, Belin, & Ciocca, 2006; Stanutz, Wapnick, & Burack, 2014), and has been associated with language learning in typically developing infants (Mueller, Friederici, & Männel, 2012). Furthermore, within the study of ASD auditory perception, a distinction between the perception of speech and non-speech stimuli has been highlighted as it is hypothesized that ASD children with severe impairments in social communication may have a specific impairment processing complex speech sounds. Studies demonstrating atypical responses to speech sounds in particular have been documented using EEG (Kuhl, Coffey-Corina, Padden, & Dawson, 2005; Yau, McArthur, Badcock, & Brock, 2015; Schwartz, Shinn-Cunningham, & Tager-Flusberg, 2018) and behavioral (Heaton, Williams, Cummins, & Happé, 2008) methods.

However, of the studies mentioned here, only one (Yau et al., 2015) has assessed auditory perception in minimally verbal subjects. Moreover, studies invariably assess features of auditory perception – e.g. pitch and rhythm – in isolation. Devising tasks that can examine multiple components of auditory perception in a comparable manner is important if we are to understand which features may be most relevant to language development. For these reasons, we aimed to develop an objective measure that could examine different features of auditory perception, and that could be used for all children including those that are minimally verbal.

1.2. Audiometry and hearing thresholds

Within psychophysics, perception of sensory signals can be quantified by absolute detection thresholds, i.e. what is the softest sound one can hear before only perceiving silence, and auditory difference thresholds, i.e. what is the smallest difference one can perceive between two tones before they are perceived as identical. In the current study, we focus solely on auditory difference thresholds.

In hearing evaluation, there are several common ways to evaluate auditory difference thresholds. Standard hearing tests require the participant to communicate with an experimenter or audiologist, with the patient verbally instructed to signal when they perceive a tone or a difference between tones. In nonverbal or minimally verbal populations, such as young children, there are two primary methodologies to evaluate thresholds without requiring the child to understand instructions or actively communicate with the researcher; one uses overt behavioral measures and the other electro-encephalography (EEG) signal recording.

1.3. Behavioral methods

Audiologists have developed a set of non-invasive observational techniques to present stimuli and evaluate orientation and physical behavior in response to the stimulus (Dievendorf & Gravel, 1996; Sabo, 1999).

In Behavioral Observation Audiometry (BOA) sounds are presented and the audiologist must observe the infants (new born to ∼5 months) behavior (eye and head orientation and/or facial expressions) when exposed to sound stimuli.

Visual Reinforcement Audiometry (VRA) and Conditioned Orientation Reflex (COR) are techniques that both rely on the child (typically 5 months – 2 years) learning a stimulus–response pairing and responding with an orientation response (eye, head and/or body turn). Initially, the audiologist will present an anchor sound and reward the child when an orientation response is made (rewards may be, for example, shaking a toy or moving a puppet). In VRA an absolute threshold may be evaluated once the child has learned to orient to a sound. The audiologist then may gradually lower the volume of the sound over a series of presentations until the orientation response is no longer elicited indicating the child no longer perceives the sound. COR expands the same premise to have two or more sound sources, useful for auditory difference thresholds. For instance, the child may be reinforced to orient to the left for one sound and to the right for another. If the child cannot perceive a difference between the two stimuli their pattern of choice behavior should be 50/50.

Conditioned Play Audiometry is similar to the standard audiology test but adds extra incentives to keep children (2–3 years) engaged. For instance, instructing the child to use a toy in a certain way if they hear a particular sound, such holding a toy block and dropping it as soon as they hear a sound.

1.4. EEG methods

Alternatively, auditory difference thresholds can be determined via auditory evoked potentials (AEP), electrical activity from populations of synchronously firing neurons (Luck, 2005), recorded via EEG. This is an extremely useful method as not only can it be used to evaluate thresholds in nonverbal participants, it works in participants that are not actively attending and can be conducted in an automated fashion not requiring an audiologist to observe the participant behavior.

In such studies, the timing of EEG responses can be accurately recorded and analyzed according to distinct portions that match to different stages of processing (Musiek & Baran, 2007). Early brainstem responses (0-15 ms) can be used to determine absolute thresholds (Paulraj, Subramaniam, Yaccob, Adom, & Hema, 2015). If a participant repeatedly hears a stimulus and then is presented with a new dissimilar stimulus, a mismatch negativity (MMN) response arises late in the record (200-400 ms) and can be used to estimate auditory difference thresholds (Näätänen, Paavilainen, Rinne, & Alho, 2007). However, MMN is problematic for clinical assessment of individual participants and typically requires across subject averaging. Schall (2016) has argued that MMN can be clinically useful for measuring sound processing impairments but not a final diagnostic and will require substantial investment to target specific patient groups and gather normative data sets. Furthermore, whilst EEG techniques are non-invasive and don’t require overt behavior, the testing environment can be intense, and the participant is required to endure a set-up procedure involving putting a cap on their head with multiple electrodes that need to be secured. This can be extremely difficult in a subject who may be touch sensitive, anxious, hyperactive and/or noncompliant.

1.5. Testing hearing in the ASD population

The sensory atypicalities that are common in ASD also contribute to the difficulties in conducting tests in this population (Egelhoff, Whitelaw & Rabidoux, 2005; Brueggeman, 2012). In the case of hearing testing, sensory atypicalities could mean that an ASD participant dislikes wearing items like headphones, or they find the volume of sound presentation overwhelming. Rosenhall, Nordin, Sandström, Ahlsen, and Gillberg (1999) outlined several additional common characteristics and challenges in testing the ASD population including: anxiety, hyperactivity and poor attention, cognitive and language comprehension impairments, difficulty in new environments, increased false positive and false negative responses, and rapid or slowed habituation. Although any individual participant may only exhibit a portion of these characteristics they all contribute to difficulty in collecting reliable experimental data. Similarly, Plesa Skwerer, Jordan, Brukilacchio, and Tager-Flusberg (2016) measured receptive language ability in ASD children and found a large amount of variety arguing for individually tailored assessment and interventions. Downs, Schmidt, and Stephens (2005) did not recommend standard auditory exams (i.e. the patient explicitly signals when a tone is heard) with the ASD population and instead advised that patient history (communicated by the parent) and behavioral observation audiometry could be more successful. Kasari, Brady, Lord, and Tager-Flusberg (2013) recommend a comprehensive inventory of behavior in minimally verbal ASD children covering both verbal and nonverbal behavior. Individualized methods could certainly be useful for gaining insight into one child's ability and potential difficulty. However, this approach is limited in that conclusions regarding auditory difference thresholds and the potential relationship with additional cognitive and linguistic abilities in a wider population is not possible when different methods are used for every participant.

1.6. Motivation for current approach

While assessment of thresholds via behavioral observation or auditory evoked potentials are well established, they have some requirements that may be difficult to meet in a non-specialist lab. Without access to an audiology clinic or auditory testing facilities many of the behavioral testing options are not feasible. Similarly, without an EEG system and a trained technician to collect and interpret the data, it is not clear how to best assess audition. As a result, reports of EEG studies in minimally verbal ASD participants are extremely sparse (Yau et al., 2015). In our case, we decided to try to convert some aspects of visual reinforcement audiometry into a ‘game’ that uses eye movements as a means of interaction. Eye tracking hardware is becoming more affordable and easier to use, and there are several additional advantages that are particularly beneficial when testing a minimally verbal ASD population: the set-up time is minimal, physical elements like attaching electrodes are not required, testing can begin as soon as calibration is achieved, and participants are permitted to make some movements during testing without adversely affecting data. In children who may have impaired motor skills, the eyes can be used effectively as means to select onscreen imagery, and once calibrated an eye tracker can be used as an intuitive real-time interaction device.

Mixed success has been reported using gaze contingent experiments in infant participants. Several researchers have demonstrated that infant eye movements can be successfully recorded and infants can learn how to interact within a gaze contingent experiment (Miyazaki, Takahashi, Rolf, Okada, & Omori, 2014; Wass, Porayska-Pomsta, & Johnson, 2011; Wang et al., 2012). These studies trained participants to control their visual attention by reinforcing ‘correct’ gaze movements with visual and auditory rewards. They demonstrated that infants learn to anticipate rewards and moderate their eye-movements to control their environment. The only study that we are aware of that attempted to conduct an auditory test using a gaze contingent eye-tracking set-up to record ‘responses’ is reported only as proceedings from a conference (Schwarz, Nazem, Olsson, Marklund, & Uhlén, 2014). These authors attempted to train 12 infants to fixate on one location on the screen in response to the presence of a sound and used visual rewards to reinforce the looking behavior. Of 6 infants who were successfully calibrated, only 2 passed the training criterion, and neither of these completed the following test phase at above-chance level.

Like infants, ASD children with impaired receptive language may have limited understanding of verbal instructions and have trouble providing verbal or manual responses. However, these ASD children are of course not infants – they are children who probably have experience with interacting with electronic devices; who can potentially concentrate for longer (given the right task); who have greater control of their eye-movements, and who usually have some understanding of cause and effect regarding their impact on the environment. As such, we aimed to learn from reports with infant participants and adapt the methods to achieve even greater success in developing a gaze contingent paradigm for minimally verbal ASD children.

1.7. *Aims*

In the current study, (1) we develop a gaze contingent test for auditory perception that could be completed by typical children and those with severe impairments in language and intellectual impairment (IQ < 70). Our paradigm requires them to discriminate between two sounds. At the easiest level, the two sounds were the most extreme examples; at the hardest level, the two sounds were the most similar perceptually. This allows us to record a perceptual threshold. (2) With such a test we can compare auditory difference thresholds for stimuli including pitch and rhythm (using both human voice and synthetic sounds). (3) Lastly, by open sourcing our Matlab test code it can be used and improved by the research community.

2. Method

2.1. *Participants*

A group of 16 typically developing children (TD; age range: 4.4–9.9 years; mean age: 8.4 years) and 9 ASD children (age range: 5.5–9.9 years; mean age: 8.2 years) from the region near Seville, Spain were tested. ASD children were recruited through ‘Autismo Sevilla’, an organization for autistic individuals, and ‘Asociación SETA’ a center for attention and early intervention. Diagnostic reports were obtained via the parents from clinical psychologists or psychiatrists. If not previously conducted, the Autism Diagnostic Observation Schedule-Generic (ADOS-G, Lord et al., 2000) was completed by the research team. Minimally verbal participants completed module 1, the remainder completed module 2. The ADOS was not completed with one participant due to anxiety and noncompliance. Because he already had a clinical diagnosis, and observations by the research team who are experienced with autism strongly indicated ASD, it was decided to terminate the ADOS assessment. The Spanish version of the Autism Quotient (AQ; Auyeung, Baron-Cohen, Wheelwright, & Allison, 2007) was completed for all participants, although final scores for one ASD participant and 6 TD participants were not available due to missing items or questionnaire not being returned. If a participant scores over 76, an assessment of ASD is recommended. However, it is not a diagnostic tool and the four ASD participants we had who scored below this cut-off (see Table 1) were still analyzed as within the spectrum as they had scores within the autism range on the ADOS. Parents of all children reported that their child had no hearing difficulties, and completed the Spanish version of The Child Sensory Profile 2 (Dunn, 2014), providing corroborative information on participants’ sensory experiences.

To measure receptive language, the Spanish version of the Test for Reception of Grammar (TROG; Bishop, 2003; Spanish version: ‘Test de Comprensión de Estructuras Gramaticales’ (CEG), Mendoza, 2005; Mendoza, Carballo, Muñoz, & Fresneda, 2005) was attempted with all participants, although it was not possible to complete the test with one ASD participant. To measure nonverbal cognitive ability, the Kaufman Brief Intelligence Test matrices (KBIT; Kaufman & Kaufman, 2004) subtest was completed for all participants. This task was considered one of the most effective measures of nonverbal IQ in this population because participants are not required to give a verbal response.

Of the ASD participants, 3 were minimally verbal (they had fewer than five productive words to communicate and receptive language scores in the impaired range), 4 participants had five or more productive words to communicate and scored in the impaired range on the receptive language task, and 2 had fluent phrase speech and scored in the normal range on the receptive language task. Details of participant demographics and cognitive measures are provided in Table 1.

2.2. *Stimuli and set-up*

Our source code and stimuli are located at: https://github.com/VisionResearchBlog/gaze-audio-threshold.

Participants were seated ~60 cm from the eye tracking setup. For ease of calibration, we used a Tobii TX300 tracker that consists of an eye tracker camera base with a 23″ monitor on top (the screen subtends roughly 45° × 25° of visual angle at that viewing distance). Note, our setup is portable to other trackers and was also tested on a Tobii X2-60, which is smaller and cheaper, but runs at a lower frame rate (but still sufficient for our gaze contingent testing). Conversion to use a different brand of tracker is possible but would require sufficient programming skill as the current software only supports trackers using the Tobii Pro software development kit. Stimulus presentation and data recording were controlled by a PC running Matlab R2015a (Mathworks, Natick, MA, USA).

The testing environment was in a quiet lab space (without sound proofing) with overhead lighting. ‘Attention getting’ calibration

Table 1
Individual behavioral measures and auditory threshold results.

Group	Age	Participant code	Nonverbal IQ	Verbal IQ	Autistic traits		Child sensory profile 2 (quadrant scores)				Auditory threshold test highest level reached (*N* training trials/*N* main experiment trials)			
			KBIT norm (40–160, mean: 100)	CEG norm (1–100, mean: 50)	ADOS scores (communica-tion & social subscales)	Autism quotient (0–150)	Seeking	Avoiding	Sensitivity	Registration	Pitch human	Pitch synthetic	Rhythm human	Rhythm synthetic
ASD (impaired receptive language, verbal)	5.5	a01	127	1	Comm: 4 Soc: 5	64	29	30	23	24	1 (20/8)	*Training 2*	*Training 2*	*Training 2*
ASD (typical receptive language, fluent phrase speech)	6.7	a02	90	57	Comm: 6 Soc: 10	84	49	49	38	42	–	–	*Training 2*	–
ASD (minimally verbal)	7.8	a03	40	–	–	–	46	56	36	63	–	–	–	*Training 2*
ASD (impaired receptive language, verbal)	8.1	a04	66	1	Comm: 8 Soc: 10	57	26	26	26	23	8 (16/49)	3 (46/16)	9 (22/54)	2 (16/21)
ASD (impaired receptive language, verbal)	8.1	a05	40	1	Comm: 6 Soc: 9	65	36	33	44	30	2 (18/23)	–	–	2 (44/29)
ASD (impaired receptive language, verbal)	8.1	a06	85	1	Comm: 5 Soc: 9	42	24	22	20	26	–	–	*Training 2*	–
ASD (typical receptive language, fluent phrase speech)	9.3	a07	133	50	Comm: 2 Soc: 5	115	49	59	53	56	9 (14/55)	–	–	7 (16/55)
ASD (minimally verbal)	9.9	a08	58	1	Comm: 6 Soc: 13	100	56	40	40	36	*Training 1*	*Training 1*	*Training 2*	1 (74/18)
ASD (minimally verbal)	9.9	a09	63	1	Comm: 5 Soc: 11	86	46	36	29	31	*Training 2*	–	–	–
TD	6.7	c01	112	40		50	11	22	21	4	2 (26/27)	2 (14/26)	3 (14/34)	3 (14/36)
TD	6.8	c02	100	55		–	–	–	–	–	–	*Training 2*	3 (20/28)	–
TD	7.3	c03	124	95		–	25	37	28	31	2 (14/23)	1 (74/9)	*Training 2*	2 (38/24)
TD	8.6	c04	87	30		–	–	–	–	–	1 (76/23)	–	–	*Training 2*
TD	8.6	c05	98	65		57	32	13	17	20	*Training 2*	2 (20/24)	8 (18/56)	5 (78/44)
TD	8.8	c06	134	98		–	30	32	25	31	9 (20/40)	–	–	6 (16/45)
TD	9	c07	99	45		67	29	26	22	20	2 (20/19)	–	–	1 (54/12)
TD	9.2	c08	83	10		–	–	–	–	–	–	2	*Training 2*	–

(*continued on next page*)

Table 1 (*continued*)

Group	Age	Participant code	Nonverbal IQ	Verbal IQ	Autistic traits		Child sensory profile 2 (quadrant scores)				Auditory threshold test highest level reached (*N* training trials/*N* main experiment trials)			
			KBIT norm (40–160, mean: 100)	CEG norm (1–100, mean: 50)	ADOS scores (communica-tion & social subscales)	Autism quotient (0–150)	Seeking	Avoiding	Sensitivity	Registration	Pitch human	Pitch synthetic	Rhythm human	Rhythm synthetic
												(20/26)		
TD	9.3	c09	114	50		40	27	25	25	22	3 (48/28)	–	–	*Training 2*
TD	9.7	c10	99	65		56	30	37	30	24	–	5 (24/53)	*Training 2*	3 (14/30)
TD	9.7	c11	99	65		56	40	31	22	29	2 (46/26)	–	9 (16/60)	6 (22/56)
TD	9.9	c12	113	75		58	39	47	30	26	–	8 (14/63)	9 (26/55)	8 (16/60)
TD	9.9	c13	98	90		54	15	7	7	2	–	9 (14/58)	9 (20/40)	–
TD	10	c14	83	4		–	–	–	–	–	–	1 (110/20)	3 (16/30)	–
TD	4.8	c15[a]	101	20		46	27	28	24	24	–	3 (72/31)	9 (18/57)	–
TD	6.6	c16[a]	99	95		28	64	47	44	52	–	3 (27/39)	4 (51/39)	–

Participant diagnosis, age, IQ scores, autistic traits, sensory profile and behavioral results for our auditory threshold game are presented. Blank entries indicate the participant was not tested.

[a] Two typically developing participants completed an older version of the task in which training phases 1A & 1B had a pass criterion of 3 out of 3 correct trials. The testing phase and all parameters and stimuli were identical.

targets designed for children (included with Tobii Studio software) were used to calibrate the tracker. The animations subtended 1.2° × 1.2° and were placed in a randomized sequence in five locations onscreen (upper left, bottom left, center, top right, and bottom right) to gather calibration data. The brief 1.5 s animations of items included drawings of toys including a bus, caterpillar, dog, chick, cat, lobster, and rattle, all paired with brief semantically unrelated sounds. For example, the toy bus shrinks and grows while one hears the sound of an orchestral fanfare, and the chick shakes right and left with the sound of a two-note piano ostinato.

Auditory stimuli were in 4 sets of 10 fixed examples that varied within one perceptual dimension. The four sets included pitch (synthetic and human voice) and rhythm (synthetic and human voice). The pitch stimuli were single sounds of 2 s duration at a frequency of: 415 Hz ($G^{\#}$); 428 Hz; 440 Hz; 453 Hz; 466 Hz; 473 Hz; 480 Hz; 488 Hz; 492 Hz; 494 Hz (B). The rhythm stimuli were a series of beats lasting 6–7 s, at a fixed frequency of 330 Hz in the synthetic set and 196 Hz in the human set. The beat patterns gradually and systematically increased in complexity. At each progression, one bar out of three would change slightly. The first stimulus was 3 bars of 4 quarter-notes. The second stimulus was 2 bars of 4 quarter-notes, and 1 bar consisting of 1 quarter-note/2 eighth-notes/1 half-note. The tenth stimulus was 3 identical bars each consisting of 3 eighth-notes/2 sixteenth-notes/1 half-note. The rhythm combinations are visually represented as a musical score in the file set which is free to download via the link above ('Rhythm Beats.pdf'). The synthetic pitch stimuli were generated using Praat (www.fon.hum.uva.nl/praat/, Boersma, 2001; Boersma & Weenink, 2017) and Adobe Audition 7.0 (Adobe, San Jose, CA, USA, 2014). The synthetic rhythm stimuli were generated in Noteflight (www.noteflight.com, Noteflight, LLC, Boston, MA, USA, 2007), an online music notation software, and mimicked the sound of a piano. The human versions were of an adult female mimicking the synthetic versions and saying the syllable 'ba'. They were recorded in a soundproof lab edited afterwards using Praat so that the final versions were as close as possible to their synthetic counterparts whilst maintaining a human quality. All stimuli were presented at 60 dB measured using a sound level meter (Velleman, DVM1326). All sounds were played using a single speaker (Sony, SRS-GU10iP) positioned in the center directly below the monitor.

Visual stimuli were hand drawn characters – a dinosaur, alien and a snail (2.8° × 2.8° square). For each character, 12 different silent 'reward animations' of ~1 s were generated using Adobe Photoshop, in which the character completed a 'positive' and entertaining action, e.g. dancing, smiling, waving. Two of the animations were designed to be particularly rewarding and encouraging, e.g. silent animations of the character cheering with multicolored fireworks in the background.

All individual auditory and visual stimuli can be downloaded from the source code link provided above. Note, an additional set of sounds varying in volume were generated to test absolute thresholds, but were not tested as the experimental sessions were becoming too long (the duration of sessions varied form 20–45 min). However, these sounds are also available in the stimulus set provided in the link.

Participants completed two versions of the task (one pitch, one rhythm) in a session, and the order was counterbalanced across participants. The experimenter was always present in the room and she offered support and guidance during calibration, and to help the participant grasp the concept of the point-of-gaze fixation with their eye gaze (see below). Once the experiment started she was silent, only interrupting if the quality of the track was poor and recalibration was necessary. ASD participants completed the experiment with a parent also present. This was necessary to reduce anxiety and ensure the participant remained seated and attentive to the task. The parent was told to not point at any images on the screen, and to only help the participant maintain attention toward the screen. For control participants it was not necessary to have a parent present.

2.3. Trial protocol

Our experimental design went through many revisions (see Section 5). The final version consisted of an initial eye tracker setup followed by four phases that gradually introduced the participant to gaze-based interaction in the context of a 3 up – 1 down staircase. First, participants were seated in front of the eye tracking and monitor setup and calibrated. After calibration, the software would display the mean and standard deviation of the calibration data for each calibration point and overall. We sought to achieve error below 2–3 degrees. Throughout the experiment a small orange square (0.25° × 0.25°) was rendered in real-time at the current point-of-gaze; the participant was encouraged to recognize that this represented their point-of-gaze by making them follow a finger of the experimenter moving around the screen. All participants grasped the concept of manipulating the orange square with their eye-gaze. This also meant that the experimenter (who silently observed behind the participant) had direct feedback on the quality of the track and was able to pause the program and recalibrate if needed.

The initial stimulus presentation screen used a gray background with three zones. This consisted of a central fixation zone displaying one of the characters and two rectangles (7.1° × 5.7°), a blue one on the left and a red one on the right, both with a central white question mark (1.9° × 1.9°). Real-time area of interest detection used a generous 4° boundary around central fixation and the 2 rectangles.

Fig. 1 shows how the trials progressed, with details of the variations from training through to main experiment protocol described in the section below. The trials always started with the presentation of a character in the central fixation zone on which the participant was required to fixate for 300 ms. If the program detected the participant had not fixated the central zone after 2 s had elapsed, the character would move left and right (± 2°) as an attention getting stimulus. Once the participant fixated, the auditory stimulus played and the two rectangles appeared.

Throughout the experiment, the inter-trial interval – defined as the time between the end of the final visual presentation of one trial (see Fig. 1A, B) and the appearance of the central fixation image for the next trial – was a number randomly selected between 0.5 s and 3 s using Matlab's *rand()* function.

The crux of the game was that participants associated one stimulus with the left-hand side, and another stimulus with this right-hand side (this stimulus-response pairing was kept consistent per participant, but randomized across).

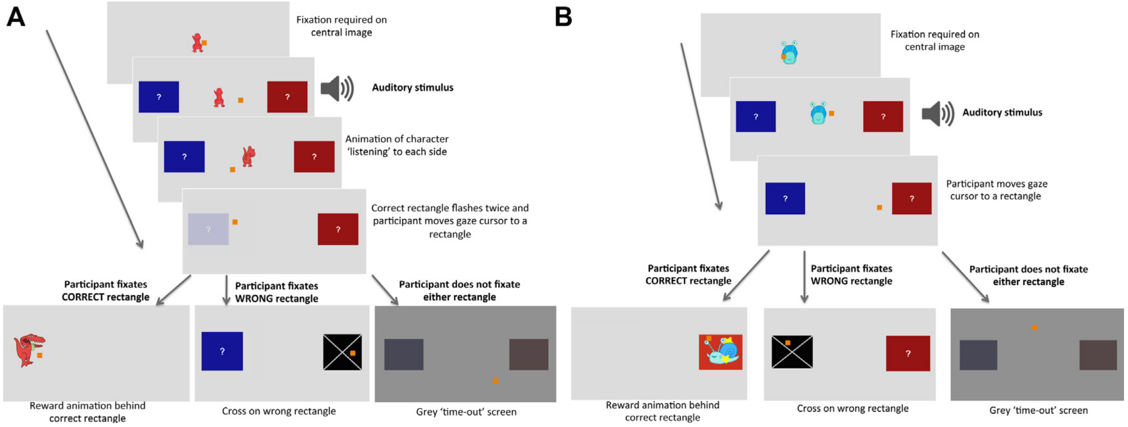
A
Fixation required on central image
Auditory stimulus
Animation of character 'listening' to each side
Correct rectangle flashes twice and participant moves gaze cursor to a rectangle
Participant fixates CORRECT rectangle
Participant fixates WRONG rectangle
Participant does not fixate either rectangle
Reward animation behind correct rectangle
Cross on wrong rectangle
Grey 'time-out' screen
B
Fixation required on central image
Auditory stimulus
Participant moves gaze cursor to a rectangle
Participant fixates CORRECT rectangle
Participant fixates WRONG rectangle
Participant does not fixate either rectangle
Reward animation behind correct rectangle
Cross on wrong rectangle
Grey 'time-out' screen

Fig. 1. (A) Starting from the top, showing the progression in time of the onscreen sequence for Phase 1A & 1B training trials. (B) The time progression for Phase 2 training trials and main experimental testing trials. Orange square represents the gaze cursor that participants learnt to control by moving their eye-gaze.

2.4. Training phases and main experiment protocol

The auditory threshold experiment consisted of two training phases that the participant had to pass in order to move onto the full staircase threshold test.

Training phase 1A: This phase taught the pairing of stimulus and location, e.g. the highest tone of the 10 pitch stimuli was paired with the left-hand side. Fig. 1A shows the progression of Training Phase 1. Once the subject had fixated the central fixation and the auditory stimulus had played, the central fixation character was animated to listen to the left and right by holding their hands to their ears and looking each way. Next, the central fixation disappeared and the 'correct' rectangle flashed twice to draw attention. If the participant then looked at the correct rectangle for the minimum duration (333 ms), the reward animation was triggered. If the participant looked at the wrong rectangle for the minimum duration (333 ms), the rectangle turned black with a cross over it. If the participant did not look at either rectangle within 6 s, the trial timed out, the screen dimmed for 1.5 s and the rectangles turned blank gray. All stimuli then disappeared, and the next trial started once the random inter-trial interval (0.5–3 s) elapsed.

In this phase, the participant was required to look to the correct location in the allotted time on 2 of 2 trials. If they did not meet this criterion, they were presented with another set of 2 trials and this evaluation was repeated as necessary.

Training phase 1B: In this phase the set-up was the converse of Phase 1A stimulus pairing, e.g. the lowest tone of the set of 10 pitch stimuli was paired with the right-hand side.

Training phases 1A & 1B were then repeated once apiece.

Training phase 2: Fig. 1B demonstrates the sequence of events. This phase continued to present only the two extreme stimuli (e.g. highest vs. lowest pitch), but the 'listening' animation was not used, nor did the rectangles flash after the sound was played. The participant needed to use the sound & location mapping they learned in Phase 1 to pass trials. After hearing the auditory stimuli, the two rectangles were presented and participants were expected to fixate one or the other. As before, if they fixated the correct rectangle for the minimum duration (333 ms) they viewed a reward animation; if they fixated the wrong rectangle they saw a cross; if they fixated neither rectangle within 6 s, the screen went gray for 500 ms and the trial ended and proceeded to the next trial. Six trials (3 of each auditory stimulus) were randomly presented and they needed to fixate the correct rectangle on 5 of 6 trials. If they did not meet this criterion, they were presented with another set of 6 and this evaluation was repeated as necessary.

The total minimum number of training trials to complete before progressing to the full experiment was 14. No maximum trial number was set, but if the participant clearly was not grasping the concept, the experiment was terminated.

Testing phase: Here the sequence of visual events was identical to Training Phase 2 (Fig. 1B), but now the three-up/one-down staircase procedure was introduced to move through the auditory levels. As the participant progressed through the trials, when they went 'up' this triggered one of the stimuli to increment one step closer in perceptual similarity. During the entire experiment, one stimulus would shift whereas the other would remain the same acting as a static reference. There were three possible ways for the testing phase to complete. (1) Repeated successful discrimination at the highest level of difficulty four trials in a row. This means their auditory difference threshold is at the highest level. (2) Repeated unsuccessful discrimination at the lowest level of difficulty six trials in a row. This means their auditory difference threshold is at the lowest level or there is some other confound preventing their success. (3) If the subject achieved 11 reversals, that is going up and then down in difficulty, or down then up, the program would come to a halt. We assume that the highest stimulus level at which the reversals occur represents the participant's difference threshold.

At the easiest level the two sounds were the most extreme examples; at the hardest level the two sounds were the most perceptually similar. Therefore, in the pitch experiments, participants distinguished between 415 Hz and 494 Hz in level 1, and between 492 Hz and 494 Hz in level 9. In the rhythm experiment, level 1 required participants to distinguish between 3 bars of equal beats and 3 bars of complex beat combinations. In level 9, both stimuli contained 3 bars of complex short and long beats and they differed only at one point where two eighth notes replaced a single quarter note.

On 1 in 10 successful trials, in addition to the character animation, participants would see an additional brief animation of gold going into a pot, jelly beans into a jar, or a star emerging from a magical box. This was intended to provide variety and entertainment for the participant to keep them engaged to perform correctly. It was decided not to use sounds as a reward to avoid interfering with learning the auditory stimuli. If the participant was very disinterested, we set up the experimental program so that a key was mapped to pause the experiment and play a silent video (in our case an episode from a child's animation program).

Experiments were counterbalanced within and between subjects in terms of side of static reference stimuli (i.e. left or right) and pairing of auditory and visual stimuli (e.g. dinosaur with pitch). Selection of experiment parameters and counterbalancing were entered as needed by the experimenter into the Matlab code before each trial block.

3. Results

Several adults completed pilot tests and demonstrated that they could perform the audition game at or near the highest level of difficulty (level 9).

All TD and ASD children were reported to have normal hearing and no history of hearing impairment, although none underwent formal testing. Table 1 provides details of individual participants' behavioral measures and their performance for the threshold experiment, including the highest level reached and the number of trials completed. As shown in the table some children were not able to exit training for some stimulus sets. Testing was not attempted for every version of the task in every child. This was due to time restraints, or because the participant was clearly unable to grasp the concept of the experiment.

All participants were able to complete several trials by manipulating the orange square on the screen with their eye-gaze to select one of the two boxes.

3.1. TD participants

In summary, all 16 TD children were able to pass training for at least one of the stimulus sets. The number of trials taken to achieve this ranged from the minimum of 14 (taking between 1–2 min to complete) to 110 (taking around 10 min to complete). Note that prolonged times during training most likely tired our participants and in extreme cases like the latter of 110 trials, it may be that training success was due to luck and the participant was not clear on the rules of the game.

TD threshold scores vary from the lowest level to the highest and are quite heterogeneous. An overall mean score was calculated for each participant averaging across all attempted versions, and allocating a ‘0’ score when training criteria was not met. Mean score across all TD participants was 3.6 (s.d. 2.8). Despite our small participant pool, which is underpowered for most statistics, we analyzed correlations. Overall mean scores were tested for a correlation with age, receptive language normed score, and nonverbal IQ normed score. Using adjusted alpha levels of 0.017, there were no significant associations between mean threshold score and age ($r(16) = 0.14$, $p = 0.6$); receptive language ($r(16) = 0.48$, $p = 0.06$); or nonverbal-IQ ($r(16) = 0.34$, $p = 0.2$). An example of a TD participant's performance on a synthetic rhythm experiment test is presented in the left-hand panel of Fig. 2.

Next, mean scores were calculated for ‘pitch’, ‘rhythm’, ‘human’ and ‘synthetic’, by taking the average highest-level when two experiments in the category were completed, or using the available highest-level if only one experiment was completed (and including ‘0’ when the training phase was not passed). To determine if the versions of the experiment were of comparable difficulty, a 2 × 2 ANOVA was conducted with stimulus types as levels: ‘pitch/rhythm’ and ‘synthetic/human’. The interaction was non-significant ($F(1,15) = 1.96$, $p = 0.18$, $\eta^2 = 0.12$), as were the main effects of pitch/rhythm ($F(1,15) = 0.01$, $p = 1.00$, $\eta^2 = 0.01$) and synthetic/human ($F(1,15) = 0.11$, $p = 0.75$, $\eta^2 = 0.01$).

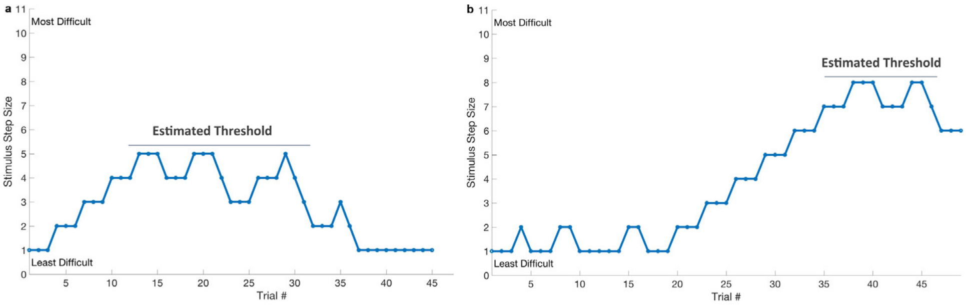

Fig. 2. Two examples of trial-by-trial performance on the test phase. In the 3-up, 1-down regime, each step increase is the result of 3 correct choices, and each decrement is due to 1 wrong choice. In the left panel, the performance of TD participant (‘C01’) is presented on the synthetic rhythm task. This participant completed 78 training trials before progressing to the main experiment; he then completed 44 test trials, of which 32 were correct. His threshold is estimated at level 5. In the right panel an ASD participant's performance (‘A04’) on the human pitch task is presented. This participant completed16 training trials before progressing to the main experiment; he then completed 49 test trials, of which 42 were correct. His threshold is estimated at level 8 (i.e. discrimination between 488 Hz and 494 Hz).

3.2. ASD participants

Only 5 of the 9 ASD children passed training for at least one of the stimulus sets. In the ASD participant group the mean threshold score was 1.8 (s.d. 3.0). Unfortunately, too few ASD children passed the training trials to conduct any meaningful analyses on the results.

In Table 2, we expand briefly on our experiences of testing each of these individuals.

4. Discussion

We have demonstrated a proof of concept gaze contingent method to evaluate auditory thresholds in the typical child population. Unfortunately, performance was mixed in both groups and often unsuccessful with the atypical children we tested. We were forced to question whether the participants had auditory impairments that prevented them from perceiving auditory differences, or if the game was too complex or confusing for our participants. Furthermore, it is unclear whether the varied performance of the control group represented real differences in auditory perception abilities, or whether the introduced variance was a product of the task. To address these concerns there is a need for corroborating information on auditory thresholds from another methodology. Traditional VRA methods with control participants would be a valuable first step, but as this is a problematic methodology in autistic participants, our group is pursuing the use of auditory evoked potentials as a baseline to compare against our method (Ruiz Martinez, Wilson, Yau, Saldaña et al., under review).

Despite apparent limitations of the paradigm, we document some notable successes. First, all participants could be calibrated. This is no small achievement as some participants were had severe intellectual impairments with significant behavioral, cognitive and motor disabilities. The infant calibration method (code for which is provided in the GitHub link) and the high-resolution eye-tracking system were both instrumental to this success. Second, the use of a visible gaze cursor that participants learnt to control was extremely effective, and all participants – regardless of age and level of IQ – successfully did this without being verbally instructed. As such, we have demonstrated that the use of eye-tracking methodology does not need to be passive, but that minimally verbal children can interact with experiments and effectively provide responses. This is a significant contribution to methodology and opens avenues for conducting much needed research in this challenging and neglected population. Third, as a result of pilot testing and several modifications, the final versions of the stimuli used in the different 'games' were of comparable difficulty for typical children. The different versions could therefore be used to compare perceptive ability of pitch versus rhythm information, as well as human versus non-human sounds. Adaptations for testing other auditory qualities (e.g. volume and duration, stimuli available in GitHub link) could easily be incorporated.

4.1. Reflections and future directions

If we assume that the auditory impairments of our participants were not so severe that they could not perceive the differences in even the most extreme stimuli, then we must accept that the difficulties progressing through the experimental 'game' were due to limitations in the protocol. Given the sparsity of our data it is difficult to unravel issues due to stimulus choice, game design, and idiosyncrasies of each participant. Auditory difference thresholds may be attainable with the right combination of stimuli, design and higher function participants, but more coarse discrimination may be possible in other groups. Alternatively, if the task proves too difficult to accurately measure thresholds, the experimental set-up may prove more useful to compare perception of different stimulus types in a within-subject design and between groups.

Nevertheless, we remain hopeful that with careful design our setup could work for auditory differences thresholds. Following the summary of specific challenges faced by each participant (Table 2), we discuss these insights on potential reasons for lack of success and provide ideas for further modifications.

1. *Participants did not understand the game.* The object of the game is for the participant to learn to associate one sound category with one location on the screen, and another sound category with another location on the screen. This concept may have simply been too complicated for some of our participants, or we may have asked too much in expecting them to master the association so quickly. Additionally, our choice of reinforcing stimuli may not have been well tuned for our participants thus inhibiting them from learning the correct associations.
2. *Problems with generalizing.* The test assumes that the child will generalize the rules learned in training phases when confronted with a novel stimulus that is similar to one of the categories they learned. It may be the case that this is confusing, as they do not encounter the generalization stimuli in the training phases and they were unable to grasp the connection between the novel stimulus and the original. Given that individuals with ASD are thought to have difficulties generalizing (Klinger & Dawson, 2001), this may adversely affect their performance (e.g. perhaps the case with participant A05). Perhaps focusing on the training of one 'anchor' stimuli, and later introducing the alternative stimuli, would alleviate this problem. This would create more of a 'same/different' testing paradigm.
3. *Quantity of testing.* A key aim was to test multiple dimensions of auditory stimuli using the same protocol. Unfortunately, this may have introduced problems as participants were expected to learn first one set of stimuli (e.g. pitch) then another (e.g. rhythm). A more effective learning procedure could have focused on one dimension and perhaps completed a training session on one day to familiarize the participant with the concept of the game and learn a visual-auditory association, then returned to complete the experimental task on the following day. Of course, introducing methods that require participants and their families to attend

Table 2
ASD participants and common problems encountered during their test runs.

Participant	Group	Problems encountered	Comments
A03	Minimally verbal ASD	0, 1, 2, 4	This was a severely autistic, low IQ 7-year-old boy with challenging behaviors and problems with **anxiety**. He found the **testing situation stressful** and needed constant reassurance from his mother to stay on task. Nevertheless, we successfully completed calibration and validation, and he went on to complete a total of 68 training phase trials. He was able to interact with the game by using his eye-gaze to 'select' a rectangle, however he never met criteria to pass Training Phase 1. He showed **little responsiveness to the reward** animations therefore it **was unclear if he could not perceive the differences** between stimuli, if he did not associate the visual and auditory stimuli, or if he was simply not motivated to learn the association.
A08	Minimally verbal ASD	1, 2	This was a severely autistic low IQ 9-year-old. He was **cooperative and was easily calibrated**. Every version of the experiment was tested over two sessions on two consecutive days. He was **pleased when he saw the reward animations**, therefore seemed to be motivated to succeed. However, he only passed training criteria in one experiment and this was after completing 74 training trials, therefore likely by chance. **Problems with perseverance** were noted during one experimental session as he repeatedly looked toward the blue rectangle on the left. We c**ould not be sure if he could not perceive the differences between the auditory stimuli or if he did not associate the auditory and visual stimuli.**
A09	Minimally verbal ASD	1,2	This was a severely autistic low IQ 9-year-old boy. He was cooperative and easily calibrated but showed **little responsiveness to the reward animations**. Only one version of the experiment was attempted. He completed 54 training trials and reached the second phase but did not meet criteria to pass. He was interacting with the game, but it was clear he **was not grasping the concept of the auditory-visual association** but was simply looking to either rectangle at random.
A01	Impaired receptive language, verbal ASD	2,4	This 5-year-old boy could communicate verbally, he had low receptive grammar ability but above average non-verbal ability. He was **co-operative and not anxious**. Every version of the task was attempted over two sessions a week apart. Calibration was easily achieved, and he **clearly grasped the concept of interacting with the game** using his eye-gaze. However, he only passed training criteria on one version of the task, and when the main test phase began his concentration was very poor, and he passed only 1 trial out of 8. The experiment was terminated as it was clear he **had lost motivation** and an accurate estimate of auditory perception would not be achieved. For this participant, it is **likely he became bored and unmotivated** before he was able to learn the auditory-visual association.
A04	Impaired receptive language, verbal ASD	None	This 8-year-old boy was the triplet of participants A05 and A06. All three have a diagnosis of ASD. A04 was **compliant, not anxious, easy to calibrate and quick to grasp the concept of the game**. He completed the synthetic rhythm, then the synthetic pitch test in the first session, reaching levels 2 and 3 respectively. He returned 2 weeks later and completed the human pitch (reaching level 8; performance visualized in the right-hand panel of Fig. 2) and finally the human rhythm, on which he passed the highest level thus completing the experiment. This demonstrated that a child with low receptive language ability could do extremely well on the task if motivated to attend and to succeed. Results suggest that either this child was able to discriminate synthetic sounds much more readily than human sounds, or that practice effects were important, as he performed better on his second session.
A05	Impaired receptive language, verbal ASD	2,3,4	This 8-year-old girl was the triplet of participants A04 and A06. She was **compliant, but less quick to grasp the concept of how to interact** with the game. On her first session she completed the synthetic rhythm task, she passed the training after 44 trials by which time she was clearly tired and loosing motivation. Although she reached level 2 of the main test phase, it is possible that this underestimated her ability to perceive the differences between the sounds. When she returned for a second session she completed the human pitch task. She passed the training in 18 trials suggesting she could discriminate the sounds and was successfully associating the visual-auditory stimuli. However, she reached only level 2 in the main test phase. This **could be a true indication of her discrimination threshold, or that she had trouble generalizing** the visual-auditory association to the gradually changing auditory stimulus. She was reluctant to complete another stimulus set so testing was terminated.
A06	Impaired receptive language, verbal ASD	0,1,2	This 8-year-old boy was the triplet of A04 and A05. He was **hyperactive during the testing session and it was difficult to maintain his attention**. With some effort, he was **calibrated successfully, but he found it difficult to move his eyes without moving his head**. He also **perseverated** on the left rectangle. He **enjoyed the reward animations** and completed 112 training trials but still could not reach criteria to pass the training phases.
A02	Typical receptive language, fluent phrase speech ASD	0,1,2	This was a 6-year-old boy with a high level of autistic features, but average receptive language and nonverbal ability levels. He was **anxious during testing and needed constant encouragement** from his mother to remain on task. He completed 86

(*continued on next page*)

Table 2 (*continued*)

Participant	Group	Problems encountered	Comments
			training trials over the course of 31 min (including a break) but was unable to pass Training Phase 2. It was decided to terminate the experiment at that stage, and his mother elected not to return to try on another date as she thought he would not be able to grasp the concept. It was **unclear whether this child was unable to hear the differences** in the rhythm stimuli, **or if the visual-auditory association was beyond his comprehension**.
A07	Typical receptive language, fluent phrase speech ASD	None	This 9-year-old boy had a high level of autistic symptoms but had average receptive language and above average non-verbal IQ. He completed two tasks in one session. He completed the synthetic rhythm task first: he made only one mistake during the training phases and went on to reach level 7. Then he completed the human pitch task: he completed the training phases without making a single mistake, then reached level 9 (although did not pass level 9).

Problems are coded from 0 to 5, with: 0 = anxiety/noncompliance; 1 = did not understand the game; 2 = problems with generalizing; 3 = quantity of testing; 4 = participant lost interest; 5 = problems with memory.

multiple sessions at fixed times can be problematic, but perhaps necessary to achieve success.

4. *Participants lost interest.* Despite our best efforts to create an interesting ‘game’, the repetitive experimental nature may still have resulted in loss of interest. Nevertheless, during development of the software we did have several ideas to increase the level of interest – some of which were adopted, others were rejected (see Section 5 below). To improve this perhaps some game elements could be emphasized, e.g. keeping score, or more attractive graphics and animations could be created.
5. *Problems with memory.* If our participants did not remember the rules or became confused as to when they applied this would result in poor performance. Additionally, the game requires a mental comparison between the current stimulus and the anchor stimulus before a choice can be made. At each step interval, odds are the participant will hear the anchor stimulus, but this was not guaranteed as stimulus selection was randomized. This presentation style may have led to scenarios where the participant had forgotten the anchor and could not correctly compare. In our opinion, this accounts for why some of the typical participants were able to reach high levels on some versions of the game, but could not pass training criteria on others: the problem was not that they could not discriminate between the extreme examples of stimuli, but more likely that they made mistakes and then became confused and were unable to re-establish which auditory stimuli was associated with which rectangle. Again – a solution may be to focus on the training of the anchor stimuli, and only once this is well established to introduce the alternative stimuli and test discrimination against the anchor.

As is clear from Table 2, the elements of the task that the ASD children found challenging were extremely varied: some were anxious or noncompliant, others were not; some were motivated by the reinforcers and enjoyed the ‘game’, but others did not; some easily grasped the audio-visual association, others never grasped it. Moreover, it was not necessarily the children with fluent speech or those with average IQ that were successful. This extreme variability in performance supports the use of individualized approaches advocated by some previous researchers (e.g. Kasari et al., 2013; Plesa Skwerer et al., 2016). Perhaps combining the core elements of the testing paradigm, with individualized reinforcers and/or cues to encourage rapid grasping of the concept, would yield better task success and more reliable indication of auditory discrimination ability.

5. Discussion on design choices

During the process of developing the game, several edits were made to stimuli and protocol. These followed initial pilot testing with adults (prompting revisions in stimuli (points 1 and 2)), and subsequent tests with typical children (leading to revisions in training and test protocol (points 3 and 4)). The key design choices are outlined here so that other researchers may benefit from lessons that were learnt.

1. *Pitch stimuli.* Originally, the pitch stimuli ranged from 330 Hz (‘E’) to 587 Hz (‘D’), with a semitone between each level. The large differences between levels were problematic because participants found it difficult to generalize across incrementing stimuli during the staircase phase. Also, the hardest level – intended to capture threshold discrimination – would have been clearly different to most people of normal hearing. Therefore, stimuli were modified so that differences between levels were much smaller and at the most difficult level the step was 2 Hz – barely perceptible to most individuals of normal hearing.
2. *Rhythm stimuli.* Originally rhythm stimuli were 3.5 s long, to be more in line with pitch stimuli. However, it proved to be much more difficult than the pitch stimuli because of the rapid presentation of beats. In order to make the versions of similar difficulty level, the rhythms were slowed down, although this was at the cost of making the experiment much longer.
3. *Training phases.* The training phases underwent several modifications, with the constant challenge of training participants as

efficiently as possible – any method that took too long would risk boring the participant; any method that had too many variables in it would risk confusing the participant.

First, we attempted showing only one rectangle in the earliest phases, but pilot tests indicated that participants were immediately confused when two rectangles were shown, probably because they had not realized the importance of the auditory stimuli until that point. Therefore, both stimuli were always present, but in Training Phase 1 the correct side flashed to draw attention to it. Second, we created a training scenario where after the audio stimulus was played the central fixation character appeared to listen, by holding up hands to ears, and would then automatically slide over to the correct side. After successful trials the character would no longer automatically slide over to the correct location but was instead controlled by the child's gaze. However, here the child often failed to learn the gaze contingency and would stare at the central fixation waiting for the character to move. Therefore, we included the animation of the character listening to each side to help the participant understand that they must attend to the sound, but did not move the character automatically. It is possible that further non-verbal instruction may have helped, while we did not implement this, it would be possible to include a graphic cue that indicates what the child should pay attention to, for instance when testing pitch an arrow pointing up or down might be useful.

Third, we piloted versions with far more trials in Training Phase 1 (blocks of the most extreme stimuli) but found that participants became bored, and also forgot the first stimuli when listening to the second. In order to balance the presentations whilst providing enough repetitions to learn the distinct stimuli, 2 sets of 2 examples on each side was decided on. Fourth, we tried using much longer inter-trial intervals so that learning of stimuli could be assimilated, but participants quickly lost interest.

4. *Test phase.* We considered having the child's progress be represented onscreen, either by having the extra reward animations with gold, jelly beans, or stars have a marker to stay onscreen to show how many awards have been collected; this was not done to avoid visual clutter. We also considered having the onscreen stimuli start at the bottom of the screen and move upwards whenever the child progressed a step up in the level of difficulty, but we reasoned this may be confusing to the child. Using sound stimuli to encourage participants or re-enforce success or failure of a trial was also considered but rejected because it may interfere with learning the target sounds stimuli.

6. Conclusion

There are many aspects to consider when turning a psychophysical test into a game that children of a variety of abilities have the capacity to enjoy and easily learn and interact with. Despite apparent limitations in our approach, we have made significant progress in this regard and document our efforts so that the field may continue to move forwards. The code we designed, and the well-controlled auditory and visual stimuli, are now freely available to other researchers to use and adapt.

Acknowledgements

CEW was supported by a Talentia Postdoc launched by the Andalusian Knowledge Agency, co-funded by the European Union's Seventh Framework Program, Marie Skłodowska-Curie actions (COFUND – Grant Agreement n° 267226) and the Ministry of Economy, Innovation, Science and Employment of the Junta de Andalucía, Spain. BTS was supported by the Marie Skłodowska-Curie actions LanPercept Innovative Training Network (FP7-PEOPLE-2012-ITN, Project ID: 316748).

We thank Joel Talcott and Santiago Benjumea for helpful discussions on method design. Also to Marina Barrio for her assistance in producing the sound stimuli.

Finally, we are extremely grateful to all the participants and their families that participated in this study.

References

Auyeung, B., Baron-Cohen, S., Wheelwright, S., & Allison, C. (2007). The autism spectrum quotient: Children's version (AQ-child). *Journal of Autism and Developmental Disorders, 38*, 1230–1240.

Bertoncini, J., Nazzi, T., Cabrera, L., & Lorenzi, C. (2011). Six-month-old infants discriminate voicing on the basis of temporal envelope cues (L). *The Journal of the Acoustical Society of America, 129*(5), 2761–2764.

Bishop, D. V. (2003). *Test for reception of grammar: Vesion 2: TROG-2.* Harcourt Assessement.

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

Boersma, P., & Weenink, D. (2017). *Praat: Doing phonetics by computer [Computer program].* Version 6.0.27, retrieved 17 March 2017 from http://www.praat.org/.

Bonnel, A., McAdams, S., Smith, B., Berthiaume, C., Bertone, A., Ciocca, V., ... Mottron, L. (2010). Enhanced pure-tone pitch discrimination among persons with autism but not Asperger syndrome. *Neuropsychologia, 48*(9), 2465–2475.

Bonnel, A., Mottron, L., Peretz, I., Trudel, M., Gallun, E., & Bonnel, A. M. (2003). Enhanced pitch sensitivity in individuals with autism: A signal detection analysis. *Journal of Cognitive Neuroscience, 15*(2), 226–235.

Brueggeman, P. (2012). 10 tips for testing hearing in children with autism. *The ASHA Leader, 17*(1), 5–7.

Carr, K. W., White-Schwoch, T., Tierney, A. T., Strait, D. L., & Kraus, N. (2014). Beat synchronization predicts neural speech encoding and reading readiness in preschoolers. *Proceedings of the National Academy of Sciences USA, 111*(40), 14559–14564.

Crane, L., Goddard, L., & Pring, L. (2009). Sensory processing in adults with autism spectrum disorders. *Autism, 13*(3), 215–228.

Dievendorf, A. O., & Gravel, J. S. (1996). Behavioral observation and visual reinforcement audiometry. In S. E. Gerber (Ed.). *The handbook of pediatric audiology, Chapter 4* (pp. 55–83). Gallaudet University Press.

Downs, D., Schmidt, B., & Stephens, T. J. (2005). *Auditory behaviors of children and adolescents with pervasive developmental disorders. Seminars in Hearing (Vol. 26, No. 04)*. 333 Seventh Avenue, New York, NY 10001, USA: Copyright© 2005 by Thieme Medical Publishers, Inc226–240.

Drullman, R., Festen, J. M., & Plomp, R. (1994). Effect of temporal envelope smearing on speech reception. *The Journal of the Acoustical Society of America, 95*(2), 1053–1064.

Dunn, W. (2014). *Sensory profile 2 assesment.* London, UK: Pearson Clinical.

Egelhoff, K., Whitelaw, G., & Rabidoux, P. (2005). *What audiologists need to know about autism spectrum disorders. Seminars in Hearing (Vol. 26, No. 04)*. 333 Seventh

Avenue, New York, NY 10001, USA: Copyright© 2005 by Thieme Medical Publishers, Inc202–209.

Elliott, T. M., & Theunissen, F. E. (2009). The modulation transfer function for speech intelligibility. *PLoS Computational Biology, 5*(3), e1000302.

Fujii, S., & Wan, C. Y. (2014). The role of rhythm in speech and language rehabilitation: The SEP hypothesis. *Frontiers in Human Neuroscience, 8*, 777.

Gordon, R. L., Shivers, C. M., Wieland, E. A., Kotz, S. A., Yoder, P. J., & Devin McAuley, J. (2015). Musical rhythm discrimination explains individual differences in grammar skills in children. *Developmental Science, 18*(4), 635–644.

Heaton, P., Williams, K., Cummins, O., & Happé, F. (2008). Autism and pitch processing splinter skills: A group and subgroup analysis. *Autism, 12*(2), 203–219.

Huss, M., Verney, J. P., Fosker, T., Mead, N., & Goswami, U. (2011). Music, rhythm, rise time perception and developmental dyslexia: Perception of musical meter predicts reading and phonology. *Cortex, 47*(6), 674–689.

Jones, C. R., Happé, F., Baird, G., Simonoff, E., Marsden, A. J., Tregay, J., ... Charman, T. (2009). Auditory discrimination and auditory sensory behaviours in autism spectrum disorders. *Neuropsychologia, 47*(13), 2850–2858.

Kasari, C., Brady, N., Lord, C., & Tager-Flusberg, H. (2013). Assessing the minimally verbal school-aged child with autism spectrum disorder. *Autism Research, 6*(6), 479–493.

Kaufman, A. S., & Kaufman, N. L. (2004). *KBIT: Kaufman brief intelligence test (KBIT Spanish version).* Madrid: TEA Editions.

Klinger, L. G., & Dawson, G. (2001). Prototype formation in autism. *Development and Psychopathology, 13*(1), 111–124.

Klintwall, L., Holm, A., Eriksson, M., Carlsson, L. H., Olsson, M. B., Hedvall, Å., ... Fernell, E. (2011). Sensory abnormalities in autism: A brief report. *Research in Developmental Disabilities, 32*(2), 795–800.

Kuhl, P. K., Coffey-Corina, S., Padden, D., & Dawson, G. (2005). Links between social and linguistic processing of speech in preschool children with autism: Behavioral and electrophysiological measures. *Developmental Science, 8*(1), F1–F12.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... Rutter, M. (2000). The autism diagnostic observation schedule—generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders, 30*(3), 205–223.

Luck, S. J. (2005). An introduction to event-related potentials and their neural origins. *An Introduction to the Event-Related Potential Technique* (pp. 1–50).

Mendoza, E. (2005). *CEG: Test de comprensión de estructuras gramaticales.* Madrid: TEA Editions.

Mendoza, E., Carballo, G., Muñoz, J., & Fresneda, M. D. (2005). Evaluación de la comprensión gramatical: un estudio translingüístico. *Revista de Logopedia, Foniatría y Audiología, 25*(1), 2–18.

Miyazaki, M., Takahashi, H., Rolf, M., Okada, H., & Omori, T. (2014). The image-scratch paradigm: A new paradigm for evaluating infants' motivated gaze control. *Scientific Reports, 4*.

Mueller, J. L., Friederici, A. D., & Männel, C. (2012). Auditory perception at the root of language learning. *Proceedings of the National Academy of Sciences USA, 109*(39), 15953–15958.

Musiek, F. E., & Baran, J. A. (2007). *The Auditory system.* Boston, MA: Pearson Education, Inc.

Näätänen, R., Paavilainen, P., Rinne, T., & Alho, K. (2007). The mismatch negativity (MMN) in basic research of central auditory processing: A review. *Clinical Neurophysiology, 118*(12), 2544–2590.

Paulraj, M. P., Subramaniam, K., Yaccob, S. B., Adom, A. H. B., & Hema, C. R. (2015). Auditory evoked potential response and hearing loss: A review. *The Open Biomedical Engineering Journal, 9*(1).

Plesa Skwerer, D., Jordan, S. E., Brukilacchio, B. H., & Tager-Flusberg, H. (2016). Comparing methods for assessing receptive language skills in minimally verbal children and adolescents with autism spectrum disorders. *Autism, 20*(5), 591–604.

Rosenhall, U., Nordin, V., Sandström, M., Ahlsen, G., & Gillberg, C. (1999). Autism and hearing loss. *Journal of Autism and Developmental Disorders, 29*(5), 349–357.

Sabo, D. L. (1999). The audiologic assessment of the young pediatric patient: The clinic. *Trends in Amplification, 4*(2), 51–60.

Samson, F., Mottron, L., Jemel, B., Belin, P., & Ciocca, V. (2006). Can spectro-temporal complexity explain the autistic pattern of performance on auditory tasks? *Journal of Autism and Developmental Disorders, 36*(1), 65–76.

Schall, U. (2016). Is it time to move mismatch negativity into the clinic? *Biological Psychology, 116*, 41–46.

Schwarz, I. C., Nazem, A., Olsson, S., Marklund, E., & Uhlén, I. (2014). Towards a contingent anticipatory infant hearing test using eye-tracking. *FONETIK 2014, the XXVIIth Swedish Phonetics Conference* (pp. 35–40).

Schwartz, S., Shinn-Cunningham, B., & Tager-Flusberg, H. (2018). Meta-analysis and systematic review of the literature characterizing auditory mismatch negativity in individuals with autism. *Neuroscience & Biobehavioral Reviews, 87*, 106–117.

Slater, J., Skoe, E., Strait, D. L., O'Connell, S., Thompson, E., & Kraus, N. (2015). Music training improves speech-in-noise perception: Longitudinal evidence from a community-based music program. *Behavioural Brain Research, 291*, 244–252.

Stanutz, S., Wapnick, J., & Burack, J. A. (2014). Pitch discrimination and melodic memory in children with autism spectrum disorders. *Autism, 18*(2), 137–147.

Stevenson, R. A., Siemann, J. K., Woynaroski, T. G., Schneider, B. C., Eberly, H. E., Camarata, S. M., & Wallace, M. T. (2014). Evidence for diminished multisensory integration in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 44*(12), 3161–3167.

Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: The neglected end of the spectrum. *Autism Research, 6*(6), 468–478.

Wang, Q., Bolhuis, J., Rothkopf, C. A., Kolling, T., Knopf, M., & Triesch, J. (2012). Infants in control: Rapid anticipation of action outcomes in a gaze-contingent paradigm. *PLoS ONE*.

Wass, S., Porayska-Pomsta, K., & Johnson, M. H. (2011). Training attentional control in infancy. *Current Biology, 21*(18), 1543–1547.

Watson, L. R., Roberts, J. E., Baranek, G. T., Mandulak, K. C., & Dalton, J. C. (2012). Behavioral and physiological responses to child-directed speech of children with autism spectrum disorders or typical development. *Journal of Autism and Developmental Disorders, 42*(8), 1616–1629.

Yau, S. H., McArthur, G., Badcock, N. A., & Brock, J. (2015). Case study: Auditory brain responses in a minimally verbal child with autism and cerebral palsy. *Frontiers in Neuroscience, 9*.