

Behavioral Study on Emotional Voice Perception in Children with Autism Spectrum Disorder

Keita Matsumoto¹ · Toshiro Sugiyama² · Chiaki Saito³ · Shiho Kato⁴ · Kikuko Kuriyama⁵ · Kousuke Kanemoto⁶ · Akinori Nakamura⁷

Received: 17 March 2016 / Revised: 26 August 2016 / Accepted: 26 September 2016 / Published online: 13 October 2016
© American Academy of Pediatric Neuropsychology 2016

Abstract Previous studies conducted on children with autism spectrum disorder (ASD) and their abilities to recognize the psychological states of others have primarily focused on visual aspects such as facial expression and perception. The aims of this study were to clarify the characteristics of auditory information processing mechanisms in individuals with ASD in order to expand our understanding of those with ASD and to contribute to improvements in their social adaptability. The behavioral characteristics of voice cognition in ASD were investigated in 12 individuals with ASD (mean age = 11.83 years) and 12 matched typical development (TD) controls (mean age = 11.58 years). Participants were asked to judge the emotional valence of vocal stimuli with happy,

angry, or neutral expressions (emotion task) and select illustrations that best correspond to the semantics of the vocal stimuli (semantic task). The ASD group had significantly ($p = 0.013$) longer reaction times in the emotion task, suggesting that they have difficulty distinguishing emotional valence. In the semantic task, the ASD group had lower understanding of the meaning of sentences when voices were presented with angry expressions. These results suggest the importance of the emotions conveyed in verbal speech during conversations with individuals with ASD.

Keywords Autism spectrum disorder · Emotional voice perception · Semantic perception · Reaction time · Auditory

✉ Keita Matsumoto
ke-matsumoto@med.osakacity-hp.or.jp

¹ Department of Child and Adolescent Psychiatry, Osaka City General Hospital, 2-13-22 Miyakojima-hondori, Miyakojima-ku, Osaka City 534-0021, Japan

² Child and Adolescent Psychiatry, Hamamatsu University School of Medicine, 1-20-1 Handayama, Higashiku, Hamamatsu, Shizuoka, Japan

³ Center for Dementia Care Research and Practices in Obu, 3-294 Hantsuki-cho, Obu City, Aichi, Japan

⁴ Department of Psychiatry, Obu Hospital, 1 Kamigengo, Morioka, Higashiura Cho, Aichi, Japan

⁵ Department of Child and Adolescent Psychiatry, Aichi Children's Health and Medical Center, 1-2 Osakada, Morioka-cho, Obu City, Aichi, Japan

⁶ Department of Neuropsychiatry, Aichi Medical University, 1-1 Yazako Karimata, Nagakute, Aichi, Japan

⁷ Department of Clinical and Experimental Neuroimaging, National Center for Geriatrics and Gerontology, 7-430, Morioka-cho, Obu City, Aichi, Japan

Introduction

The semantics of words are not the only aspects that are exchanged during communication. Whether the speaker is conscious of it or not, a message containing visual information such as facial expressions and gestures, as well as auditory information such as tone and pitch of the voice, is projected to the listener. From the listener's perspective, while there may be some differences depending on the time and place, an effort is made to receive nonverbal messages that accompany the semantics of words, in order to accurately understand the speaker's intention.

Birdwhistell (1970) suggested that only 35 % of all dialogs between two individuals contain verbal information that can be communicated through semantics of words and that informed decisions through nonverbal information account for the remaining 65 %. Nonverbal information that is shared during conversations with others contains the following: visual information such as facial expressions, eye contact, and gestures; phonetic information through prosody such as dynamics, intonation, and pitch of the voice; and non-visual and

non-phonetic information that is based on the assumption and background of the conversation. In most cases, nonverbal information accompanies spoken words and serves as supplementary information.

Autism spectrum disorder (ASD) is a neurological condition that is characterized by a persisting impairment in mutual social communication or interpersonal interaction and by a limited and repetitive form of behavior, interest, or activity (Diagnostic and Statistical Manual of Mental Disorders Fifth edition [DSM-V], American Psychiatric Association [APA], 2013). In individuals with ASD, their lack of interpersonal interaction to share ideas, emotions, and interests with others causes difficulties in processing and reacting to social cues, and this consequently leads to social adaptability issues (DSM-V, American Psychiatric Association 2013). Even if the individual is highly intelligent, those with ASD are known to have communication issues such as implicit, automatic, and intuitive mentalizing difficulty (Lai, Lombardo, and Baron-Cohen 2014). Previous work has suggested that the deficiency of these communication abilities is one of the factors that lead to the prevention of social success, hindrance in forming friendships, and consequently isolation from others (Corbett, Qualls, Valencia, Fecteau, and Swain 2014). The causes of these difficulties have been explained by several hypotheses, e.g., inherent difficulties in recognizing the emotions of other people resulting in impaired interpersonal mutual reaction (Hobson 1988), developmental problems with “theory of mind,” which is the ability to infer the psychological state of others (Baron-Cohen 2004), deficits in cognitive flexibility which is the ability to switch rapidly between multiple task problems of the joint attention (de Vries and Geurts 2012), and problems in joint attention to share a focus of interest with others (Gernsbacher, Stevenson, Khandakar, and Goldsmith 2008). Moreover, while correlations between the difficulties individuals with ASD face in terms of social adaptability and problems with perceiving or expressing prosody have been demonstrated (Paul et al. 2005), most studies on the recognition of the mental state of others in children with ASD have primarily focused on visual aspects such as recognizing facial expressions, and only a handful of studies have targeted the cognitive ability of auditory aspects (Lartseva, Dijkstra, and Buitelaar 2015; O’Connor 2012).

In the present study, we therefore aimed to clarify the following two questions through a detailed behavioral investigation in order to delineate the characteristics of children with ASD in perceiving emotional voice.

The first question is are there differences between those with ASD and typical development (TD) in the ability to distinguish different emotions expressed in voice? Views on the ability of children with ASD to decipher emotions expressed in voice differ depending on the reported study, and a clear conclusion has yet to be reached. Mazefsky and Oswald (2007) presented sentences that were each read with four different types of emotions to 16 children with Asperger’s

syndrome (mean age = 11.47 years) and to 14 children with high-functioning ASD (mean age = 11 years) and asked them to classify the emotion on a multiple-choice list. Results showed that the children with high-functioning ASD had a markedly lower mean score compared to those with Asperger’s syndrome or TD. Additionally, Golan, Sinai-Gavrilov, and Baron-Cohen (2015) presented voice recordings of nine complex emotions using video clips to 30 children with high-functioning ASD (mean age: 9.7 years) and 25 children with TD (mean age = 10 years) and asked them to choose the correct emotion. The results of that study showed that the accuracy of children with ASD was markedly lower compared to those with TD. There are multiple reports that demonstrate individuals with ASD have difficulty distinguishing different emotions in voice (Golan et al. 2015; Hobson 1986; Lindner and Rosen 2006; Mazefsky and Oswald 2007; Philip et al. 2010). On the other hand, Jones et al. (2011) presented voice recordings of numbers read in one of six types of emotions and non-verbal sounds such as crying or laughter to 99 adolescents with ASD (mean age = 15.6 years) and 57 adolescents with TD (mean age = 15.6 years) and asked the participants to discriminate different emotions; they concluded that significant differences in accuracy were not observed between the two groups. Similarly, Grossman, Bemis, Plesa-Skwerer, and Tager-Flusberg (2010) presented sentences that had been filtered to eliminate semantic content and normal sentences that were read with three types of emotions to 16 children with high-functioning ASD (mean age = 12.4 years) and 15 peers with TD (mean age = 12.7 years) and asked the participants to disambiguate the emotion of each voice sample. The authors reported that both ASD and TD groups had similar abilities to determine emotion in either task (Grossman et al. 2010).

In the present study, we anticipated that we would be able to answer this question from a new angle by focusing not only on the accuracy of deducing the correct emotion, but also on the reaction time. Because assessing answers that are either correct or incorrect is a mere two-point evaluation, indicating that if the IQ is within normal range and the individual is adapted to everyday life to some degree, we postulated that a difference in accuracy may be difficult to detect only by assessing the accuracy. Additionally, supposing that emotions are determined through different strategies between individuals with adapted ASD vs TD (Lartseva et al. 2015; O’Connor 2012), it is predicted that a longer time is required for individuals with ASD to reach an accurate decision, and this indicates the possibility that a difference in the reaction time could be extracted between the two groups of individuals. Although there have been several studies that assessed the reaction time in perceiving facial expression from a visual aspect (Fink et al. 2014; Rump et al. 2009; Tracy et al. 2011), very few, including the aforementioned studies, have focused on the reaction time related to the auditory perception of expressions.

Our second question is does an emotional voice affect the auditory perception of semantic information in individuals with ASD? In previous studies on the effects of changes in voice on the perceptual processing of auditory meaning in ASD, it has been reported that the ability to recognize the meaning of words declines in ASD compared to TD depending on stress patterns (Grossman et al. 2010) and that the sentence semantic recognition ability declines in ASD depending on intonation (Peppé et al. 2007); however, very few studies have focused on how an emotionally toned voice affects the auditory perception of semantic information. Since there are many individuals with ASD who experience difficulties in exchanging emotionally toned verbal communication in everyday life (Lartseva et al. 2015; O'Connor 2012; Williams 1992), we consider it essential and meaningful to answer this question.

Methods

Participants

Twenty-four right-handed native Japanese children aged 10–15 years participated in all experiments. Twelve of these participants had ASD (six boys, six girls; mean age = 11.83 years, SD = 1.90), and the other 12 had TD (six boys, six girls; mean age = 11.58 years, SD = 1.68). The ASD group was randomly selected from patients who were receiving care at the Aichi Children's Health and Medical Center on an outpatient basis and were diagnosed with high-functioning ASD but were not undergoing treatment with medication. We conducted an interview with all patients and used DSM fourth edition (APA 1994) and the Pervasive Developmental Disorders Autism Society Japan Rating Scale (PARS; Autism Society Japan 2008) for diagnosis. PARS is the standard scale for evaluating ASD in Japan with established reliability and validity, and it has also been reported for its correlation with the Autism Diagnostic Interview-Revised (ADI-R; Ito et al. 2012; Lord et al. 1994). All participants in the ASD group in this study were verified to exceed the cut-off score for ASD diagnostic

criteria based on the PARS score. The TD group was recruited from the general population within nearby communities. We confirmed through interviews that all participants in both groups attended a regular elementary or junior high school in Japan, had satisfactorily adapted to everyday life without any specific support, had normal vision with or without correction, and did not have any comorbid hearing impairments or other mental disorders. There was no statistical difference between the two groups in the male-to-female ratio. In this study, IQ was measured using Wechsler Intelligence Scale for Children third edition (WISC-III)(Wechsler 1991) or fourth edition (WISC-IV)(Wechsler 2003). In the Japanese version of these two tests, the correlations are known to be reliably high ($r=0.81$ for Full Scale IQ; FSIQ) (WISC-IV, Nihon Bunka Kagakusha, Tokyo, Japan). There were no significant differences in FSIQ or in subtests for the Verbal Comprehension Index (VCI) and Perceptual Reasoning Index (PRI) between the two groups (t test, FSIQ: $p=0.057$, VCI: $p=0.209$, PRI: $p=0.258$; Table 1). However, the processing speed index (PSI) was significantly lower in the ASD group than the TD group ($p=0.001$).

This study was planned in accordance with the Declaration of Helsinki and was approved by the first author's institutional ethics committee. After detailed explanations of the study were provided, written informed consent was obtained from both participants and their guardians.

Stimuli

Selection of Stimulus Words Using the Nippon Telegraph and Telephone Corporation (NTT) Database Series "Lexical Properties of Japanese" (Amano and Kondo 1999), nouns with three moras and present-tense verb phrases with six moras were selected (60 nouns and six verbs) from words that have the greatest familiarity among Japanese people. The NTT Database Series comprehensively manages nearly all Japanese words and is a Japanese language database that is used in research and information analysis. Familiarity is a subjectively assessed value pertaining to the degree in which humans are familiar with a specific word. Additionally, among

Table 1 Participant demographics

	ASD	TD	Statistical differences p value (Student t test)	Effect size (Cohen's d)
Participants (male)	12 (6)	12 (6)		
Age (SD)	11.83 (1.90)	11.58 (1.68)	0.736	0.14
FSIQ (SD)	96.92 (13.69)	106.75 (9.99)	0.057	0.82
VCI (SD)	97.58 (13.74)	106.00 (17.86)	0.209	0.53
PRI (SD)	97.42 (14.64)	103.67 (11.57)	0.258	0.47
PSI (SD)	95.92 (14.25)	117.00 (12.98)	0.001	1.55

Values represent mean (SD). ASD autism spectrum disorder, TD typical development, FSIQ Full Scale IQ, PRI Perceptual Reasoning Index, VCI Verbal Comprehension Index, PSI Processing Speed Index

variables that influence the comprehension and generation of words, familiarity is widely accepted as a variable that determines the speed and degree of difficulty in lexical access. Familiarity scoring was established by Amano and Kondo (2000) and is categorized into four levels (Max 7.0, 7.0–5.5, 5.5–4.0, 4.0–2.5, 2.5–1.0). In the present study, words and phrases that were thought to be highly recognizable especially to children were selected from words with a familiarity score of 7.0–5.5. Mora is a unit of phonology in Japanese that determines syllable weight, and the combination of consonant phoneme and vowel phoneme forms one mora. In Japanese, kana are syllabic scripts and one kana character equates to one mora.

Creation of Stimulus Sets The selected words were recorded by a professional female announcer whose native language is Japanese. The voice recordings consisted of three different emotions (a happy voice, an angry voice, and a neutral voice). The recorded vocal audio files were processed with Sound Engine Free 4.60 (Coderium, Sapporo, Japan) on a personal computer to equalize the volume. Each word was extracted as a Waveform (RIFF waveform Audio Format) file in a 1100-msec fragment by aligning the audio (voice) start times and inserting spaces at the end of the word. Sentences (all were simple sentences composed of a subject-object-verb word order) were constructed using a combination of these fragments with a living organism as the subject so that they had a natural meaning and form. By inserting spaces between words, sentences were output into vocal audio files that were all 3800 msec in length (this length is the most natural-sounding and is easy to understand). A total of 60 words and 64 sentences constructed from four types of animals as the subject (fox, panda, gorilla, raccoon dog), four verbs (hide, bite, wash, see), and four objects (apple, cucumber, banana, mandarin orange) ($4 \times 4 \times 4 = 64$) were used as auditory stimuli.

In order to establish whether the emotional valence and semantics of the voice stimuli presented were properly recognized, the stimuli were presented to healthy student volunteers with normal hearing. There were four male and 17 female healthy volunteers who were between 20 and 29 years old

(mean age=21.1 years). The healthy volunteers were asked to write down the emotion and semantics of the stimuli presented. As a result of this pre-screening of stimuli, 45 words and 60 sentences with an accuracy of >95 % were ultimately chosen as the auditory stimuli (see Table 2). In addition, illustrations [400 (height) × 600 (width) pixels] that matched the semantics of the word or sentence were created as visual stimuli.

Experimental Procedure

The experiment was conducted on a one-on-one basis in a quiet room. Participants were asked to sit in front of a 23-inch desktop liquid crystal display (RDT233WX, Mitsubishi Electric Corporation, Japan). The distance between the liquid crystal display and participant was set at approximately 40 cm with a viewing angle of approximately 40°.

The experiment consisted of an “emotion task” where participants were asked to distinguish between the emotional valence (happy, angry, or neutral) of a voice presenting stimuli, as well as a “semantic task” where participants were asked to select the illustration that best corresponded to the semantics of the vocal stimuli. There were two sessions for each of the tasks, one session for word stimuli and another for sentence stimuli.

The presentation of stimuli was controlled by a computer using Presentation (Neurobehavioral Systems, Inc., USA) software, and the voices were presented using audio speakers (Companion 2 Series II Multimedia Speaker System, BOSE Corporation, USA) set next to the liquid crystal display. The volume was set such that the voice could be heard clearly but did not cause discomfort.

Three custom-made buttons were used to measure reaction times; participants were instructed to press the buttons with their dominant, right hand. Prior to the actual experiment, practice sessions were conducted using stimuli different from the ones used in the experiment to familiarize the participants with appropriately maneuvering the buttons.

Emotion Task—Word Session The sequence of word stimuli was randomized, and the same stimulus sequence was presented to all participants. During the sessions, participants

Table 2 Examples of word and sentence stimuli

(A) Word session Japanese (pronunciation)	/ English	(B) Sentence session Japanese (pronunciation)	/ English
ブドウ (budou)	/ grape	キツネがリンゴをかくしている (Kitsune ga ringo wo kakushiteiru)	/ A fox is hiding an apple.
トケイ (tokei)	/ clock	パンダがキュウリをかじっている (Panda ga kyuuri wo kajitteiru)	/ A panda is biting a cucumber.
イルカ (iruka)	/ dolphin	ゴリラがバナナをあらっている (Gorilla ga banana wo aratteiru)	/ A gorilla is washing a banana.

were asked to infer the emotion that was expressed by the voice and respond by pressing the corresponding button. Buttons were arranged in the following manner: “happy” on the left, “neutral” in the middle, and “angry” on the right. Reaction time was measured with the start of voice presentation as the zero point, and participants were asked to respond at the time in which they determined the answer, even if that occurred during the voice presentation.

Emotion Task—Sentence Session Similar to the word session, sentences were presented, and participants were asked to deduce the emotional valence of the stimuli presented by the voice and answer accordingly using the pre-assigned buttons for each emotion. Reaction time was measured as time between voice presentation (zero point) and when the participants pressed the button. The reaction time was measured from the time in which the voice was presented, and participants were asked to respond as soon as they recognized the emotion, even if that occurred before the end of the stimulus presentation.

Semantic Task—Word Session Similar to the emotion task, words were presented using an emotional tone, and three illustrations that corresponded to the semantics of the audio presentation were displayed on the monitor 0.3 s later. Participants were asked to select the illustration that matched the voice using the button placed in front of each illustration (Fig. 1a). Reaction time was measured as the time between visual stimuli presentation (zero point) and when the participants pressed the button.

Semantic Task—Sentence Session Using a similar method to the word session, sentences were presented instead of words (Fig. 1b), and three illustrations that corresponded to the semantics of the audio presentation were displayed on the monitor 0.3 s later. Participants were asked to select the

illustration that matched the voice using the button placed in front of each illustration. Reaction time was measured as the time between visual stimuli presentation and when the participants pressed the button (Fig. 1b).

Statistical Analysis

For accuracy of the emotion and semantic tasks, we first checked the normality of the data using the Kolmogorov-Smirnov test. Since the data were not normally distributed, Mann-Whitney *U* tests were conducted for every stimulus (word, sentence) and emotion presented (happy, neutral, angry) in order to compare differences between groups. Bonferroni corrections were applied to the *p* values. Regarding measurement of reaction time, three-way ANOVAs were performed using three factors—stimulus, emotion, and diagnosis (ASD group or TD group)—and three-way interaction as well as simple interactions and main effects were analyzed. Sphericity was also validated at the same time using Mauchly’s test, and Greenhouse-Geisser corrections were applied when the results were significant. The Bonferroni correction was used for post hoc analysis. The effect size was evaluated with η^2_p . SPSS statistical software version 21 (IBM, New York, USA) was used for all analyses.

Results

Emotion Task

The results of accuracy and reaction time are presented in Table 3. Accuracy was $\geq 97\%$ regardless of stimulus or emotion, and there were no significant differences between the groups. The three-way ANOVA results on reaction time are displayed in Table 4.

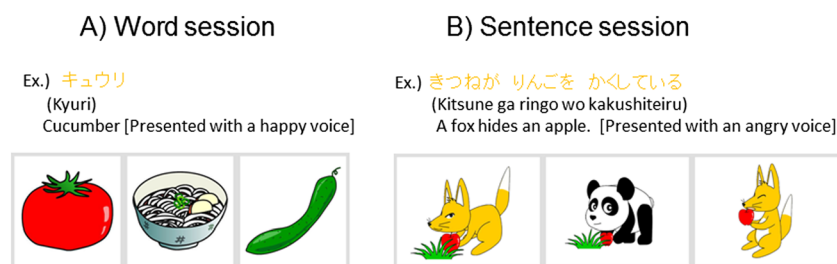


Fig. 1 Examples of word (a) and sentence (b) stimuli in the semantic task. A word or a sentence was spoken with one of three emotions (happy, neutral, or angry). An illustration showing the meaning of the word or the sentence was then displayed on a computer monitor, and participants were asked to press the button corresponding to the picture that they thought was correct out of the three options presented. In the sentence session, three images were selected based on the following criteria: using this particular illustration as an example, the voice stimulus “A fox hides an apple” is presented in Japanese as “Kitsune ga ringo wo kakushiteiru.”

The correct answer cannot be reached by only listening and understanding the first word “kitsune” (fox) or the subsequently presented “ringo” (apple) or even the combination of “kitsune + ringo,” but can be deduced only after listening all the way to the end of the sentence to “kakushiteiru” (hide). As described here, the combination of the three simultaneously presented illustrations was counterbalanced and arranged such that the answer was not limited to one choice during the middle of the sentence

Table 3 Accuracy and reaction times during the emotion task

	Word session						Sentence session					
	Accuracy (%)			Reaction time (msec)			Accuracy (%)			Reaction time (msec)		
	Neutral	Happy	Angry	Neutral	Happy	Angry	Neutral	Happy	Angry	Neutral	Happy	Angry
ASD	97	99	97	1322	1250	1283	100	98	98	1134	1153	1098
(SD)	(4)	(3)	(4)	(281)	(206)	(221)	(0)	(3)	(4)	(309)	(348)	(280)
TD	97	100	99	1072	1031	1055	100	99	98	913	937	887
(SD)	(4)	(0)	(4)	(184)	(179)	(249)	(1)	(3)	(2)	(169)	(153)	(161)

Values represent mean (SD). SD standard deviation, ASD autism spectrum disorder, TD typical development

The three-way ANOVA on reaction time did not yield a significant three-way interaction for stimulus × diagnosis × emotion. However, main effects for diagnosis were observed, and the perception of emotion was significantly delayed in the ASD group (word average: 1284 ± 226 msec, sentence average: 1129 ± 305 msec) compared to the TD group (word average: 1051 ± 195 msec, sentence average: 913 ± 155 msec) regardless of stimulus or emotion. Specifically, these delays were 233 msec for words and 216 msec for sentences (Fig. 2).

Although the two-way ANOVA showed an interaction effect for stimulus × emotion ($p=0.003$, $\eta^2_p=0.244$), post hoc tests did not show significant differences between emotions in either stimulus. However, simple main effects were observed for stimulus ($p=0.002$, $\eta^2_p=0.371$), and sentence stimuli (average: 1020 ± 261 msec) exhibited shorter reaction times compared to word stimuli (average: 1169 ± 238 msec), regardless of emotion or diagnosis.

Interactions for diagnosis were not observed for three-way or two-way ANOVAs.

Semantic Task

The results from accuracy measurements and reaction times are presented in Table 5. For the sentence stimuli, when voices with an angry emotion were presented, the accuracy was

94.6 % in the ASD group, which was significantly lower compared to the TD group (Mann-Whitney U test $p=0.024$, Bonferroni corrected) (Fig. 3).

The results of a three-way ANOVA on reaction time are displayed in Table 6. While the results did not show a three-way interaction between stimulus × diagnosis × emotion ($p=0.115$, $\eta^2_p=0.093$), main effects for diagnosis ($p=0.031$, $\eta^2_p=0.195$) and two-way interactions for stimulus × diagnosis were observed ($p=0.036$, $\eta^2_p=0.185$). The difference between the ASD and TD groups regarding reaction time was influenced by the stimulus, and while there were no differences between the groups regarding word stimuli (ASD average: 1234 ± 192 msec, TD average: 1153 ± 190 msec), the ASD group was shown to have significantly delayed (ASD average: 1819 ± 692 msec, TD average: 1282 ± 360 msec) semantic perception compared to the TD group on sentence stimuli ($p=0.026$, $\eta^2_p=0.206$) (Fig. 4). Specifically, these delays in the ASD group compared to the TD group were 81 msec for words and 537 msec for sentences.

Similar to the emotion task, a two-way interaction was observed between stimulus and emotion. Post hoc analysis showed that, regardless of diagnosis, the reaction time for word stimuli was the slowest when neutral emotions were presented (average: 1228 ± 193 msec). Additionally, reaction time for sentence stimuli was significantly delayed with the

Table 4 ANOVA results from the emotion task

	Degree of freedom	F	p value		η^2_p
Main effect					
Diagnosis	1	7.244	0.013	*	0.248
Stimulus	1	12.958	0.002	**	0.371
Emotion	2	1.203	0.310		0.052
3-way interaction					
Stimulus × diagnosis × emotion	2	0.148	0.840		0.007
2-way interaction					
Stimulus × diagnosis	1	0.038	0.847		0.002
Emotion × diagnosis	2	0.146	0.865		0.007
Stimulus × emotion	1.599	7.108	0.003	**	0.244

* < 0.05 ** < 0.01

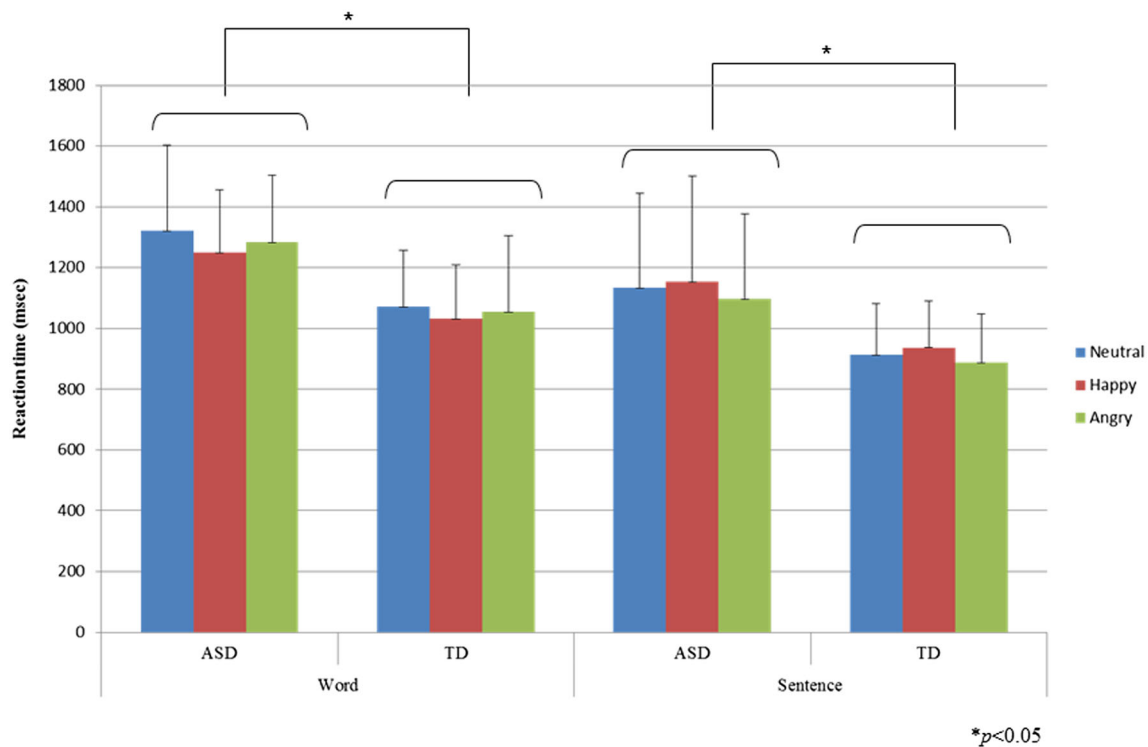


Fig. 2 Group differences in reaction time for both word and sentence sessions in the emotion task. The averaged reaction times to all emotions

for each session and group are displayed. *ASD* autism spectrum disorder; *TD* typical development

presentation of angry emotions compared to happy emotions (average: 1581 ± 641 msec).

In contrast to the emotion task, the simple main effect for stimulus ($p = 0.002$, $\eta^2_p = 0.356$) demonstrated that the reaction time was longer for sentences compared to words.

Discussion

The aims of the present study were to clarify the characteristics of auditory information processing mechanisms in individuals with ASD in order to expand our understanding of individuals with ASD and to contribute to improvements in their social adaptability. By focusing on reaction time, the present study was able to clarify the difficulties that individuals with ASD

experience in the perceptual processing of emotional tone, which previously yielded inconclusive study findings.

During the emotion task, we assessed differences in the abilities of individuals with ASD compared to TD in reading emotional tone during voice presentations of stimuli. In terms of accuracy alone, there were no differences between the ASD group and the TD group. However, when reaction time was analyzed, results indicated that individuals with ASD required a significantly longer time to read emotions, regardless of whether it was a short stimulus such as a word or a long and complex stimulus such as a sentence. Previously, Doyle-Thomas et al. (2013), Eigsti et al. (2012), Grossman et al. (2010), Heikkinen et al. (2010), and Jones et al. (2011) argued that individuals with ASD and TD were equipped with equal abilities to distinguish different emotions expressed in voices,

Table 5 Accuracy and reaction times during the semantic task

	Word session						Sentence session					
	Accuracy (%)			Reaction time (msec)			Accuracy (%)			Reaction time (msec)		
	Neutral	Happy	Angry	Neutral	Happy	Angry	Neutral	Happy	Angry	Neutral	Happy	Angry
ASD	100	100	100	1264	1230	1211	98	97	95	1816	1762	1881
(SD)	(0)	(0)	(0)	(196)	(196)	(199)	(4)	(4)	(4)	(645)	(709)	(733)
TD	99	100	99	1192	1137	1132	99	98	99	1296	1268	1281
(SD)	(2)	(0)	(2)	(191)	(183)	(204)	(2)	(3)	(2)	(396)	(341)	(357)

Values represent mean (SD). *SD* standard deviation, *ASD* autism spectrum disorder, *TD* typical development

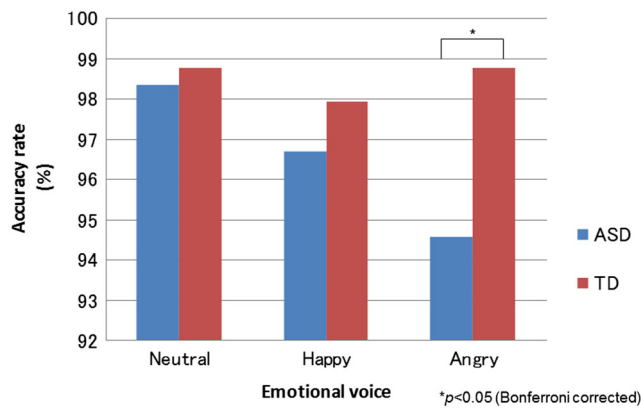


Fig. 3 Group differences in accuracy to the valence of each emotional voice during the semantic task. *ASD* autism spectrum disorder, *TD* typical development

while Golan et al. (2015), Hobson (1986), Lindner and Rosen (2006), Mazefsky and Oswald (2007), and Philip et al. (2010) disproved the ability of individuals with ASD to distinguish among different emotional tones, indicating contradictory results (as described in the “Introduction”). Our results support the claims of Golan et al. up until the time of response. However, it is necessary to verify the possibility that the difference in reaction time reflects the general delay in reaction by those with ASD or that such difference is due to the effects of IQ. Because reaction time was not different between groups in the word session of the semantic task in the present study and because reaction times were not different between the ASD group and TD group in a meta-analysis of 32 studies on reaction time (Ferraro 2014), we reasoned that it is unlikely that the results we observed were due to a general delay in motor responses in individuals with ASD. However, PSI in the ASD group was significantly lower than that in the TD group, which is known as a general characteristic of ASD (Oliveras-Rentas et al. 2012). Also, the mean FSIQ was considerably lower in the ASD group with a relatively large effect size (Cohen’s $d=0.82$), although the difference was not statistically significant ($p=0.057$). Therefore, it is necessary to

more carefully assess the effects of IQ. We conducted ANCOVA analysis using FSIQ, VCI, PRI, and PSI as covariates in order to adjust for the effects of IQ. Results verified that the main effect of diagnosis, with regards to the reaction time of emotion recognition in the emotion task, was statistically significant even after adjusting for FSIQ ($p=0.038$, $\eta^2_p=0.188$), VCI ($p=0.016$, $\eta^2_p=0.247$), and PRI ($p=0.029$, $\eta^2_p=0.208$), with the exception of PSI, which did not reach significance ($p=0.078$, $\eta^2_p=0.141$). To further elucidate the effects of PSI on reaction time, we conducted a multiple regression analysis with reaction time as the dependent variable, and PSI and diagnosis for predictors and found that the model was significant ($p=0.048$) with $R^2=0.251$. In this model, the effect of PSI was much lower (standardized beta= -0.075 , $p=0.759$, partial $r=-0.068$) compared with that of diagnosis (standardized beta= 0.450 , $p=0.078$, partial $r=0.375$). These results suggest that the slower reaction time of the ASD group in the emotion task observed in this study was independent of PSI. In addition, we analyzed the correlation between PARS scores and reaction time or PSI within the ASD group. The correlation coefficient to reaction time with PARS ($r=0.484$) was larger than that to PSI with PARS ($r=-0.235$), although neither value as statistically significant. Further, there was a significant correlation between reaction time for the sentence session in the emotion task and PARS scores ($r=0.636$, $p=0.026$). Therefore, it is possible that the longer reaction time needed to recognize emotions from voices in the ASD group was not simply attributed to low PSI, but rather reflected difficulty in social interaction.

In the semantic task, we examined the effects of emotional voice presentation on perceptual processing mechanisms. Notably, our results demonstrated that accuracy was significantly lower in the ASD group compared to the TD group when angry emotions were presented during the sentence session. These findings suggested that when a voice with an angry emotion was presented, individuals with ASD became confused regarding its semantic perception. Consequently, semantic information was not as accurately conveyed to individuals with

Table 6 ANOVA results from the semantic task

	Degree of freedom	F	<i>p</i> value		η^2_p
Main effect					
Diagnosis	1	5.330	0.031	*	0.195
Stimulus	1	12.168	0.002	**	0.356
Emotion	2	4.378	0.018	*	0.166
3-way interaction					
Stimulus × diagnosis × emotion	2	2.269	0.115		0.093
2-way interaction					
Stimulus × Diagnosis	1	5.005	0.036	*	0.185
Emotion × Diagnosis	2	1.562	0.221		0.066
Stimulus × Emotion	2	5.285	0.009	**	0.194

* < 0.05 ** < 0.01

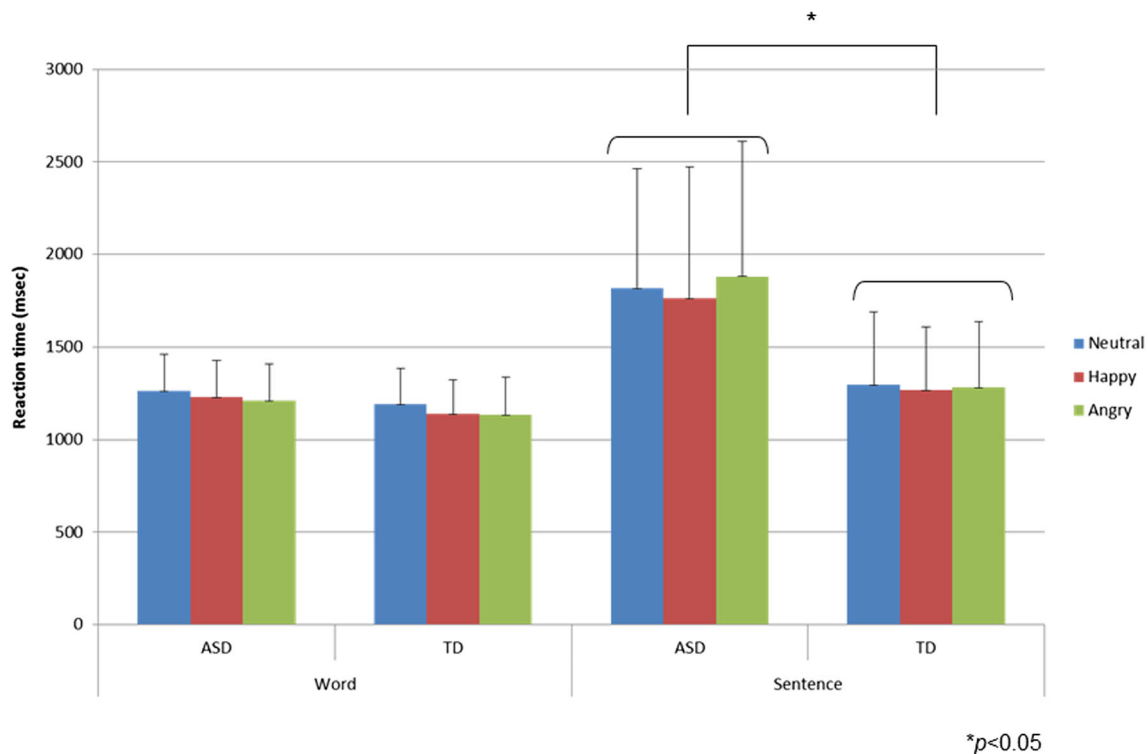


Fig. 4 Group differences in reaction time for the word and sentence sessions during the semantic task. *ASD* autism spectrum disorder, *TD* typical development

ASD compared to the TD group. Individuals with ASD are widely known to experience discomfort with specific pitch ranges; Williams (1992) stated in her book that when individuals with ASD encounter a high-pitched voice or sound, they experience unbearable pain. Grandin and Scariano (1986) also reported a similar view in her work. It is possible that the pitch of the vocal stimuli also affected our results. We therefore determined the first formant of all words used in the voice stimuli sets and analyzed the data by emotion. We found that the primary frequencies (mean) of each emotionally toned voice were as follows: 436 ± 58 Hz for happy, 278 ± 61 Hz for neutral, and 297 ± 31 Hz for angry, indicating that “happy” had a significantly higher pitch than the other two emotions ($p < 0.001$) and that “angry” and “neutral” were not significantly different ($p = 0.395$). Consequently, it was suggested that it is unlikely that the present results were due to a simple difference in the pitch of the voice, but that the angry toned voice itself had affected the semantic recognition process in the ASD group. Van del Kolk (1994) reported that 20 % of post-traumatic stress disorder participants exhibited a flashback of a traumatic experience when they were presented with acoustic startle stimuli; it is therefore possible that the angry voice affected the memory system of individuals with ASD who are often scolded on a daily basis, consequently confusing their semantic recognition function. Eigsti et al. (2012) presented affirmative and interrogative sentences with either angry or neutral emotion to 16 children with ASD (mean age = 13.7 years) and 11 with TD (mean age = 13.7 years) and asked them to determine the emotion,

sentence structure, and whether or not the semantic content was about a living creature. They reported that there were no differences in accuracy between the two groups in any of the tasks. Although those results are contradictory to our findings, the difference may be attributed to the fact that Eigsti et al. study only used two types of emotions, angry and neutral. Eigsti et al. also found in the same report that brain activity was detected in a broader area in the ASD group through functional magnetic resonance imaging conducted during the presentation of emotionally toned voice. This indicates that greater brain resources are required for individuals with ASD to recognize emotionally toned voices, suggesting that their information processing mechanism is different from individuals with TD. In the analysis of reaction time for sentence stimulus, both ASD and TD groups had different speeds of semantic understanding depending on the emotion. For example, expressing a happy emotion led to the fastest conveyance of semantics while expressing an angry emotion led to the slowest conveyance of semantics. This suggested that the children’s semantic perception processing mechanism is affected by the emotion expressed in voices, regardless of their diagnosis. In addition, although differences between groups were not observed in the overall reaction times for the word sessions, the reaction times of individuals in the ASD group were significantly prolonged in the sentence session, which consisted of more complex semantics. To assess the effects of IQ on this difference in reaction time, we performed ANCOVA similarly to the emotion task described above. The results showed that, even after adjusting for IQ, differences

between groups were observed in the sentence session alone (FSIQ: $p=0.048$, $\eta^2_p=0.173$, VCI: $p=0.037$, $\eta^2_p=0.192$, PRI: $p=0.039$, $\eta^2_p=0.187$), with the exception of PSI, which did not reach significance ($p=0.054$, $\eta^2_p=0.166$). Similar to the emotion task, an additional multiple regression analysis showed that the effect of PSI was much smaller than that of disease (standardized beta = 0.088 and 0.509, $p=0.728$ and 0.054, partial $r=-0.077$ and 0.407, respectively). This observation is consistent with the neural complexity hypothesis described by Bertone et al. (2005), which states that a greater complexity in perceptual processing, due to an increased amount of information, leads to confusion in individuals with ASD during perceptual processing in visual processing. Additionally, this finding can also be explained by the weak central coherence theory (Frith 1989), as individuals with ASD tend to put too much thought into the details of a sentence, leading to difficulties in ascertaining the overall picture.

The neuroanatomical correlate of the difficulty in emotional recognition in children with ASD is still not fully understood. Several studies have suggested that anatomical or functional abnormalities in the amygdala of individuals with ASD are possible causes of this symptom (Baron-Cohen et al. 2000; Kim et al. 2010; Nacewicz et al. 2006). However, several studies have also reported that abnormalities in the amygdala cannot explain this symptom (Amaral, Bauman, and Schumann 2003; Dziobek et al. 2006). Further research is needed to understand the neuronal basis of the difficulty in emotional recognition in individuals with ASD.

Limitations of This Study

There are several significant limitations to this study. (1) The sample size was small. We determined η^2_p and verified that the effect size was statistically adequate. Nonetheless, further investigation with a greater sample size is desirable. (2) The mean accuracy was $\geq 94.6\%$ for all tasks, and the sensitivity of accuracy evaluation may not be sufficient due to a ceiling effect. (3) In the present study, we were not able to use ADI-R and the Autism Diagnostic Observation Schedule (Lord et al. 2000), which are standard instruments for the confirmatory diagnosis of ASD, as their usage for research purposes is restricted within Japan. Also, the participants were not administered the same IQ test, as both WISC-III or WISC-IV were used. (4) Regarding the stimuli, we only used “happy,” a representative positive emotion, “angry,” a notable example of a negative emotion, and “neutral,” which is in between the former two emotions. If too many types of stimuli were used, the necessary number of stimuli would also increase, resulting in a greater burden on the participants. We therefore limited the number of stimuli to three types. However, it will be necessary to conduct further investigations with other emotions in the future. (5) The study was conducted with native Japanese children, and one may argue about the generalizability of the findings because Japanese

language, especially for its writing system, is very different from western languages. However, we consider the effect to be minimal because all of the stimuli consisted of pictures and voices.

Future Directions

In order to understand the social difficulties of individuals with ASD more in detail, comprehensive studies involving assessments of behavioral, neuroanatomical, genetic, and environmental factors should be conducted. Such studies are expected to advance behavioral therapy and educational programs for individuals with ASD.

Conclusion

Our experimental results suggested that individuals with ASD possess a different information processing mechanism from those with TD to determine different emotions that are presented vocally. Additionally, it was suggested that angry toned voices may affect semantic recognition. Based on our findings, we believe that it is necessary to pay close attention to the emotion and amount of information contained in words, in addition to understanding the characteristics of ASD in perceiving emotional voices, when communicating with individuals with ASD.

Compliance with ethical standards

Conflict of Interest The authors declare that they have no conflict of interest.

No funding was received for this study.

Ethical approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the institutional and national research committee and with the 1964 Helsinki declaration and its later amendments or comparable ethical standards. This study was approved by the first author’s institutional ethics committee. After detailed explanations of the study were provided, written informed consent was obtained from all individual participants included in the study and their guardians.

References

- Amano, S., & Kondo, T. (1999). *Nihongo-no Goi-Tokusei (Lexical properties of Japanese)*. Tokyo: Sanseido.
- Amano, S., & Kondo, T. (2000). On the NTT psycholinguistic databases “Lexical Properties of Japanese”. *Journal of the Phonetic Society of Japan*, 4(2), 44–50.
- Amaral, D. G., Bauman, M. D., & Schumann, C. M. (2003). The amygdala and autism: implications from non-human primate studies. *Genes Brain Behavioral*, 2(5), 295–302.
- American Psychiatric Association. (1994). *Diagnostic and statistical manual of mental disorders* (4th ed.). Washington, DC: APA Press.
- American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). Washington, DC: APA Press.

- Autism Society Japan. (2008). *Pervasive Developmental Disorders Autism Society Japan Rating Scale*. Tokyo: Spectrum Publishing Company.
- Baron-Cohen, S. (2004). *The Essential Difference: Male and Female Brains and the Truth about Autism*. London: Penguin.
- Baron-Cohen, S., Ring, H. A., Bullmore, E. T., Wheelwright, S., Ashwin, C., & Williams, S. C. R. (2000). The amygdala theory of autism. *Neuroscience and Biobehavioural Reviews*, 24, 355–364.
- Bertone, A., Mottron, L., Jelenic, P., & Faubert, J. (2005). Enhanced and diminished visuo-spatial information processing in autism depends on stimulus complexity. *Brain*, 128(Pt 10), 2430–2441.
- Birdwhistell, R. L. (1970). *Kinesics and context-Essays on body-motion*. Philadelphia: University of Pennsylvania Press.
- Corbett, B. A., Qualls, L. R., Valencia, B., Fecteau, S. M., & Swain, D. M. (2014). Peer-mediated theatrical engagement for improving reciprocal social interaction in autism spectrum disorder. *Frontiers in Pediatrics*, 2, 110.
- de Vries, M., & Geurts, H. M. (2012). Cognitive Flexibility in ASD; Task Switching with Emotional Faces. *Journal of Autism and Developmental Disorders*, 42(12), 2558–2568.
- Doyle-Thomas, K. A., Goldberg, J., Szatmari, P., & Hall, G. B. (2013). Neurofunctional underpinnings of audiovisual emotion processing in teens with autism spectrum disorders. *Frontiers in Psychiatry*, 4, 48.
- Dziobek, I., Fleck, S., Rogers, K., Wolf, O. T., & Convit, A. (2006). The ‘amygdala theory of autism’ revisited: linking structure to behavior. *Neuropsychologia*, 44(10), 1891–1899.
- Eigsti, I. M., Schuh, J., Mencl, E., Schultz, R. T., & Paul, R. (2012). The neural underpinnings of prosody in autism. *Child Neuropsychology*, 18(6), 600–617.
- Ferraro, F. R. (2014). No evidence of reaction time slowing in autism spectrum disorder. *Autism*, 20(1), 116–122.
- Fink, E., de Rosnay, M., Wierda, M., Koot, H. M., & Begeer, S. (2014). Brief report: Accuracy and response time for the recognition of facial emotions in a large sample of children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 44(9), 2363–2368.
- Frith, U. (1989). *Autism: Explaining the Enigma*. Oxford: Blackwell.
- Gernsbacher, M. A., Stevenson, J. L., Khandakar, S., & Goldsmith, H. H. (2008). Why Does Joint Attention Look Atypical in Autism? *Society for Research in Child Development*, 2(1), 38–45.
- Golan, O., Sinai-Gavrilov, Y., & Baron-Cohen, S. (2015). The Cambridge Mindreading Face-voice Battery for Children (CAM-C): Complex emotion recognition in children with and without autism spectrum conditions. *Molecular Autism*, 6, 22-015-0018-z. eCollection 2015.
- Grandin, T., & Scariano, M. (1986). *Emergence, labeled autistic*. Novato, CA: Arena Press.
- Grossman, R. B., Bemis, R. H., Plesa-Skwerer, D., & Tager-Flusberg, H. (2010). Lexical and affective prosody in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 53(3), 778–793.
- Heikkinen, J., Jansson-Verkasalo, E., Toivanen, J., Suominen, K., Vayrynen, E., Moilanen, I., & Seppänen, T. (2010). Perception of basic emotions from speech prosody in adolescents with Asperger’s syndrome. *Logopedics, Phoniatrics, Vocology*, 35(3), 113–120.
- Hobson, R. P. (1986). The autistic child’s appraisal of expressions of emotion. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 27(3), 321–342.
- Hobson, R. P., Ouston, J., & Lee, A. (1988). Emotion recognition in autism: Coordinating faces and voices. *Psychological Medicine*, 18(4), 911–923.
- Ito, H., Tani, I., Yukihiro, R., Adachi, J., Hara, K., Ogasawara, M., & Tsujii, M. (2012). Validation of an interview-based rating scale developed in Japan for pervasive developmental disorders. *Research in Autism Spectrum Disorders*, 6(4), 1265–1272.
- Jones, C. R., Pickles, A., Falcaro, M., Marsden, A. J., Happe, F., Scott, S. K., & Charman, T. (2011). A multimodal approach to emotion recognition ability in autism spectrum disorders. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, 52(3), 275–285.
- Kim, J. E., Lyoo, I. K., Estes, A. M., Renshaw, P. F., Shaw, D. W., Friedman, S. D., & Dager, S. R. (2010). Laterobasal amygdalar enlargement in 6- to 7-year-old children with autism spectrum disorder. *Archives of General Psychiatry*, 67(11), 1187–1197.
- Lai, M. C., Lombardo, M. V., & Baron-Cohen, S. (2014). Autism. *Lancet*, 8, 383(9920), 896–910.
- Lartseva, A., Dijkstra, T., & Buitelaar, J. K. (2015). Emotional language processing in autism spectrum disorders: A systematic review. *Frontiers in Human Neuroscience*, 8, 991.
- Lindner, J. L., & Rosen, L. A. (2006). Decoding of emotion through facial expression, prosody and verbal content in children and adolescents with Asperger’s syndrome. *Journal of Autism and Developmental Disorders*, 36(6), 769–777.
- Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism Diagnostic Interview-Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5), 659–685.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Jr., Leventhal, B. L., DiLavore, P. C., Pickles, A., & 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.
- Mazefsky, C. A., & Oswald, D. P. (2007). Emotion perception in Asperger’s syndrome and high-functioning autism: the importance of diagnostic criteria and cue intensity. *Journal of Autism and Developmental Disorders*, 37(6), 1086–1095.
- Nacewicz, B. M., Dalton, K. M., Johnstone, T., Long, M. T., McAuliff, E. M., Oakes, T. R., Alexander, A. L., & Davidson, R. J. (2006). Amygdala volume and nonverbal social impairment in adolescent and adult males with autism. *Archives of General Psychiatry*, 63(12), 1417–1428.
- O’Connor, K. (2012). Auditory processing in autism spectrum disorder: a review. *Neuroscience and Biobehavioral Reviews*, 36(2), 836–854.
- Oliveras-Rentas, R. E., Kenworthy, L., Roberson, R. B., Martin, A., & Wallace, G. L. (2012). WISC-IV Profile in High-Functioning Autism Spectrum Disorders: Impaired Processing Speed is Associated with Increased Autism Communication Symptoms and Decreased Adaptive Communication Abilities. *Journal of Autism and Developmental Disorders*, 41(1), 102–109.
- Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 35(2), 205–220.
- Peppé, S., McCann, J., Gibbon, F., O’Hare, A., & Rutherford, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 50(4), 1015–1028.
- Philip, R. C., Whalley, H. C., Stanfield, A. C., Sprengelmeyer, R., Santos, I. M., Young, A. W., & Hall, J. (2010). Deficits in facial, body movement and vocal emotional processing in autism spectrum disorders. *Psychological Medicine*, 42(5), 655–664.
- Rump, K. M., Giovannelli, J. L., Minshev, N. J., & Strauss, M. S. (2009). The development of emotion recognition in individuals with autism. *Child Development*, 80(5), 1434–1447.
- Tracy, J. L., Robins, R. W., Schriber, R. A., & Solomon, M. (2011). Is emotion recognition impaired in individuals with autism spectrum disorders? *Journal of Autism and Developmental Disorders*, 41(1), 102–109.
- van der Kolk, B. A. (1994). The body keeps the score: Memory and the evolving psychobiology of posttraumatic stress. *Harvard Review of Psychiatry*, 1(5), 253–265.
- Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children* (3rd ed.). San Antonio, TX: The Psychological Corporation.
- Wechsler, D. (2003). *Manual for the Wechsler Intelligence Scale for Children* (4th ed.). London: Pearson Assessment.
- Williams, D. (1992). *Nobody nowhere*. London: Jessica Kingsley Publishers.