frontiers | Frontiers in Communication

TYPE Original Research
PUBLISHED 08 November 2024
DOI 10.3389/fcomm.2024.1483135

OPEN ACCESS

EDITED BY
Martina Micai,
National Institute of Health (ISS), Italy

REVIEWED BY
Helene Kreysa,
Friedrich Schiller University Jena, Germany
Catherine Caldwell-Harris,
Boston University, United States

*CORRESPONDENCE
Martine Grice
✉ martine.grice@uni-koeln.de

†PRESENT ADDRESS
Francesco Cangemi,
International Education Division, Center for Global Education, The University of Tokyo, Bunkyō, Japan

‡These authors have contributed equally to this work and share senior authorship

RECEIVED 19 August 2024
ACCEPTED 22 October 2024
PUBLISHED 08 November 2024

CITATION
Zimmermann JT, Ellison TM, Cangemi F, Wehrle S, Vogeley K and Grice M (2024) Lookers and listeners on the autism spectrum: the roles of gaze duration and pitch height in inferring mental states. *Front. Commun.* 9:1483135. doi: 10.3389/fcomm.2024.1483135

Lookers and listeners on the autism spectrum: the roles of gaze duration and pitch height in inferring mental states

Juliane T. Zimmermann[1], T. Mark Ellison[2], Francesco Cangemi[2†], Simon Wehrle[2], Kai Vogeley[1,3‡] and Martine Grice[2*‡]

[1]Department of Psychiatry, Faculty of Medicine and University Hospital Cologne, University of Cologne, Cologne, Germany, [2]IfL – Phonetics, University of Cologne, Cologne, Germany, [3]Institute of Neuroscience and Medicine, Division of Cognitive Neuroscience (INM-3), Research Centre Juelich, Juelich, Germany

Although mentalizing abilities in autistic adults without intelligence deficits are similar to those of control participants in tasks relying on verbal information, they are dissimilar in tasks relying on non-verbal information. The current study aims to investigate mentalizing behavior in autism in a paradigm involving two important nonverbal means to communicate mental states: eye gaze and speech intonation. In an eye-tracking experiment, participants with ASD and a control group watched videos showing a virtual character gazing at objects while an utterance was presented auditorily. We varied the virtual character's gaze duration toward the object (600 or 1800 ms) and the height of the pitch peak on the accented syllable of the word denoting the object. Pitch height on the accented syllable was varied by 45 Hz, leading to high or low prosodic emphasis. Participants were asked to rate the importance of the given object for the virtual character. At the end of the experiment, we assessed how well participants recognized the objects they were presented with in a recognition task. Both longer gaze duration and higher pitch height increased the importance ratings of the object for the virtual character overall. Compared to the control group, ratings of the autistic group were lower for short gaze, but higher when gaze was long but pitch was low. Regardless of an ASD diagnosis, participants clustered into three behaviorally different subgroups, representing individuals whose ratings were influenced (1) predominantly by gaze duration, (2) predominantly by pitch height, or (3) by neither, accordingly labelled "Lookers," "Listeners" and "Neithers" in our study. "Lookers" spent more time fixating the virtual character's eye region than "Listeners," while both "Listeners" and "Neithers" spent more time fixating the object than "Lookers." Object recognition was independent of the virtual character's gaze duration towards the object and pitch height. It was also independent of an ASD diagnosis. Our results show that gaze duration and intonation are effectively used by autistic persons for inferring the importance of an object for a virtual character. Notably, compared to the control group, autistic participants were influenced more strongly by gaze duration than by pitch height.

KEYWORDS

autism spectrum disorder, theory of mind, pitch height, eye gaze duration, intonation, perception and gaze behavior, mentalizing, adult

1 Introduction

Autism spectrum disorder (ASD) is characterized by difficulties in social communication and interaction (American Psychiatric Association, 2013). These difficulties might in part be explained by impaired perspective-taking or mentalizing skills (Baron-Cohen et al., 1985; Frith et al., 1991; Frith and Frith, 2006). However, adults with autism without intelligence deficits perform similarly to control participants in mentalizing tasks – inferring mental states of others – that strongly rely on verbal abilities (Bowler, 1992; Happé, 1994; Scheeren et al., 2013; Gernsbacher and Yergeau, 2019), whereas they show difficulties in non-verbal mentalizing tasks (*cf.* Baron-Cohen et al., 2001a; Baron-Cohen et al., 2001b; Ponnet et al., 2004; Dziobek et al., 2006; White et al., 2011), for example, when inferring mental states of people depicted in videos of social interactions (Ponnet et al., 2004; Dziobek et al., 2006). Accordingly, autistic adults tend to rely on verbal information (e.g., the words spoken) more than on non-verbal information (e.g., the body language accompanying the words and the way they are spoken) (Kuzmanovic et al., 2011; Stewart et al., 2013). However, the interplay between non-verbal modalities has not been studied systematically in this context. For the current study, we will focus on the interplay of two powerful means to communicate nonverbally in face-to-face interactions: eye gaze and intonation.

In human communication as well as in the communication between humans and virtual characters, eye gaze can be very informative, as it is closely linked to attention: people tend to look at objects (Buswell, 1935; Yarbus, 1967; DeAngelus and Pelz, 2009) or locations they attend to (Ferreira et al., 2008; Theeuwes et al., 2009). The relevance (Klami et al., 2008; Klami, 2010) of and the preference for an object (Shimojo et al., 2003; Chuk et al., 2016) is indicated by the time one spends looking at the object. This implies that another person's gaze behavior is key to inferring their intentions or attentional state (Baron-Cohen et al., 1995; Lee et al., 1998; Freire et al., 2004; Einav and Hood, 2006; Jording et al., 2019a, 2019b). Observing another person gazing towards an object in their environment can re-direct the observer's attention and increase the duration the observer spends looking at the respective object themselves (Freeth et al., 2010). However, adults with autism tend to have difficulties inferring emotions and mental states based on another person's eye region (Hobson et al., 1988; Baron-Cohen et al., 1997; Baron-Cohen et al., 2001a). They look at gaze-indicated objects less often (Wang et al., 2015) and tend to spend less time fixating those objects (Fletcher-Watson et al., 2009; Freeth et al., 2010). One explanation for this could be reduced attention towards gaze cues in individuals on the autism spectrum (Itier et al., 2007). Certainly, overt attention towards social stimuli in general is reduced in persons with autism (Chita-Tegmark, 2016), who tend to fixate the eye region for a shorter amount of time than control participants (Setien-Ramos et al., 2022), while differences for other parts of the face are less pronounced (Klin et al., 2002; Pelphrey et al., 2002; Dalton et al., 2005; Nakano et al., 2010; Auyeung et al., 2015). Irrespective of an ASD diagnosis, the time spent fixating a person's eyes is linked to the observer's mentalizing abilities (Müller et al., 2016). In autism, a decreased fixation duration on the eye region is associated with impaired social functioning and increased autism symptom severity (Riddiford et al., 2022). However, attention towards social stimuli in autism is dependent on the stimulus at hand (Guillon et al., 2014; Chita-Tegmark, 2016), and eye gaze behavior is influenced by the experimental task and task instructions (Del Bianco et al., 2018; Setien-Ramos et al., 2022). In classical false-belief tasks, which test the ability of an observer to understand that other people can believe things which the observer knows to be untrue (most famously the "Sally-Anne" test), eye gaze behavior can indicate impaired mentalizing in autism (Senju et al., 2009; Schneider et al., 2013; Schuwerk et al., 2015). By including eye-tracking in our study, we aimed to investigate the influence of an ASD diagnosis in combination with behavioral differences on participants' gaze behavior.

Prosody—referring to the non-verbal aspects of speech—is an important aspect of spoken language, as it adds an additional layer of information to the verbal content of an utterance, and can significantly change the meaning, and consequently the interpretation, of what is being said. This is important, for example, when deciphering emotions. Most prosodically expressed basic emotions, such as fear or sadness, can be recognized well by persons with autism (O'Connor, 2012; Stewart et al., 2013; Ben-David et al., 2020; Zhang et al., 2022). However, the identification of prosodically expressed emotions that are complex, such as curiosity or concern (Kleinman et al., 2001; Rutherford et al., 2002; Golan et al., 2007; Hesling et al., 2010; Rosenblau et al., 2017), or low-intensity (Globerson et al., 2015) has been reported to be impaired in autistic adults, possibly due to difficulties with the perception and interpretation of vocal pitch modulation (how the speech melody is changed) during speech (Schelinski et al., 2017; Schelinski and von Kriegstein, 2019; see Grice et al., 2023 for a review). Moreover, the imitation of vocal pitch can also be impaired in autistic adults (Wang et al., 2021).

Aspects of conversation that are important, new, or in focus are often highlighted prosodically by the speaker. In German, this can be achieved through pitch accent placement and type, cued *inter alia* by fundamental frequency, which is perceived as pitch height (Grice and Baumann, 2007; Féry and Kügler, 2008). The raising of pitch conveys prosodic prominence and importance for the listener (Arnold et al., 2013; Baumann and Winter, 2018). Autistic listeners have been reported to take pitch accents into account to a lesser extent than control persons when judging the givenness of a word, i.e., judging whether the object it denotes is known to the interlocutors in a given context or has not been previously introduced (Grice et al., 2016). Findings from the general population show that an attenuated sensitivity to pitch accent types is associated with poor pragmatic skills, i.e., the appropriate use of language in social situations (Bishop, 2016; Hurley and Bishop, 2016; Bishop et al., 2020).

Analogously to gaze, prosody (and pitch accents in particular) can function as a deictic cue (referring or "pointing" to an entity) and orient a listener's attention (Dahan et al., 2002; Weber et al., 2006; Ito and Speer, 2008; Watson et al., 2008). Studying overt attention in children with autism, Ito et al. (2022) found that, although the autistic group responded relatively slowly and weakly to a target word denoting an object, both the control group and the autistic group looked at the respective object longer if the referring utterance received an emphatic pitch accent (i.e., it was produced with longer duration and higher pitch). This demonstrates that autistic children can shift overt attention towards an important object in their environment. No comparable study has been performed with autistic adults to date.

In a previous web-based study (Zimmermann et al., 2020), we showed that both gaze duration and pitch height are used as cues by the general population when interpreting how a virtual character

conveys the importance of an object being referred to. In that study, participants rated objects as having greater importance for the virtual character both when the character looked at the object for a longer period of time (as opposed to a shorter period of time) as well as when she produced the word referring to it with higher vocal pitch (as opposed to lower pitch). Based on the tendency of participants to take into account only one of the two cues, we subdivided the sample into three behavioral clusters: (i) "Lookers," who based their ratings primarily on gaze duration, (ii) "Listeners," who based their ratings primarily on pitch height, and (iii) a group of "Neithers," who did not predominantly base their ratings on either cue.

Continuing this line of work on the influence of gaze duration and pitch height, the present study is a lab-based experiment investigating not only participants' responses but also their eye gaze fixation durations using a desk-mounted eye-tracker. We carried out a comparative analysis of participants with and without a diagnosis of ASD. In particular, we investigated whether similar behavioral patterns can be found in both groups. We hypothesized that the autistic group would rely on the gaze and pitch cues to a lesser extent, based on reports of difficulties in autism with using social gaze and intonation as cues for mentalizing (as summarized above). We also expected this to be reflected in the participants' own gaze behavior. Additionally, we examined how participants' gaze behavior, the character's gaze and pitch cues, as well as the presence of an ASD diagnosis affected performance in a memory task involving recognition of the objects used as visual stimuli (i.e., participants had to indicate whether an object had been or had not been present in the previous part of the experiment).

2 Materials and methods

2.1 Participants

For the autistic group, we recruited 24 monolingual German native speakers within an age range from 18 to 55 who had been diagnosed with Asperger syndrome (ICD-10 identifier: F.84.5) or with childhood autism (ICD-10 identifier: F.84.0) by the outpatient clinic for autism in adulthood or by the pediatric outpatient clinic for autism of the University Hospital Cologne. For the control group, we recruited 24 age-matched (within a range of 5 years) native German speakers. All participants of both groups had normal or corrected-to-normal vision as well as hearing. The study was conducted in accordance with the Declaration of Helsinki (World Medical Association, 2013) and approved by the Ethics Committee of the University Hospital Cologne.

To ensure that results were not influenced by lower cognitive performance, we only included participants with verbal and total intelligence scores of at least 85, as measured with the *WIE-III*, (Aster et al., 2006), with attentional scores greater 80, as measured with the *D2* (Brickenkamp, 2002), and for participants in the control group with maximally moderate depressive symptoms as measured with the *Beck Depression Inventory* (*BDI-II*, Beck et al., 1996), i.e., with *BDI-II* scores <18. Sample characteristics are provided in Table 1.

Verbal intelligence scores as measured with the *WIE* indicated average or above-average verbal intelligence for all participants (Table 1). Diagnostic groups did not differ significantly regarding the *WIE* verbal scores [two-samples *t*-test, $t(46) = -1.50$, $p = 0.140$] or the *WIE* performance scores [two-samples *t*-test, $t(46) = -1.73$, $p = 0.091$]. Scores indicating depression or depressive tendencies as measured with the *BDI* were significantly higher in participants with autism compared to the control group [Welch two-samples *t*-test, $t(32.03) = -4.25$, $p < 0.001$]. Attention scores measured with the *D2* tended to be somewhat higher in the autistic sample [two-samples *t*-test, $t(46) = -1.88$ $p = 0.066$].

2.2 Experimental design

We used a paradigm established in the previous web-based study referred to above (Zimmermann et al., 2020). The material and procedures were adjusted for the laboratory setting.

We tested the individual and combined influence of *gaze duration* of a virtual character towards an object and *pitch height* of an utterance on the rating of how important the object appeared to the virtual character. In addition, we obtained object memory scores by assessing recognition rates for all objects in a subsequent recognition task. To create a socially "plausible" and at the same time standardized situation, we presented videos of a virtual character's face positioned above an object. The object was different in each trial, and each object was only shown once. The movements performed by the virtual character were limited to the eyes. The character's attention towards the object, suggesting greater importance, was operationalized as longer *gaze duration* directed towards the object alongside an auditorily presented utterance characterized by a pitch accent with a fundamental frequency peak located on the stressed syllable of the target word. We systematically varied the factors *gaze duration* and *pitch height* on two levels. *Gaze duration* towards the object was either comparatively short (600 ms) or long (1800 ms). *Pitch height* on the accented syllable was either low or high, characterized by f0 peak

TABLE 1 Sample characteristics, general.

	Sex	Age	*WIE* IQ verbal	*WIE* IQ performance	*WIE* IQ total	*D2* total error corrected	*BDI-II*
ASD ($N = 24$)	13 men 10 women 1 not indicated	18–55 years $M = 39.4$ ($SD = 11.7$)	$M = 115.9$ ($SD = 14.1$)	$M = 110.7$ ($SD = 16.2$)	$M = 114.9$ ($SD = 14.6$)	$M = 105.6$ ($SD = 9.7$)	$M = 15.5$ ($SD = 10.8$)
Control ($N = 24$)	14 men 10 women	21–58 years $M = 38.9$ ($SD = 11.9$)	$M = 110.1$ ($SD = 12.5$)	$M = 103.2$ ($SD = 13.7$)	$M = 107.4$ ($SD = 12.1$)	$M = 100.5$ ($SD = 8.8$)	$M = 5.3$ ($SD = 4.9$)

WIE IQ, Hamburg-Wechsler-Intelligenz-Test für Erwachsene III (intelligence test for adults); D2, d2 Aufmerksamkeits-Belastungs-Test (attention load test); BDI-II, Beck Depression Inventory, 2nd Version (questionnaire on depressive symptom severity).

height, which was raised by 45 Hz in the respective high-pitch-height condition. Thus, we effectively created four conditions, establishing a 2 × 2 experimental design (Table 2).

2.3 Video material

Videos were created by combining images and sound material using *Python* and the *FFmpeg* module (FFmpeg Developers, 2018). The videos used in the rating task showed a female character's face positioned above the center of the screen (screen dimensions: 1,920 × 1,200 px). The face and its position were always the same during the entire experiment. One object was presented below the center of the screen (see Figure 1 for image positions and the time course of a single trial). The background color was white. At the beginning and the end of the video, the virtual character exhibited idle gaze behavior, i.e., she performed gaze movements in the direction of random locations in the environment, but neither fixated the object nor the participant during this phase. All images of the virtual character's face were taken from previous studies investigating the perception of gaze direction (Eckert, 2017; Jording et al., 2019a). The virtual character's face was created using Poser R (Poser 8, Smith Micro Software, Inc., Columbia, USA) using Python 2.4. For the idle gaze phases, we chose eight images of gaze directions that were diverted horizontally to the left or to the right as well as diverted slightly to the bottom. The choice of the female character was based on the decision to use a female speaker after pretesting for production of the auditory stimuli.

After 2.0 s of idle gaze, the virtual character made three fixations establishing a social situation: (1) looking at the participant for 1.0 s, (2) looking towards the object for either 0.6 (short gaze) or 1.8 s (long gaze), and (3) looking at the participant again for 1.0 s. The onset of the virtual character looking towards the object was also the onset of the auditory utterance. The durations of 0.6 and 1.8 s of the virtual character's gaze toward the object were chosen based on a previous study of human–robot interaction (Pfeiffer-Lessmann et al., 2012), where the durations of 0.6 s and 1.8 s were associated with different perceptions of the robot's intention to make the participant follow their gaze. Importantly, in that study, 1.8 s was the participants' own preferred gaze duration towards an object with the intention of making the robot follow their gaze.

This set of three gazing actions (looking at the participant, looking towards the object, and looking at the participant again) conveying communicative intent was both preceded and followed by a blink simulated by presenting an image of the virtual character's face with their eyes closed for 0.1 s to simulate naturalistic interblink-interval durations (Doughty, 2001). However, to make the character's blinking behavior appear less mechanical, the videos were created by randomly

TABLE 2 2 × 2 experimental design.

		Gaze duration	
		Short	Long
Pitch height	Low	Low pitch height and short gaze	Low pitch height and long gaze
	High	High pitch height and short gaze	High pitch height and long gaze

Variation of gaze duration and pitch height resulted in four conditions.

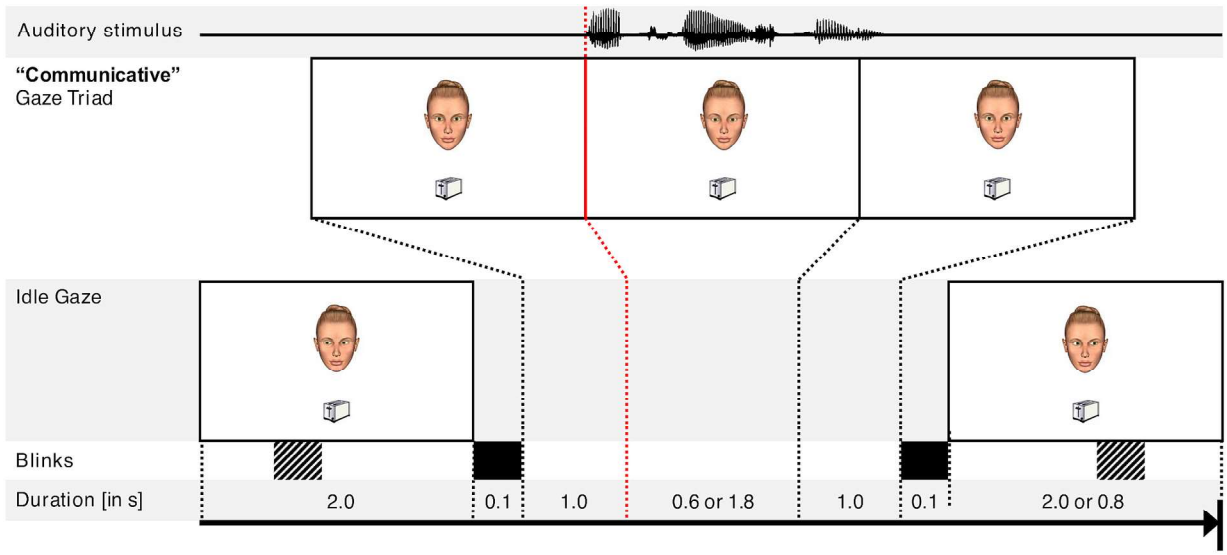

FIGURE 1
Schematic time course of a video depicting the different phases of an example trial. The object in this example is a toaster, the auditory stimulus is the utterance "der Toaster" (English: the toaster). Total duration of each video was 6.6 s. Blinks are demarcated as either fixed (solid black) or appearing at random (striped). The red line indicates the simultaneous onset of the auditory stimulus and the virtual character's gaze towards the object.

including either no blink or only one additional blink during the first and second idle (i.e., "non-communicative") phases. Following the "communicative" gaze triad, the virtual character continued gazing at random locations until the end of the video, i.e., for 2.0 s (short gaze conditions) or for 0.8 s (long gaze conditions) in order to keep the total presentation duration of the object constant in all videos.

On the basis of our experience with the web-based study (Zimmermann et al., 2020), we excluded 14 problematic items from the previous stimulus set. These exclusions resulted in a final set of 92 test items, with each participant observing 23 items per condition. Four of the discarded stimuli were used for practice trials in the current study, but did not enter analysis.

2.4 Object images

Object images used for video creation were selected from the set described in Rossion and Pourtois (2004). Images were selected based on the phonology of their referential German expressions (Genzel et al., 1995). To reduce any possible influence of the number of syllables on the perception of word prominence, only words with two syllables and initial stress were included in the subset, such as "Toaster," "Hammer," "Meißel," "Sofa" (respectively *toaster, hammer, chisel, sofa*). The full list of object names and their English translations can be found in the Supplementary material. Additionally, we partly excluded homonyms if the homonym-partner was present in the image set or if one homonym-partner was semantically clearly more salient (e.g., the German homonym "Mutter" is semantically more salient when referring to "mother" than to "nut" as the counterpart of a screw).

2.5 Auditory stimulus material

The auditory stimuli were produced by a trained female speaker, who uttered each of the 92 target phrases including the definite article (e.g., "der Toaster": *the toaster*) with an H*-accented rendition [following the categorization of German accent types by Grice et al., 2005]. The H* accent type has been found to be generic, and can be used for different focus types in German, namely broad focus, narrow focus and contrastive focus (Grice et al., 2017). Recordings took place in a soundproof booth, using an *AKG C420L* headset microphone connected to a computer running *Adobe Audition* via a USB audio interface (PreSonus AudioBox 22VSL). Stimuli were recorded with a sampling rate of 44,100 Hz, 16 bit. The resulting speech stimuli were normalized to equal loudness using *Myriad* (Aurchitect Audio Software, LLC, 2018). The editing was performed using *Praat* (Boersma and Weenink, 2018). Sound was faded in and out (Winn, 2014) to avoid any salient on- and offset of noise. Fundamental frequency (f0) contours were extracted, manually corrected, and smoothed according to an established procedure (Cangemi, 2015). The resulting pitch contours were stylized to a resolution of one semitone. These stylized versions were used directly as the audio stimuli for the low-pitch-height condition. The pitch contours of utterances for the high-pitch-height condition were resynthesized: *pitch height* maxima on the accented vowels were raised by 45 Hz. This difference between *pitch height* maxima for the different conditions was based on the individual production characteristics of the speaker for a subset of 15 words. These words were selected due to their ability to bear pitch (i.e., the amount of periodic energy, typically high in vowels and low in, e.g., fricatives and stops such as /f/ and /d/ respectively).

The speaker was asked to produce all utterances in two versions: (1) applying an H*-accent and (2) applying an L + H*-accent, the latter of the two resulting in a perceptually more strongly accented utterance which expresses greater prominence. We extracted the following measures for characterization of speaker-specific production parameters for utterances bearing an H*-accent and those bearing an L + H* -accent: onset and duration of the accented syllable, height of f0 contour peak on the accented vowel as well as the associated timepoint, height of f0 contour trough within the timeframe starting at voice onset and ending at f0 peak on the accented vowel as well as the associated timepoint. Timepoints for the onset of the accented syllable were set manually; for f0 peaks and troughs, they were set automatically and corrected manually. Subsequently, f0 peak alignment was calculated as the percentage of duration from accented syllable onset until the timepoint of the f0 contour peak in relation to the total duration of the accented syllable. F0 trough alignment was calculated analogously to f0 peak alignment. F0 onglide was calculated as the difference between f0 trough height and f0 peak height (Figure 2). We plotted the following parameters to examine to what degree they contributed to distinguishing between utterances bearing an H*-accent and those bearing an L + H*-accent in our speaker: (a) f0 peak and f0 peak alignment (Figure 3A), (b) f0 trough and f0 trough alignment (Figure 3B), (c) f0 onglide and the duration of the accented syllable (Figure 3C).

Visual inspection of production parameters showed that *pitch height* most reliably separated the two stimulus conditions for the selected speaker (Figure 3). *Pitch height* was on average 205.4 Hz ($SD = 0.65$) for the utterances produced with an H*-accent, and 251.1 Hz ($SD = 1.08$) for the utterances produced with an L + H*-accent. To mirror this difference, a positive adjustment of 45 Hz was chosen to simulate an otherwise comparably accented L + H*-like version of our stylized H*-accented utterances.

The resulting auditory stimuli were submitted to a perceptual pretest: The original H*-accented utterances and their stylized versions were rated by six trained phoneticians for "similarity." All stylized stimuli were rated for "naturalness" and accent type. Details on the pretest's methods and results as well as auditory and video stimuli can be found in the Supplementary material.

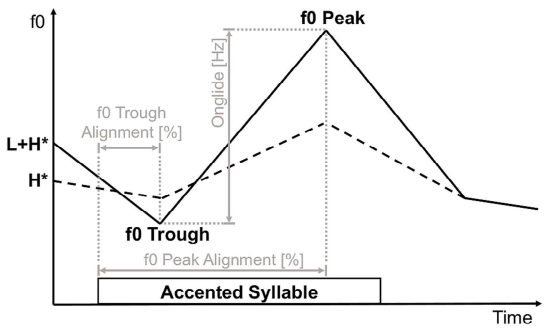

FIGURE 2
Schematic illustration of acoustic measures of an utterance. Extracted measures were onset and offset of the accented syllable, f0 contour peak on the accented vowel and f0 contour trough. Calculated measures are depicted in gray.

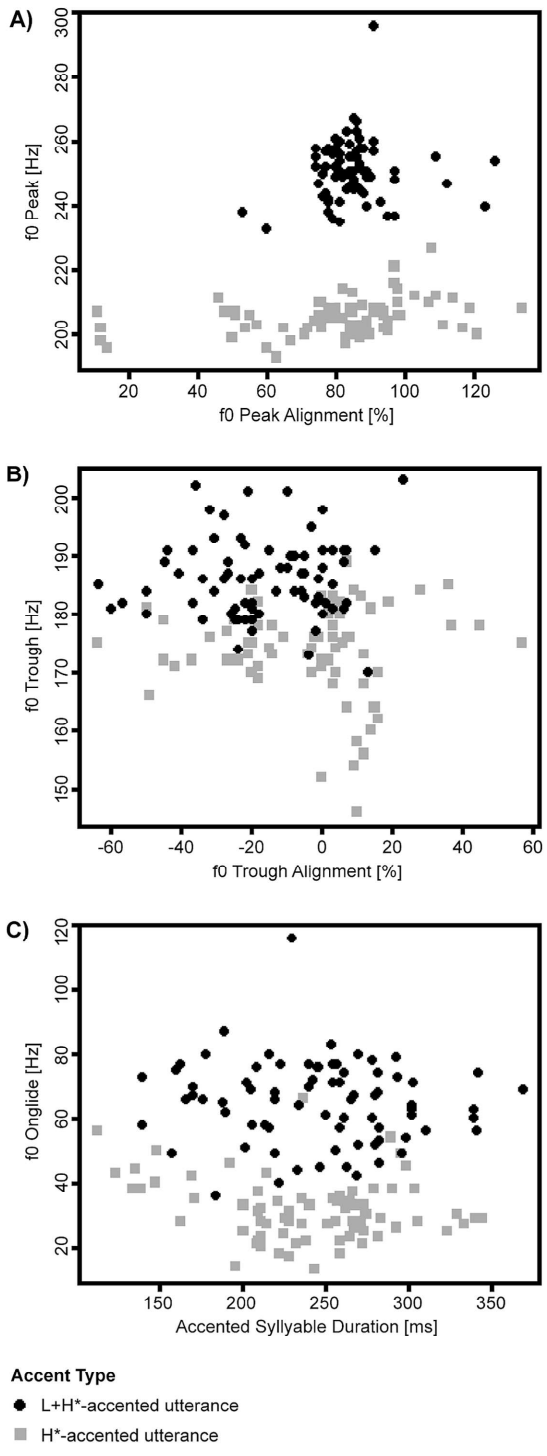

FIGURE 3
Acoustic parameters of the speaker's utterances: **(A)** F0 peak and f0 peak alignment, **(B)** f0 trough and f0 trough alignment, **(C)** f0 onglide and the duration of the accented syllable. Gray squares and black dots distinguish between the utterance produced with lesser (H*) and greater (L + H*) prominence.

2.6 Selection of distractor words for the recognition task

The 92 distractor words presented alongside the 92 target words in the recognition task were selected by identifying words of similar word frequency compared to the words we used in the rating task (Brysbaert et al., 2011). Since animacy has been reported to lead to better recognition (Leding, 2020), we included an equal number of animals in the list of distractor words and target words.

2.7 Psychological tests

To infer mentalizing abilities, we employed the *"Reading the Mind in the Eyes" test* (Baron-Cohen et al., 2001a; Baron-Cohen et al., 2001b), henceforth referred to as *Eyes-test*. For a proxy of sensory perception we included a German translation of the *Sensory Perception Quotient* (*SPQ*, Bierlich et al., 2024). As indicators for autistic traits, we included the *Autism Quotient* (*AQ*, Baron-Cohen et al., 2001b), the *Empathy Quotient* (*EQ*, Baron-Cohen and Wheelwright, 2004) and the *Systemizing Quotient* (*SQ*, Baron-Cohen et al., 2003).

2.8 Procedure

The study was conducted at the Department for Psychiatry of the University Hospital of Cologne. Participants provided informed consent and filled in the *AQ, EQ, SQ, SPQ* and a questionnaire on demographic data as well as information on (their history of) visual, auditory, psychological or speech impairments. Afterwards, they filled in the *BDI-II* and were tested with the *Eyes-test* and the *D2* (as described above). For the duration of the rating task, participants were seated in front of a desk-mounted eye-tracker. Head movement was minimized with the use of a fixed chin rest. They were instructed to imagine that the utterances they heard were produced by the character on screen and were informed that the character could convey the importance of the object. Participants were then instructed to answer the same question after each trial: "How important does the character find the depicted object?" (original German instruction: "Wie wichtig findet die Figur das abgebildete Objekt?"). Each trial of the rating task consisted of a video and its subsequent rating. To ensure that each of the 92 videos was viewed by the same number of participants, they were randomly assigned to one of four experimental groups. Items were presented in randomized order. Before and after the video presentation, a fixation cross was presented in the center of the screen for a random duration in the range 500–1,000 ms. Each video sequence was followed by a screen asking for ratings on a scale from 1 to 4 (through keyboard presses): 1 = "not important at all" (German: "unwichtig"); 2 = "rather unimportant" (German: "eher unwichtig"); 3 = "rather important" (German: "eher wichtig"); 4 = "very important" ("sehr wichtig"). Four items were used as practice trials.

The rating study was followed by a recognition task. Here the words from the rating task and the same number of distractor words were presented on screen alongside their respective definite articles in the nominative case. Participants were instructed to indicate whether the respective object had been presented during the rating task or not (through keyboard presses). Thus, this task was designed to test whether they recognized the objects used in the rating task. After the recognition task, participants filled in a questionnaire regarding their experience with the tasks and stimuli as well as possible rating strategies. Finally, participants were debriefed and reimbursed for their participation.

2.9 Eye-tracking

Eye-tracking was carried out using an *SR Research Eyelink 1,000 plus* configured for desktop mount. The distance from the chin rest was 55 cm to the eye-tracker and 90 cm to the screen. The sampling rate was 1,000 Hz. Calibration and validation were performed before the rating task with a 9-point calibration procedure. During the rating task, we additionally included a drift check after every tenth trial to improve the quality of the eye-tracking data. Blinks were excluded from the analysis. Eye-tracking data of 3 participants (2 controls, 1 autistic) had to be discarded due to technical problems and did not enter the relevant analyses, i.e., the analysis of fixation durations and the Bayesian models for object recognition rates.

2.10 Analysis

The permutation software was implemented in *R* (R Core Team, 2023). Other analyses were implemented in *R* (R Core Team, 2019) and *RStudio* (RStudio Team, 2016). When reporting significance of t-tests, we assumed a 95% confidence interval.

For the analysis of participants' ratings, we performed non-parametric permutation tests (Odén and Wedel, 1975; Pesarin and Salmaso, 2010; Berry et al., 2011; Good, 2013) to determine likelihoods of the effects of conditioning arising by chance. These tests explored the effect of the virtual character's gaze duration and pitch height on the participant's rating as to how important an object was considered to be for the character. The dependent variable predicted in these tests was the raw rating data. Corresponding to the four experimental groups, participants' data sets were grouped into four sets of equal size, with the same number of participants with an ASD diagnosis and control participants. Within each experimental group, participants were arranged into pairs, each containing one person of each diagnostic group, with the pairs aligned for maximum age similarity. Thus, experimental group, age-pair, and diagnosis together served to specify a single participant. The conditions of gaze and pitch variation were assessed by using within-subject permutations, while the effect of diagnosis was assessed by permuting data between participants matched for group and age-pair.

For each condition, we ran 1,000,000 permutations. Permutation evaluations were treated as independent samples from a distribution, and the beta function was used to assess the extent of the 95% confidence interval for the likelihood p of a permutated value for the rating exceeding the actual value. This upper limit on the confidence value is reported as p below.

For the analysis of eye-tracking data three regions of interest were defined: The eye region was defined by a rectangle (212 × 110 px) containing the eyes and a small area around the eyes, including the eyebrows. The head region was defined by a rectangle (280 × 414 px) fitting the virtual character's head and excluding the region of interest defined for the eye region. The object region was defined as a square (280 × 280 px) that included the object and a small area around the object to account for the slightly different objects' proportions while at the same time keeping this region of interest constant across trials. Further, for the analysis of eye-tracking data and the recognition task (i.e., the correctness of the responses as to whether an object had appeared in the main experiment or not), Bayesian models (package *brms*; Bürkner, 2017; Bürkner and Vuorre, 2019) were fitted to the data. If not otherwise stated, dichotomous factors were deviation-coded, and continuous factors were z-transformed. In each model, we included random intercepts and slopes for *subject* as well as random intercepts for *object*. Estimated parameters are reported in terms of posterior means and 95% credibility intervals. The *emmeans* package (Lenth et al., 2021) was used to extract contrast coefficients. To investigate the evidence for or against the investigated effects, we compared models by calculating Bayes factors applying the *bayesfactor_models* function from the *bayestestR* package (Makowski et al., 2019) which uses bridge sampling (Gronau et al., 2020). We report respective Bayes factors of model comparisons and follow the interpretation by Lee and Wagenmakers (2014). All models ran with four sampling chains of 12,000 iterations each including a warm-up period of 2,000 iterations.

For the analysis of the influence of *diagnosis* and *cluster* and their *interaction* on the duration of fixations within the three regions of interest, we included eye-tracking data starting at the onset of the gaze cue (= onset of the auditory stimulus). We modelled a proportional value for fixation duration, namely fixation duration directed towards the region of interest divided by the video duration starting at cue onset, separately for each region. Bayesian linear zero-inflated beta models [r package *brms*; Bürkner, 2017; Bürkner and Vuorre, 2019) were fitted to the data. Fixed effects were *diagnosis* and *cluster*. Weakly informative priors were used (intercept prior: normal distribution, $M = 0.5$, $SD = 0.5$; slope priors: normal distribution, $M = 0$, $SD = 0.5$; SD priors: normal distribution, $M = 0$, $SD = 0.5$; phi priors: normal distribution, $M = 0.5$, $SD = 0.5$; zi prior: $M = 0.2$, $SD = 0.5$).

The Bayesian logistic binomial regression model for object recognition in the recognition task was fitted exclusively to data pertaining to stimuli presented in one of the four conditions. Thus, false positive responses or true rejections following the presentation of distractors were not analyzed. We included fixed effects previously identified as important in the general population: the untransformed, proportional values for *participant's gaze duration* towards the object region during the rating task; the logarithmized values of *word frequency*; and the *number of trials that had passed since object presentation*. We compared this model with models that additionally included *ASD diagnosis,* the virtual character's *gaze duration* towards the object and *pitch height*. Weakly informative priors were used (intercept prior: normal distribution, $M = 0$, $SD = 0.5$; slope priors: normal distribution, $M = 0$, $SD = 0.5$; SD priors: normal distribution, $M = 0$, $SD = 0.5$; LKJ prior: 1). Results are reported on the log-odds scale.

3 Results

Scores indicating autistic traits, measured with the AQ, were significantly higher in participants with autism compared to the control group [Table 3, two-samples t-test, $t(46) = -20.18$, $p < 0.001$]. Scores indicating empathetic traits, measured with the EQ, were significantly lower in autistic participants compared to the control group [Table 3, Welch two-samples t-test, $t(37.07) = 14.42$, $p < 0.001$]. Scores indicating tendencies to systemize, as measured with the *SQ*, were significantly higher in autistic participants compared to the control group [Table 3, two-samples t-test, $t(46) = -5.64$, $p < 0.001$]. Mentalizing abilities, as indicated by the *Eyes-test* scores, were

TABLE 3 Psychological screening scores.

	AQ	EQ	SQ	Eyes-test
ASD ($N = 24$)	$M = 42.1$ ($SD = 4.3$)	$M = 11.5$ ($SD = 6.0$)	$M = 45.0$ ($SD = 13.8$)	$M = 16.0$ ($SD = 4.6$)
Control ($N = 24$)	$M = 14.2$ ($SD = 5.3$)	$M = 46.3$ ($SD = 10.2$)	$M = 23.6$ ($SD = 12.4$)	$M = 19.0$ ($SD = 2.8$)

AQ, autism spectrum quotient; EQ, empathy quotient; SQ, systemizing quotient; Eyes-test, "Reading the mind in the Eyes" test.

significantly higher in the control group than in the autistic group [two-samples t-test, $t(46) = 2.76$, $p = 0.008$]. These results further support the clinical diagnosis.

3.1 Rating behavior

The condition characterized by short gaze duration and low pitch height yielded the lowest mean ratings in both the autistic group ($M = 1.90$, $SD = 0.53$) and the control group ($M = 2.14$, $SD = 0.37$). The condition with both long gaze and high pitch yielded the highest mean ratings in both groups (ASD: $M = 2.70$, $SD = 0.54$; control persons: $M = 2.65$, $SD = 0.48$). The conditions with either longer gaze duration (ASD: $M = 2.55$, $SD = 0.61$; control persons: $M = 2.45$, $SD = 0.46$) or increased pitch height (ASD: $M = 2.13$, $SD = 0.74$; control persons: $M = 2.45$, $SD = 0.58$) yielded mean ratings between the two afore-mentioned conditions. Mean ratings (see Figure 4) therefore replicate the general pattern reported for a sample from the general population in our previous web-based study (Zimmermann et al., 2020).

We assessed the significance of the differences in ratings as a function of condition and diagnosis by means of permutation tests. Long gaze significantly increased participants' ratings ($p < 0.001$). This held true regardless of the combination of diagnosis and pitch, i.e., both in the autistic and non-autistic group, ratings in conditions in which the virtual character looked towards the object for a long duration were higher than those for conditions in which the gaze was short, both for the high-pitch and low-pitch conditions. Pitch height had a slightly weaker impact on the ratings but again significantly increased participants' ratings ($p < 0.001$) for all combinations of diagnosis and gaze duration, i.e., both in the autistic and non-autistic group, ratings in conditions in which pitch was high, were higher than those for conditions in which pitch was low, both for the long-gaze and the short-gaze conditions. The only exception from this general pattern were participants diagnosed with ASD looking at long gaze: for this latter combination, the effect was also significant ($p = 0.001$), but potentially more likely to have occurred by chance.

Finally, we examined the impact of diagnosis on distinct combinations of pitch height and gaze duration. For short gaze, regardless of pitch height, ratings of autistic participants were significantly lower than those of the control group ($p < 0.001$). When gaze was long but pitch was low, ratings of autistic participants were significantly higher than those of the control group ($p = 0.004$). When both gaze was long and pitch was high, ratings of autistic and non-autistic participants did not differ significantly ($p = 0.130$). Rating differences in response to the two different gaze cue durations were thus greater in the autistic group than in the non-autistic group, indicating that different gaze durations of the virtual character towards the object had a greater effect in the autistic group.

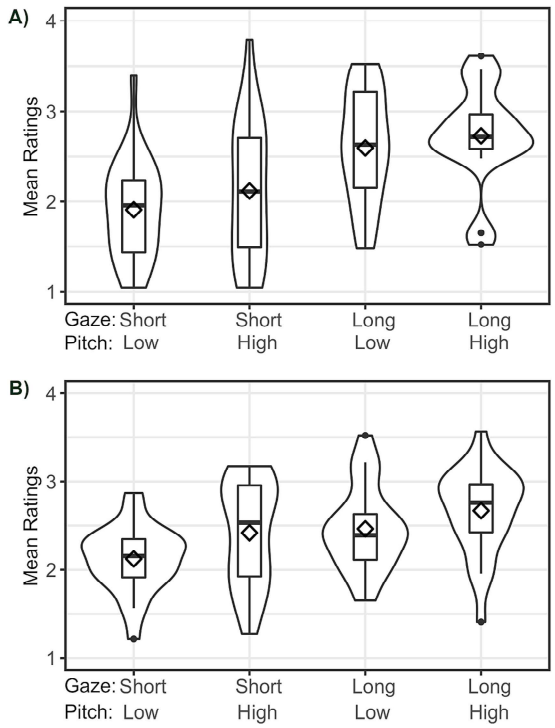

FIGURE 4
Distribution of participants' mean ratings of stimuli: **(A)** ASD group, **(B)** Control group. The range of the y-axis equals the total rating scale (1–4). Diamonds indicate means across subjects. Horizontal lines indicate medians.

These results reflect the visible differences by condition and diagnosis seen in Figure 4.

3.2 Individuality

Similar to the findings in our web-based study (Zimmermann et al., 2020), there was substantial inter-individual variability. Participants' ratings were predominantly influenced by either one or the other factor rather than by both factors in combination. Figure 5 shows each participant's individual cue use behavior regarding *gaze duration* and *pitch height*, indicated by the difference between their mean ratings for long vs. short gaze duration conditions and the difference between their mean ratings for high vs. low pitch height conditions. For each participant, we carried out two Wilcoxon rank sum tests—including the expectation that longer gaze and higher pitch would each increase ratings—on the ratings for the long- versus short-gaze and high- versus low-pitch conditions, respectively. The resulting

10.3389/fcomm.2024.1483135

p-values indicating significant differences at the 5% level were used as indicators that the individual made use of the respective cue. Participants were subsequently categorized as "Lookers" if ratings were significantly higher in the long-gaze conditions than in the short-gaze conditions, as "Listeners" if ratings were significantly higher in the high-pitch conditions than in the low-pitch conditions, as "Neithers" if ratings did not differ significantly for either factor, and as "Both" if ratings significantly differed for both factors. However, no participant was categorized as "Both" in this dataset, irrespective of diagnosis, mirroring results from our previous study (Zimmermann et al., 2020). The resulting three clusters are color-coded in Figure 5. Participants of both diagnostic groups can be found across all three clusters.

Interestingly, three participants that clustered as "Lookers" (two of these autistic) reported initially having used the virtual character's (tone of) voice for their ratings, but switching to concentrating on the character's gaze towards the objects, once they had detected this cue. In the "Listeners" cluster, only one participant reported also having used the character's gaze towards the object for their ratings.

Participants clustered as "Neithers" were not consistently influenced by either gaze duration or pitch height. However, when asked for their rating strategy in free-text form, some of the "Neithers" reported having taken into account the gaze behavior of the virtual character or the voice stimulus for their ratings, few specifically referred to the virtual character's gaze duration towards the objects or the tone of voice. However, none of the participants in the "Neithers" cluster reported exclusively having taken into account either intonation or gaze duration towards the object (or both). Instead, they attended to more than one source of information, amongst them the character's blinking behavior, the duration of the second idle gaze phase, gaze direction and loudness. One participant reported that the different durations of the character's gaze towards the object did not influence their rating behavior as it did not affect their perception as to how important the objects appeared to be for the character. Only one (autistic) participant in the "Neithers" cluster reported sometimes having guessed. Across both autistic and non-autistic participants, some reported having concentrated on the object itself (its animacy, entertaining quality, potential benefit or danger) or their personal perception of the object's importance as well as the object's presumed importance for the virtual character based on her age, gender and appearance.

3.3 Fixation durations

Overall, both the autism group and the control group spent more time looking at the eye region (ASD: $M = 3.88$ s, $SD = 1.61$; control group: $M = 4.10$ s, $SD = 1.26$) than at the object (ASD: $M = 1.01$ s, $SD = 0.74$; control group: $M = 1.04$ s, $SD = 0.74$) and head region (ASD: $M = 1.05$ s, $SD = 0.65$; control group: $M = 1.05$ s, $SD = 0.60$) (see Figure 6).

Across diagnostic groups, within the three clusters, rating behavior was reflected by fixation durations within the three regions: Compared to the other groups, the group of "Lookers" looked longer at the eye region ("Lookers": $M = 4.85$ s, $SD = 1.04$; "Listeners": $M = 3.01$ s, $SD = 1.36$; "Neithers": $M = 3.67$ s, $SD = 1.30$), but spent less time fixating the object region ("Lookers": $M = 0.55$ s, $SD = 0.24$; "Listeners": $M = 1.49$ s, $SD = 0.85$; "Neithers": $M = 1.30$ s, $SD = 0.69$). Fixation durations within the head region (not including the eye region) were similar between clusters ("Lookers": $M = 0.92$ s, $SD = 0.64$; "Listeners": $M = 1.09$ s, $SD = 0.61$; "Neithers": $M = 1.24$ s, $SD = 0.59$). Visual inspection suggested that within the "Listeners" cluster, the difference between mean fixation durations towards the eye region for participants with an ASD diagnosis and control participants tended to be greater than the respective difference within the other two

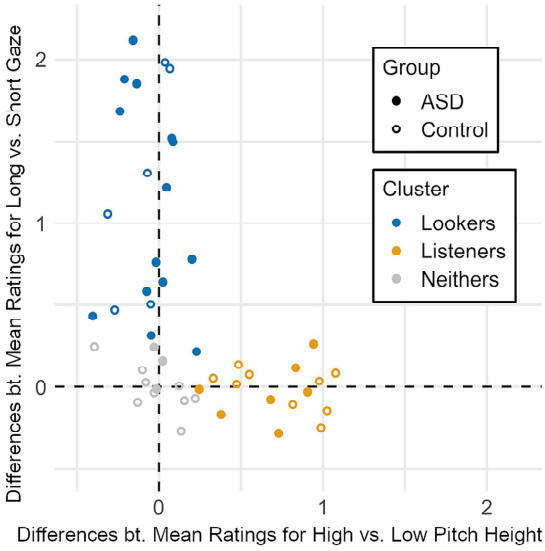

FIGURE 5
Individual differences between mean ratings for high vs. low pitch height conditions (*x*-axis) and individual differences between mean ratings for long vs. short gaze duration conditions (*y*-axis) combined to one coordinate for each participant. Cluster labelling is based on the significance of these differences at the 5% level. There are no participants with significant differences on both axes.

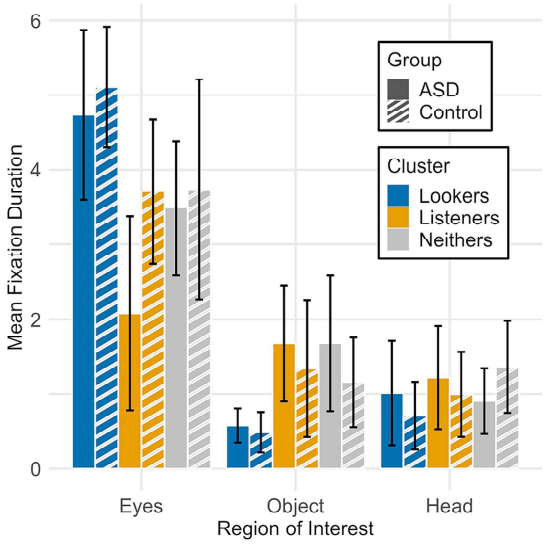

FIGURE 6
Mean fixation durations and standard deviations for the ASD and control group and the three clusters within the three regions of interest. The total possible per-trial fixation duration is 6.6 s.

clusters. This observation was, however, not supported by statistical analysis.

We analyzed the influence of *diagnosis* and *cluster* on fixation durations beginning at gaze cue onset (which coincides with the onset of the auditory stimulus), separately for each region of interest. In sum, *cluster* was identified as a statistical reliable influence on total fixation duration within the eye region and the object region, however, *diagnosis* was not: we found only anecdotal evidence for a *diagnosis* effect for the eye region ($b = -0.36$; 95% CI = [−0.77, 0.05], BF = 1.93), but anecdotal evidence against an effect in the object region ($b = 0.19$; 95% CI = [−0.10, 0.50], BF = 0.73) and in the head region ($b = 0.01$; 95% CI = [−0.37, 0.40], BF = 0.39). In support of the finding of cluster-dependent gaze patterns reported above, extreme evidence for an effect of *cluster* was found in the eye region (BF > 100) and object region (BF > 100), while anecdotal evidence against an effect was found in the head region (BF = 0.57). Specifically, and irrespective of diagnostic group, "Lookers" spent more time fixating the eye region than "Listeners" ($b = 0.95$; 95% CI = [0.50, 1.39], BF > 100) and tended to also spend more time fixating this region than "Neithers" ($b = 0.61$; 95% CI = [0.09, 1.14], BF = 4.90), while "Listeners" tended to spend less time fixating the eye region than "Neithers" ($b = -0.35$; 95% CI = [−0.82, 0.16], BF = 1.33). In comparison to the "Lookers," the "Listeners" ($b = 0.63$; 95% CI = [0.30, 0.96], BF > 100) and "Neithers" ($b = 0.67$; 95% CI = [0.28, 1.05], BF = 70.42) spent more time fixating the object region, while there was no difference between "Listeners" and "Neithers" ($b = -0.04$; 95% CI = [−0.40, 0.33], BF = 0.38). We found anecdotal evidence against an interaction effect of *diagnosis* and *cluster* in the eye region (BF = 0.70) and object region (BF = 0.40). Moderate evidence for an interaction effect was found in the head region (BF = 3.71). Further investigation of this effect revealed that it was mainly driven by tendencies within the control sample: participants in the "Neither" cluster tended to fixate the head region for a longer duration than both "Lookers" (BF = 4.97) and—to a lesser extent—"Listeners" (BF = 3.27), while no difference was found between the "Listeners" and "Lookers" (BF = 0.62). Within the autism sample, no statistically reliable differences between clusters were found for the head region (0.48 < BFs < 1.0).

3.4 Object recognition

Recognition rates for target words tended to be slightly lower in the autistic group compared to the control group, but were similar within groups for all four conditions (Table 4). Correct identification of distractor words was comparable between groups (ASD: $M = 93.0\%$, $SD = 5.4$; Controls: $M = 93.3\%$, $SD = 5.4$).

For target words, we found extreme evidence for an effect of *participants' fixation* duration towards the object on their memory performance, with longer fixation of an object increasing recognition ($b = 0.49$; 95% CI = [−0.03, 0.98], BF > 1,000). The *number of trials that had passed since object presentation* also had a statistically robust effect on memory performance: The fewer trials passed since object presentation, the greater the likelihood the respective word was recognized correctly in the recognition task ($b = -0.28$, 95% CI = [−0.37, −0.19]; BF > 1,000). Additionally, very strong evidence was found for an effect of *word frequency*: more frequent words tended to lead to better recognition ($b = 0.06$, 95% CI = [−0.09, 0.22]; BF > 100). Anecdotal evidence was found for including the factor *ASD diagnosis* (BF = 1.44). Including the factors *gaze duration* or *pitch height* did not improve model fit (gaze duration: BF = 0.07; pitch height: BF = 0.22).

3.5 Exploratory correlation analysis

Within the two diagnostic groups, we performed exploratory correlation analyses for differences between mean ratings (for high vs. low *pitch height* conditions and for long vs. short *gaze duration* conditions; see Figure 5) in combination with the *SPQ visual* and *auditory* scores. We found a statistically noteworthy relationship for the *SPQ* regarding *gaze duration*: In the control group, higher *SPQ visual* scores (indicating lower visual sensitivity) were significantly linked to taking gaze duration into account to lesser extent ($r_S = -0.444$, $p = 0.030$), which was not the case in the autistic group ($r_S = -0.183$, $p = 0.393$). No significant correlation between *SPQ visual* scores and differences between mean ratings for *pitch height* conditions was observed in the autistic and control group (ASD: $r_S = 0.012$, $p = 0.956$; Controls: $r_S = 0.290$, $p = 0.169$). No significant correlation was found between *SPQ auditory* scores and the differences between mean ratings for *gaze duration* conditions (ASD: $r_S = 0.165$, $p = 0.442$; Controls: $r_S = -0.092$, $p = 0.669$) and *pitch height* (ASD: $r_S = 0.047$, $p = 0.828$; Controls: $r_S = 0.150$, $p = 0.486$).

4 Discussion

4.1 Rating behavior

At the group-level, participants from both the autism group and the control group rated the importance of the object to the virtual character to be higher when any of the two deictic signals (gaze or pitch accent) suggested that the virtual character was more interested in the particular object (through longer gaze or higher pitch), confirming the results of our previous web-based study (Zimmermann et al., 2020).

Compared to the control group, autistic participants took gaze duration into account to a greater extent than pitch height: They

TABLE 4 Object recognition rates.

	Group	Low pitch	High pitch
Short gaze duration	ASD ($N = 24$)	$M = 54.2\%$ ($SD = 21.0$)	$M = 56.6\%$ ($SD = 24.5$)
	Control ($N = 24$)	$M = 67.6\%$ ($SD = 19.5$)	$M = 68.5\%$ ($SD = 21.8$)
Long gaze duration	ASD ($N = 24$)	$M = 60.0\%$ ($SD = 22.3$)	$M = 59.2\%$ ($SD = 20.3$)
	Control ($N = 24$)	$M = 66.0\%$ ($SD = 18.6$)	$M = 68.5\%$ ($SD = 20.0$)

judged the object's importance to the virtual character to be lower than the control group when it was gazed at for a short duration. They rated the importance higher than the control group when the object was gazed at for a long duration if presented with low pitch – and as high as the control group if it was presented with high pitch.

One explanation for the fact that participants with autism in our paradigm assigned more weight to the virtual character's gaze (as opposed to pitch height) might be an impaired interpretation of vocal pitch, both in speech (Grice et al., 2016, 2023; Schelinski and von Kriegstein, 2019) and non-speech (Schelinski et al., 2017). The study by Grice et al. (2023) suggests that the interpretation of prosody (amongst others intonation) is similar in autistic listeners and non-autistic listeners when it is used by the speaker to convey rule-based information such as syntactic structure. However, when it is used to convey less rule-based and more intuitive pragmatic aspects, such as the importance of a certain word, the interpretation of prosodic information seems to be more difficult for autistic listeners. An example for the latter is an investigation of intonation perception in autism (Grice et al., 2016): In this study, autistic listeners were less sensitive to intonation than the non-autistic group. Instead, they used other information about the words themselves, such as semantic information (human-non-human for instance), to judge whether a word presented in an auditorily presented sentence was new information or not. If participants in our paradigm found it difficult to interpret pitch height, this might be a reason for them to instead search for other information to solve the task.

Another reason for autistic participants to more strongly weigh the gaze cue rather than the pitch cue could be greater auditory capacity in comparison to control participants (Remington and Fairnie, 2017). In this study, autistic listeners were able to detect more auditory stimuli than the non-autistic group, regardless of whether they were distractors to the main task or not. Perceiving a wealth of auditory information might be beneficial in certain scenarios but could also be detrimental or exhausting in others. In our paradigm, the auditory information is arguably more complex than the visual information: The speech stimulus was a different one in each trial. Furthermore, since we used natural speech, the intonation pattern slightly varied for each item: Even if the accented syllable of each high-pitch stimulus is always 45 Hz higher relative to its low-pitch counterpart, low-pitch stimuli exhibit small fluctuations in their absolute Hz values. Additionally, other prosodic factors might influence prominence perception, such as the length of the utterance. The gaze cue, on the other hand, is comparably simple to perceive and categorize, as it was always set to either 0.6 or 1.8 s in a binary fashion. Therefore, a person processing the abundance of information presented with the auditory cue might find it easier to pay attention to the gaze cue instead, either because they do not detect the manipulated cue amongst the noise of other auditory information, or because this is more effortful than focusing on gaze duration.

The findings of the exploratory analyses showed that, in part, rating tendencies could plausibly be linked to sensory perception: within the control group, lower visual sensitivity was linked to less focus on gaze. This suggests that general visual sensitivity affects participants' ratings. One reason for a lack of this relationship in the autistic group could be that – instead of relying on their default perception – they attuned to the task's systematic structure more strongly than the control group did, which could also explain why they weighed the gaze cue more strongly than the pitch cue.

Other studies have reported that autistic participants had difficulties in solving mentalizing tasks that rely on nonverbal information (Baron-Cohen et al., 2001a; Baron-Cohen et al., 2001b; Ponnet et al., 2004; Dziobek et al., 2006; White et al., 2011). Those tasks involved more than two signals that varied in more than two steps, so that it was unclear which cue was informative. Additionally, the response required more complex mentalizing tasks than the current experiment (e.g., identifying different mental states from a selection of alternatives, or freely inferring mental states). In contrast, our task provides a much more structured setting, with only two cues varying by two different degrees. Moreover, the simple question to be answered is the same throughout. The most obvious strategy to solve the task is to identify (at least) one varying source of information and preferentially rely on that source.

4.2 Individuality

Gaze cues (Bayliss et al., 2007) and pitch height cues (Roy et al., 2017; Baumann and Winter, 2018) are not perceived and processed in the same way by every individual. Participants' ratings in our study tended to be influenced by either one or the other factor rather than by both factors in combination. Based on their rating behavior, participants clustered into three subgroups: (i) "Lookers," who based their ratings primarily on gaze duration, (ii) "Listeners," who based their ratings primarily on pitch height, and (iii) "Neithers," whose ratings were not predominantly influenced by either of these two cues. Participants of both diagnostic groups were found across all three clusters. The observation discussed above that autistic participants were more strongly influenced by the gaze cue was reflected in the distribution of clusters as well: autistic participants were identified as "Lookers" twice as often as they were identified as "Listeners." This pattern was not visible in control participants: six participants were categorized as "Lookers," whereas nine were categorized as "Listeners" in the control group. Based on previous findings of the high relevance of verbal at the expense of nonverbal information in autism (Kuzmanovic et al., 2011; Stewart et al., 2013) and of a reliance on invariant characteristics of words at the expense of intonation (Grice et al., 2016), we expected more autistic participants to cluster as "Neithers." However, this was not the case.

It is striking that none of the participants was considerably influenced by both gaze duration and pitch height together. Several studies that investigated the perception of pitch accents in combination with salient facial movement, head or hand gestures in the general population have shown that they can, in fact, lead to greater prominence perception compared to the presentation of only one modality (Krahmer et al., 2002; Swerts and Krahmer, 2008; Mixdorff et al., 2013; Prieto et al., 2015; Ambrazaitis et al., 2020). A possible explanation for the finding that, at the individual level, a combination of long gaze and high pitch in the current paradigm did not lead to higher ratings of object importance compared to when only one of the two cues was rendered prominent, might be that participants default to efficient cue use in this task. The instruction did not specify whether the virtual character would communicate via eye gaze, prosody or other behavior. Accordingly, participants had the freedom to use one, two, multiple or no cues at all. Increased multimodal cue use has been

reported in audiovisual studies in which auditory information is insufficient or difficult to understand (Munhall et al., 2004; Dohen and Lœvenbruck, 2009; Moubayed and Beskow, 2009; Macdonald and Tatler, 2013). For example, in a demanding, but highly structured task (Macdonald and Tatler, 2013), participants from the general population made use of the instructor's gaze behavior only, if the auditory information was not informative enough. Comparably, in the current paradigm, there was no need for participants to identify additional cues, as long as they found at least one cue that helped them solve the task. Identifying one cue and sticking to it may be the most efficient way to solve this task. Participants' feedback regarding their rating strategies lends anecdotal support for this idea: Four participants explicitly reported having focused on the virtual character's gaze towards the object as well as intonation. Three of these participants (two of them autistic) reported having used primarily the gaze cue for the remainder of the experiment, which exemplifies the efficiency of participants' cue use in this task.

The finding that participants in the "Neithers" cluster did not demonstrate a preference for either the gaze cue or the pitch cue does not necessarily imply that these did not affect their ratings at all, but that they weighed other cues more strongly. Feedback from these participants on their rating strategies suggests that some focused on the object's properties and the virtual character's characteristics when carrying out their ratings. Others did, in fact, attend to the character's gaze behavior and the voice stimuli, but considered aspects of gaze and voice other than the manipulated cues, such as gaze directions, blinking or voice loudness. Those that actually took into account the manipulated cues, additionally paid attention to other cues that were not manipulated, which may have attenuated potential effects of gaze duration or pitch height on their ratings.

4.3 Eye-tracking

Both diagnostic groups spent more time fixating the eye region than the object and head region. This finding is in line with previous eye-tracking studies: in the general population, a tendency to fixate the eye region for longer than either other parts of the face or objects in the environment has been reported across different tasks (Henderson et al., 2005; Freeth et al., 2010; Fedor et al., 2018). Similar fixation tendencies have been reported for individuals on the autism spectrum (Dalton et al., 2005; Hernandez et al., 2009; Freeth et al., 2010; Auyeung et al., 2015; Fedor et al., 2018).

We did not find reliable statistical evidence for differences between the autism and the control group regarding fixation durations for the eye region or the object region. A meta-analysis of 22 studies has reported shorter fixation durations for the eye region as opposed to objects in adult participants with autism in free viewing tasks (Setien-Ramos et al., 2022). Our paradigm was not suited to induce gaze aversion in autism as it required participants to search for potentially informative cues. Information variation was limited to the eyes, voice and object, and only the eye region showed visual change within a given trial (eye blinks, changing gaze direction). Thus, avoiding the character's gaze (and assuming the eye region is not processed via peripheral vision) would entail ignoring one of three relevant channels of information. Presenting only one rather static virtual character as well as a relatively long trial duration may further have shifted attention towards the eye region in our study.

Across both diagnostic groups, we were able to show that participants' rating behavior was in line with their gaze behavior: "Lookers" spent more time looking at the virtual character's eyes than "Listeners" and tended to also spend more time looking at the eyes than "Neithers." "Listeners" spent less time looking at the eyes than "Neithers." "Lookers" spent less time looking at the object than both other clusters, which did not differ in this regard. This finding corroborates the well-established notion that attention is closely linked to gaze direction (Buswell, 1935; Yarbus, 1967; DeAngelus and Pelz, 2009), leaving the "Lookers" no choice but to fixate the eye region and mostly ignore the object, while "Listeners" and "Neithers" were free to visually explore other areas as well.

Exclusively for the eye region, visual inspection—but not the statistical analysis—showed a small tendency for shorter fixation durations in "Listeners" with an ASD diagnosis compared to "Listeners" from the control group. It is possible that an underlying trend was not detected in the analysis. If present, it could suggest different strategies for solving the task: "Listeners" need to pay attention to the acoustic signal and do not depend on gathering information from the eyes. Especially for people with autism, who may experience mutual gaze as threatening or stressful, this could result in avoiding mutual gaze (Tottenham et al., 2014). In our study, we did not ask about uneasiness while fixating the eye region. Only one participant in the autism group reported exhaustion due to looking at the virtual character's face and the eye region in particular. A tendency within the autism group for the "Listeners" to look at the eye region for a shorter total duration could also indicate that persons with autism by default perceive the eyes as deictic cues but not as mutual gaze, which is a stronger social cue (Ristic et al., 2005; Caruana et al., 2018). Riby et al. (2013), who included children and adolescents with autism in their study, reported that the eye region was fixated for a shorter duration by their autistic group in comparison to a control group. In the autistic group, fixation duration towards the eye region, unsurprisingly, increased upon instruction to detect what the person in the photo was looking at.

In the control group, we found a tendency towards shorter fixations of the head region (not including the eyes) in the "Lookers" and "Listeners" compared to the "Neithers." The behavior in the rating task and our eye-tracking data support the idea that participants were actively monitoring their chosen input modality, searching for informativeness in these cues. Accordingly, we interpret the tendency of the "Neither" cluster as more strongly than the other clusters using the head region as a source of information. Three participants in the "Neither" cluster reported having taken into account virtual-character-related characteristics such as gender and age for their rating. Only one subject from the "Listeners" cluster reported potentially having been influenced in a similar fashion.

4.4 Object recognition

To detect possible memory traces of attention directed towards the objects, we included an object recognition task after completion of the rating task. Findings regarding word or object recognition in autistic adults without intelligence deficits have been mixed so far, with some studies reporting comparable performance in autism (Bowler et al., 2000; Boucher et al., 2005; Ring et al., 2015) and others showing worse recognition rates in autism (O'Hearn et al.,

2014). We found no reliable evidence for different recognition rates in participants with autism compared to the control group. Across groups, object recognition was better for objects that had previously been fixated by participants for a longer duration, which is in line with previous research on visual memory: the longer we look at an object, scene or face, the better we can later remember it (Melcher, 2001, 2006; Droll and Eckstein, 2009; Martini and Maljkovic, 2009). We also found a serial-position effect: participants could better recognize objects they had seen more recently, which is in line with previous research (Brady et al., 2008; Konkle et al., 2010). Importantly for our purposes, implicit memory in autism is considered comparable to that in the non-autistic population (Ring et al., 2015). Our results stand in contrast to other studies that reported an influence of gaze and pitch on object memory (Fraundorf et al., 2010, 2012; Dodd et al., 2012; Adil et al., 2018; Wahl et al., 2019; Ito et al., 2022). However, these studies are not directly comparable to our study because they manipulate neither gaze duration nor pitch excursion specifically.

In our study, low word frequency did not improve object memory. Instead, participants could better recognize objects described with more frequent words. Our paradigm is primarily designed to probe a very simple mentalization task requiring a judgment of how important an object appears to be for the virtual "person." Comparably, in our online study (Zimmermann et al., 2020), participants' ratings for the importance of the object for the virtual character increased with higher word frequency. We assume that word frequency in our stimulus set is confounded with other object properties (Zimmermann et al., 2020). The five most frequent words in our dataset were the German words for "car," "plane," "window," "sun" and "church." The five least frequent words were the German words for "spinning wheel," "doorknob," "spinning top," "chisel" and "roller skate." We assume that, among other factors, general object importance might have affected the ratings.

4.5 Limitations

To summarize the limitations of our paradigm discussed in previous work (Zimmermann et al., 2020), the most pertinent issue is the reductionistic design employed and its effect on the perception of the virtual character's mental state. This entails that the relevant experimental findings cannot be easily transferred to everyday social situations, which are much more complex. Additionally, the task instructions may have led participants to actively search for a cue and to then stop searching once a valid cue was found. In the following section, we will focus on issues specific to autism and the findings of the current study.

It could be argued that autistic participants may concentrate on gaze more than on intonation, because eye contact is a common target in early interventions in autism. However, most participants in our study were recruited in the outpatient clinic for autism in adulthood. This implies that they did not receive any autism-specific therapy before their diagnosis in adulthood. As we have not systematically asked every participant whether he or she received any specific training in nonverbal communication skills including mutual gaze, it cannot be ruled out that they did, but it is unlikely. Assessing a potential influence of such a training may nevertheless be informative in future studies.

It is possible that the external validity of our results is not only limited, but also differs between diagnostic groups. A study including children and adolescents has shown that the gaze behavior of participants with autism in reaction to a computer screen differed from gaze behavior in reaction to a live interaction, which was not the case for the control group (Grossman et al., 2019). No difference was, however, reported for gaze behavior in reaction to static images of virtual characters' faces as compared to photographs of real people in autistic adolescents and adults (Hernandez et al., 2009). In real-life scenarios, the problems persons with autism face when interpreting eye gaze do not only arise from difficulties with deciphering the "correct" social implications, but also from understanding when eye gaze may contain social implications in the first place, beyond, e.g., deictic information, which is itself problematic in autism (Pantelis and Kennedy, 2017; Griffin and Scherf, 2020). The latter was not part of this experiment, as the task (according to the interpretation of most participants) implicitly called for a social reading. Due to the limited stimulus variability, attending to the relevant cues was, moreover, easier than in real-life scenarios.

Future studies investigating the perception of gaze duration and intonation in a non-verbal mentalizing task in autism should aim to increase ecological validity by (1) using more natural social scenarios as stimulus material that does not only vary with respect to two isolated cues, and (2) by using different instructions or questions in each trial. We expect that this may reduce potential strategic cue searching strategies.

5 Conclusion

The current study aimed to investigate mentalizing behavior based on eye gaze and speech intonation in autism. Comparably to control participants, autistic persons used both gaze duration and intonation as cues for inferring the importance of an object for another (virtual) person. Compared to the control group, autistic participants were, however, influenced more strongly by gaze duration than by pitch height. Across both diagnostic groups, participants used either gaze or intonation as predominant cues, while some did not show this cue preference but might have used other cues predominantly to make their decision. Further investigations are required to accurately characterize differences in mentalizing abilities in autism in the nonverbal domain.

Data availability statement

The datasets presented in this article are not readily available because data handling is restricted to the collaborating institutes by our ethics proposal to secure sensitive data such as psycho(patho-)logical data. Requests to access the datasets should be directed to kai.vogeley@uk-koeln.de.

Ethics statement

The studies involving humans were approved by the Ethics Committee of the University Hospital Cologne. The studies were conducted in accordance with the local legislation and institutional requirements. The participants provided their written informed consent to participate in this study.

Author contributions

JZ: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Validation, Visualization, Writing – original draft, Writing – review & editing, Software, Supervision. TE: Data curation, Formal analysis, Writing – review & editing. FC: Methodology, Software, Writing – review & editing, Conceptualization. SW: Methodology, Writing – review & editing, Conceptualization. KV: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing. MG: Conceptualization, Funding acquisition, Methodology, Resources, Supervision, Writing – review & editing.

Funding

The author(s) declare that financial support was received for the research, authorship, and/or publication of this article. The study was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) -Project-ID 281511265 -SFB 1252.

Acknowledgments

We would like to thank C. Bloch, N. Oliveira-Ferreira, P. Da Silva Vilaca, and C. Esser for their support during data-acquisition, C. Röhr for producing the audio stimuli, H. Hanekamp for extracting acoustic parameters from these stimuli and the staff of the Phonetics Department of the University of Cologne for rating the processed audio stimuli.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Publisher's note

Supplementary material

The Supplementary material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fcomm.2024.1483135/full#supplementary-material

SUPPLEMENTARY TABLE 1
Details on the pretest's methods and results.

SUPPLEMENTARY TABLE 2
Object names and translations.

SUPPLEMENTARY DATE SHEET 1
Auditory stimuli.

SUPPLEMENTARY DATE SHEET 2
Video stimuli.

References

Adil, S., Lacoste-Badie, S., and Droulers, O. (2018). Face presence and gaze direction in print advertisements: how they influence consumer responses—an eye-tracking study. *J. Advert. Res.* 58, 443–455. doi: 10.2501/JAR-2018-004

Ambrazaitis, G., Frid, J., and House, D. (2020). Word prominence ratings in Swedish television news readings – effects of pitch accents and head movements, *The respective conference was the 10th International Conference on Speech Prosody in* Tokyo, Japan. 314–318. doi: 10.21437/SpeechProsody.2020-64

American Psychiatric Association (2013). Diagnostic and statistical manual of mental disorders. DSM-5. Washington, DC: American Psychiatric Association.

Arnold, D., Wagner, P., and Baayen, H. (2013). "Using generalized additive models and random forests to model prosodic prominence in German" in INTERSPEECH 2013, Eds. F. Bimbot, C. Cerisara, C. Fougeron, G. Gravier, L. Lamel, F. Pellegrino, P. Perrier (Lyon, France: ISCA) 272–276.

Aster, M., Neubauer, A., and Horn, R. (2006). Hamburg-Wechsler-Intelligenz-Test für Erwachsene III. Bern, Switzerland: Huber.

Aurchitect Audio Software, LLC (2018). Myriad. Aurchitect Audio Software is now onwed by Zynaptiq, Hannover, Germany: Aurchitect Audio Software, LLC.

Auyeung, B., Lombardo, M. V., Heinrichs, M., Chakrabarti, B., Sule, A., Deakin, J. B., et al. (2015). Oxytocin increases eye contact during a real-time, naturalistic social interaction in males with and without autism. *Transl. Psychiatry* 5, e507. doi: 10.1038/tp.2014.146

Baron-Cohen, S., Campbell, R., Karmiloff-Smith, A., Grant, J., and Walker, J. (1995). Are children with autism blind to the mentalistic significance of the eyes? *Br. J. Dev. Psychol.* 13, 379–398. doi: 10.1111/j.2044-835X.1995.tb00687.x

Baron-Cohen, S., Leslie, A. M., and Frith, U. (1985). Does the autistic child have a "theory of mind"? *Cognition* 21, 37–46. doi: 10.1016/0010-0277(85)90022-8

Baron-Cohen, S., Richler, J., Bisarya, D., Gurunathan, N., and Wheelwright, S. (2003). The systemizing quotient: an investigation of adults with Asperger syndrome or high-functioning autism, and normal sex differences. *Philos. Trans. R. Soc. Lond. Ser. B Biol. Sci.* 358, 361–374. doi: 10.1098/rstb.2002.1206

Baron-Cohen, S., and Wheelwright, S. (2004). The empathy quotient: an investigation of adults with Asperger syndrome or high functioning autism, and normal sex differences. *J. Autism Dev. Disord.* 34, 163–175. doi: 10.1023/B:JADD.0000022607.19833.00

Baron-Cohen, S., Wheelwright, S., Hill, J., Raste, Y., and Plumb, I. (2001a). The "Reading the mind in the eyes" test revised version: a study with normal adults, and adults with Asperger syndrome or high-functioning autism. *J. Child Psychol. Psychiatry* 42, 241–251. doi: 10.1111/1469-7610.00715

Baron-Cohen, S., Wheelwright, S., and Jolliffe, T. (1997). Is there a "language of the eyes"? Evidence from Normal adults, and adults with autism or Asperger syndrome. *Vis. Cogn.* 4, 311–331. doi: 10.1080/713756761

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., and Clubley, E. (2001b). The autism-Spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *J. Autism Dev. Disord.* 31, 5–17. doi: 10.1023/A:1005653411471

Baumann, S., and Winter, B. (2018). What makes a word prominent? Predicting untrained German listeners' perceptual judgments. *J. Phon.* 70, 20–38. doi: 10.1016/j.wocn.2018.05.004

Bayliss, A. P., Frischen, A., Fenske, M. J., and Tipper, S. P. (2007). Affective evaluations of objects are influenced by observed gaze direction and emotional expression. *Cognition* 104, 644–653. doi: 10.1016/j.cognition.2006.07.012

Beck, A. T., Steer, R. A., and Brown, G. K. (1996). Manual for the Beck depression inventory-II. San Antonio, TX: Psychological Corporation.

Ben-David, B. M., Ben-Itzchak, E., Zukerman, G., Yahav, G., and Icht, M. (2020). The perception of emotions in spoken language in undergraduates with high functioning autism Spectrum disorder: a preserved social skill. *J. Autism Dev. Disord.* 50, 741–756. doi: 10.1007/s10803-019-04297-2

Berry, K. J., Johnston, J. E., and Mielke, P. W. Jr. (2011). Permutation methods. *WIREs Comput. Statistic.* 3, 527–542. doi: 10.1002/wics.177

Bierlich, A. M., Bloch, C., Spyra, T., Lanz, C., Falter-Wagner, C. M., and Vogeley, K. (2024). An evaluation of the German version of the sensory perception quotient from an expert by experience perspective. *Front. Psychol.* 15:1252277. doi: 10.3389/fpsyg.2024.1252277

Bishop, J. (2016). Focus projection and prenuclear accents: evidence from lexical processing. *Lang. Cognit. Neurosci.* 32, 236–253. doi: 10.1080/23273798.2016.1246745

Bishop, J., Kuo, G., and Kim, B. (2020). Phonology, phonetics, and signal-extrinsic factors in the perception of prosodic prominence: evidence from rapid prosody transcription. *J. Phon.* 82:100977. doi: 10.1016/j.wocn.2020.100977

Boersma, P., and Weenink, D. (2018). Praat: doing phonetics by computer [Computer program]. Available at: http://www.praat.org/ (Accessed April 3, 2020)

Boucher, J., Cowell, P., Howard, M., Broks, P., Farrant, A., Roberts, N., et al. (2005). A combined clinical, neuropsychological, and neuroanatomical study of adults with high functioning autism. *Cogn. Neuropsychiatry* 10, 165–213. doi: 10.1080/13546800444000038

Bowler, D. M. (1992). "Theory of mind" in Asperger's syndrome. *J. Child Psychol. Psychiatry* 33, 877–893. doi: 10.1111/j.1469-7610.1992.tb01962.x

Bowler, D. M., Gardiner, J. M., and Grice, S. J. (2000). Episodic memory and remembering in adults with Asperger syndrome. *J. Autism Dev. Disord.* 30, 295–304. doi: 10.1023/A:1005575216176

Brady, T. F., Konkle, T., Alvarez, G. A., and Oliva, A. (2008). Visual long-term memory has a massive storage capacity for object details. *Proc. Natl. Acad. Sci.* 105, 14325–14329. doi: 10.1073/pnas.0803390105

Brickenkamp, R. (2002). Test d2: Aufmerksamkeits-Belastungs-test. *9th, revised and newly standardized* Edn. Göttingen: Hogrefe.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., and Böhl, A. (2011). The word frequency effect. *Exp. Psychol.* 58, 412–424. doi: 10.1027/1618-3169/a000123

Bürkner, P.-C. (2017). Brms: an R package for Bayesian multilevel models using Stan. *J. Stat. Softw.* 80, 1–28. doi: 10.18637/jss.v080.i01

Bürkner, P.-C., and Vuorre, M. (2019). Ordinal regression models in psychology: a tutorial. *Adv. Methods Pract. Psychol. Sci.* 2, 77–101. doi: 10.1177/2515245918823199

Buswell, G. T. (1935). How people look at pictures: A study of the psychology of perception in art. Chicago: University of Chicago Press.

Cangemi, F. (2015). Mausmooth [Praat script]. Available at: http://ifl.phil-fak.uni-koeln.de/sites/linguistik/Phonetik/mitarbeiterdateien/fcangemi/mausmooth.praat (Accessed April 13, 2019).

Caruana, N., Stieglitz Ham, H., Brock, J., Woolgar, A., Kloth, N., Palermo, R., et al. (2018). Joint attention difficulties in autistic adults: an interactive eye-tracking study. *Autism* 22, 502–512. doi: 10.1177/1362361316676204 (Accessed April 13, 2019).

Chita-Tegmark, M. (2016). Social attention in ASD: a review and meta-analysis of eye-tracking studies. *Res. Dev. Disabil.* 48, 79–93. doi: 10.1016/j.ridd.2015.10.011

Chuk, T., Chan, A. B., Shimojo, S., and Hsiao, J. H. (2016). Mind reading: discovering individual preferences from eye movements using switching hidden Markov models., in Proceedings of the 38th Annual Conference of the Cognitive Science Society 2016, 182–187.

Dahan, D., Tanenhaus, M. K., and Chambers, C. G. (2002). Accent and reference resolution in spoken-language comprehension. *J. Mem. Lang.* 47, 292–314. doi: 10.1016/S0749-596X(02)00001-3

Dalton, K. M., Nacewicz, B. M., Johnstone, T., Schaefer, H. S., Gernsbacher, M. A., Goldsmith, H. H., et al. (2005). Gaze fixation and the neural circuitry of face processing in autism. *Nat. Neurosci.* 8, 519–526. doi: 10.1038/nn1421

DeAngelus, M. A., and Pelz, J. (2009). Top-down control of eye movements: Yarbus revisited.

Del Bianco, T., Mazzoni, N., Bentenuto, A., and Venuti, P. (2018). An investigation of attention to faces and eyes: looking time is task-dependent in autism Spectrum disorder. *Front. Psychol.* 9:2629. doi: 10.3389/fpsyg.2018.02629

Dodd, M. D., Weiss, N., McDonnell, G. P., Sarwal, A., and Kingstone, A. (2012). Gaze cues influence memory…but not for long. *Acta Psychol.* 141, 270–275. doi: 10.1016/j.actpsy.2012.06.003

Dohen, M., and Lœvenbruck, H. (2009). Interaction of audition and vision for the perception of prosodic contrastive focus. *Lang. Speech* 52, 177–206. doi: 10.1177/0023830909103166

Doughty, M. J. (2001). Consideration of three types of spontaneous eyeblink activity in normal humans: during reading and video display terminal use, in primary gaze, and while in conversation. *Optom. Vis. Sci.* 78, 712–725. doi: 10.1097/00006324-200110000-00011

Droll, J. A., and Eckstein, M. P. (2009). Gaze control and memory for objects while walking in a real world environment. *Vis. Cogn.* 17, 1159–1184. doi: 10.1080/13506280902797125

Dziobek, I., Fleck, S., Kalbe, E., Rogers, K., Hassenstab, J., Brand, M., et al. (2006). Introducing MASC: a movie for the assessment of social cognition. *J. Autism Dev. Disord.* 36, 623–636. doi: 10.1007/s10803-006-0107-0

Eckert, H. (2017). Erzeugung von Blickreizen virtueller Charaktere mit ambiger kommunikativer Absicht mittels systematischer Variierung zweier Faktoren einer Blickbewegung - Anfangsblick und Blickziel (Medical dissertation, University of Cologne). University of Cologne.

Einav, S., and Hood, B. M. (2006). Children's use of the temporal dimension of gaze for inferring preference. *Dev. Psychol.* 42, 142–152. doi: 10.1037/0012-1649.42.1.142

Fedor, J., Lynn, A., Foran, W., DiCicco-Bloom, J., Luna, B., and O'Hearn, K. (2018). Patterns of fixation during face recognition: differences in autism across age. *Autism* 22, 866–880. doi: 10.1177/1362361317714989

Ferreira, F., Apel, J., and Henderson, J. M. (2008). Taking a new look at looking at nothing. *Trends Cogn. Sci.* 12, 405–410. doi: 10.1016/j.tics.2008.07.007

Féry, C., and Kügler, F. (2008). Pitch accent scaling on given, new and focused constituents in German. *J. Phon.* 36, 680–703. doi: 10.1016/j.wocn.2008.05.001

FFmpeg Developers (2018). FFmpeg Tool [Software]. Available at: http://ffmpeg.org/

Fletcher-Watson, S., Leekam, S. R., Benson, V., Frank, M. C., and Findlay, J. M. (2009). Eye-movements reveal attention to social information in autism spectrum disorder. *Neuropsychologia* 47, 248–257. doi: 10.1016/j.neuropsychologia.2008.07.016 (Accessed June 15, 2019).

Fraundorf, S. H., Watson, D. G., and Benjamin, A. S. (2010). Recognition memory reveals just how CONTRASTIVE contrastive accenting really is. *J. Mem. Lang.* 63, 367–386. doi: 10.1016/j.jml.2010.06.004

Fraundorf, S. H., Watson, D. G., and Benjamin, A. S. (2012). The effects of age on the strategic use of pitch accents in memory for discourse: a processing-resource account. *Psychol. Aging* 27, 88–98. doi: 10.1037/a0024138

Freeth, M., Chapman, P., Ropar, D., and Mitchell, P. (2010). Do gaze cues in complex scenes capture and direct the attention of high functioning adolescents with ASD? Evidence from eye-tracking. *J. Autism Dev. Disord.* 40, 534–547. doi: 10.1007/s10803-009-0893-2

Freire, A., Eskritt, M., and Lee, K. (2004). Are eyes windows to a Deceiver's soul? Children's use of Another's eye gaze cues in a deceptive situation. *Dev. Psychol.* 40, 1093–1104. doi: 10.1037/0012-1649.40.6.1093

Frith, C. D., and Frith, U. (2006). The neural basis of mentalizing. *Neuron* 50, 531–534. doi: 10.1016/j.neuron.2006.05.001

Frith, U., Morton, J., and Leslie, A. M. (1991). The cognitive basis of a biological disorder: autism. *Trends Neurosci.* 14, 433–438. doi: 10.1016/0166-2236(91)90041-R

Genzel, S., Kerkhoff, G., and Scheffter, S. (1995). PC-gestützte Standardisierung des Bildmaterials von Snodgrass & Vanderwart (1980): I. Deutschsprachige Normierung. *Neurolinguistik* 9, 41–53.

Gernsbacher, M. A., and Yergeau, M. (2019). Empirical failures of the claim that autistic people lack a theory of mind. *Arch. Sci. Psychol.* 7, 102–118. doi: 10.1037/arc0000067

Globerson, E., Amir, N., Kishon-Rabin, L., and Golan, O. (2015). Prosody recognition in adults with high-functioning autism spectrum disorders: from psychoacoustics to cognition. *Autism Res.* 8, 153–163. doi: 10.1002/aur.1432

Golan, O., Baron-Cohen, S., Hill, J. J., and Rutherford, M. D. (2007). The "Reading the mind in the voice" test-revised: a study of complex emotion recognition in adults with and without autism spectrum conditions. *J. Autism Dev. Disord.* 37, 1096–1106. doi: 10.1007/s10803-006-0252-5

Good, P. (2013). Permutation tests: A practical guide to resampling methods for testing hypotheses (springer series in statistics) - Good, Phillip: 9783540940975 - ZVAB. Berlin and Heidelberg: Springer-Verlag.

Grice, M., and Baumann, S. (2007). "An introduction to intonation – functions and models" in Non-native prosody (Berlin: De Gruyter Mouton), 25–52.

Grice, M., Baumann, S., and Benzmüller, R. (2005). German intonation in autosegmental-metrical phonology - Oxford scholarship. *Prosodic Typol.* 1, 55–83. doi: 10.1093/acprof:oso/9780199249633.003.0003

Grice, M., Krüger, M., and Vogeley, K. (2016). Adults with Asperger syndrome are less sensitive to intonation than control persons when listening to speech. *Cult. Brain* 4, 38–50. doi: 10.1007/s40167-016-0035-6

Grice, M., Ritter, S., Niemann, H., and Roettger, T. B. (2017). Integrating the discreteness and continuity of intonational categories. *J. Phon.* 64, 90–107. doi: 10.1016/j.wocn.2017.03.003

Grice, M., Wehrle, S., Krüger, M., Spaniol, M., Cangemi, F., and Vogeley, K. (2023). Linguistic prosody in autism spectrum disorder—an overview. *Lang. Linguist. Compass* 17:e12498. doi: 10.1111/lnc3.12498

Griffin, J. W., and Scherf, K. S. (2020). Does decreased visual attention to faces underlie difficulties interpreting eye gaze cues in autism? *Mol. Autism.* 11:60. doi: 10.1186/s13229-020-00361-2

Gronau, Q. F., Singmann, H., and Wagenmakers, E. (2020). Bridgesampling: an R package for estimating normalizing constants. *J. Stat. Softw.* 92, 1–29. doi: 10.18637/jss.v092.i10

Grossman, R. B., Zane, E., Mertens, J., and Mitchell, T. (2019). Facetime vs. Screentime: gaze patterns to live and video social stimuli in adolescents with ASD. *Sci. Rep.* 9:12643. doi: 10.1038/s41598-019-49039-7

Guillon, Q., Hadjikhani, N., Baduel, S., and Rogé, B. (2014). Visual social attention in autism spectrum disorder: insights from eye tracking studies. *Neurosci. Biobehav. Rev.* 42, 279–297. doi: 10.1016/j.neubiorev.2014.03.013

Happé, F. G. E. (1994). An advanced test of theory of mind: understanding of story characters' thoughts and feelings by able autistic, mentally handicapped, and normal children and adults. *J. Autism Dev. Disord.* 24, 129–154. doi: 10.1007/BF02172093

Henderson, J. M., Williams, C. C., and Falk, R. J. (2005). Eye movements are functional during face learning. *Mem. Cogn.* 33, 98–106. doi: 10.3758/BF03195300

Hernandez, N., Metzger, A., Magné, R., Bonnet-Brilhault, F., Roux, S., Barthelemy, C., et al. (2009). Exploration of core features of a human face by healthy and autistic adults analyzed by visual scanning. *Neuropsychologia* 47, 1004–1012. doi: 10.1016/j.neuropsychologia.2008.10.023

Hesling, I., Dilharreguy, B., Peppé, S., Amirault, M., Bouvard, M., and Allard, M. (2010). The integration of prosodic speech in high functioning autism: a preliminary fMRI study. *PLoS One* 5:e11571. doi: 10.1371/journal.pone.0011571

Hobson, R. P., Ouston, J., and Lee, A. (1988). What's in a face? The case of autism. *Br. J. Psychol.* 79, 441–453. doi: 10.1111/j.2044-8295.1988.tb02745.x

Hurley, R., and Bishop, J. (2016). Prosodic and individual influences on the interpretation of only. *Speech Prosody*. 8, 193–197. doi: 10.21437/SpeechProsody.2016-40

Itier, R. J., Villate, C., and Ryan, J. D. (2007). Eyes always attract attention but gaze orienting is task-dependent: evidence from eye movement monitoring. *Neuropsychologia* 45, 1019–1028. doi: 10.1016/j.neuropsychologia.2006.09.004

Ito, K., Kryszak, E., and Ibanez, T. (2022). Effect of prosodic emphasis on the processing of joint-attention cues in children with ASD. Lisbon, Portugal: ISCA. 110–114.

Ito, K., and Speer, S. R. (2008). Anticipatory effects of intonation: eye movements during instructed visual search. *J. Mem. Lang.* 58, 541–573. doi: 10.1016/j.jml.2007.06.013

Jording, M., Engemann, D., Eckert, H., Bente, G., and Vogeley, K. (2019a). Distinguishing social from private intentions through the passive observation of gaze cues. *Front. Hum. Neurosci.* 13:442. doi: 10.3389/fnhum.2019.00442

Jording, M., Hartz, A., Bente, G., Schulte-Rüther, M., and Vogeley, K. (2019b). Inferring interactivity from gaze patterns during triadic person-object-agent interactions. *Front. Psychol.* 10:1913. doi: 10.3389/fpsyg.2019.01913

Klami, A. (2010). "Inferring task-relevant image regions from gaze data" in 2010 IEEE international workshop on machine learning for signal processing, Eds. S.l. Kaski, D. J. Miller, E. Oja, A. Honkela (IEEE). 101–106.

Klami, A., Saunders, C., de Campos, T. E., and Kaski, S. (2008). "Can relevance of images be inferred from eye movements?" in Proceedings of the 1st ACM international conference on Multimedia information retrieval (New York, NY, USA: Association for Computing Machinery), 134–140.

Kleinman, J., Marciano, P. L., and Ault, R. L. (2001). Advanced theory of mind in high-functioning adults with autism. *J. Autism Dev. Disord.* 31, 29–36. doi: 10.1023/a:1005657512379

Klin, A., Jones, W., Schultz, R., Volkmar, F., and Cohen, D. (2002). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry* 59, 809–816. doi: 10.1001/archpsyc.59.9.809

Konkle, T., Brady, T. F., Alvarez, G. A., and Oliva, A. (2010). Conceptual distinctiveness supports detailed visual long-term memory for real-world objects. *J. Exp. Psychol. Gen.* 139, 558–578. doi: 10.1037/a0019165

Krahmer, E., Ruttkay, Z., Swerts, M., and Wesselink, W. (2002). "Perceptual evaluation of audiovisual cues for prominence" in INTERSPEECH. Denver, Colorado, USA: ISCA.

Kuzmanovic, B., Schilbach, L., Lehnhardt, F.-G., Bente, G., and Vogeley, K. (2011). A matter of words: impact of verbal and nonverbal information on impression formation in high-functioning autism. *Res. Autism Spectr. Disord.* 5, 604–613. doi: 10.1016/j.rasd.2010.07.005

Leding, J. K. (2020). Animacy and threat in recognition memory. *Mem. Cogn.* 48, 788–799. doi: 10.3758/s13421-020-01017-5

Lee, K., Eskritt, M., Symons, L. A., and Muir, D. (1998). Children's use of triadic eye gaze information for "mind reading.". *Dev. Psychol.* 34, 525–539. doi: 10.1037//0012-1649.34.3.525

Lee, M. D., and Wagenmakers, E.-J. (2014). Bayesian cognitive modeling: a practical course. Cambridge: Cambridge University Press.

Lenth, R. V., Bürkner, P., Herve, M., Love, J., Riebl, H., and Singmann, H. (2021). Emmeans: estimated marginal means, aka least-squares means. Available at: https://cran.r-project.org/web/packages/emmeans/index.html (Accessed February 3, 2023).

Macdonald, R. G., and Tatler, B. W. (2013). Do as eye say: gaze cueing and language in a real-world social interaction. *J. Vis.* 13:6. doi: 10.1167/13.4.6

Makowski, D., Ben-Shachar, M. S., and Lüdecke, D. (2019). bayestestR: Describing effects and their uncertainty, existence and significance within the Bayesian framework. *JOSS*, 4, 1541. doi: 10.21105/joss.01541

Martini, P., and Maljkovic, V. (2009). Short-term memory for pictures seen once or twice. *Vis. Res.* 49, 1657–1667. doi: 10.1016/j.visres.2009.04.007

Melcher, D. (2001). Persistence of visual memory for scenes. *Nature* 412, 401. doi: 10.1038/35086646

Melcher, D. (2006). Accumulation and persistence of memory for natural scenes. *J. Vis.* 6, 2–17. doi: 10.1167/6.1.2

Mixdorff, H., Hönemann, A., and Fagel, S. (2013). Integration of acoustic and visual cues in prominence perception. Proceedings of AVSP 2013. Available at: https://pub.uni-bielefeld.de/record/2752439 (Accessed August 16, 2023).

Moubayed, S., and Beskow, J. (2009). Effects of visual prominence cues on speech intelligibility. *Proc Int Conf Auditory Visual Speech Process.* (Norwich, UK) 9:16.

Müller, N., Baumeister, S., Dziobek, I., Banaschewski, T., and Poustka, L. (2016). Validation of the movie for the assessment of social cognition in adolescents with ASD: fixation duration and pupil dilation as predictors of performance. *J. Autism Dev. Disord.* 46, 2831–2844. doi: 10.1007/s10803-016-2828-z

Munhall, K. G., Jones, J., Callan, D., Kuratate, T., and Vatikiotis-Bateson, E. (2004). Visual prosody and speech intelligibility head movement improves auditory speech perception. *Psychol. Sci.* 15, 133–137. doi: 10.1111/j.0963-7214.2004.01502010.x

Nakano, T., Tanaka, K., Endo, Y., Yamane, Y., Yamamoto, T., Nakano, Y., et al. (2010). Atypical gaze patterns in children and adults with autism spectrum disorders dissociated from developmental changes in gaze behaviour. *Proc. Biol. Sci.* 277, 2935–2943. doi: 10.1098/rspb.2010.0587

O'Connor, K. (2012). Auditory processing in autism spectrum disorder: a review. *Neurosci. Biobehav. Rev.* 36, 836–854. doi: 10.1016/j.neubiorev.2011.11.008

O'Hearn, K., Tanaka, J., Lynn, A., Fedor, J., Minshew, N., and Luna, B. (2014). Developmental plateau in visual object processing from adolescence to adulthood in autism. *Brain Cogn.* 90, 124–134. doi: 10.1016/j.bandc.2014.06.004

Odén, A., and Wedel, H. (1975). Arguments for Fisher's permutation test. *Ann. Stat.* 3, 518–520. doi: 10.1214/aos/1176343082

Pantelis, P. C., and Kennedy, D. P. (2017). Deconstructing atypical eye gaze perception in autism spectrum disorder. *Sci. Rep.* 7:14990. doi: 10.1038/s41598-017-14919-3

Pelphrey, K. A., Sasson, N. J., Reznick, J. S., Paul, G., Goldman, B. D., and Piven, J. (2002). Visual scanning of faces in autism. *J. Autism Dev. Disord.* 32, 249–261. doi: 10.1023/A:1016374617369

Pesarin, F., and Salmaso, L. (2010). The permutation testing approach: a review. *Statistica* 70, 481–509. doi: 10.6092/issn.1973-2201/3599

Pfeiffer-Lessmann, N., Pfeiffer, T., and Wachsmuth, I. (2012). An operational model of joint attention - timing of gaze patterns in interactions between humans and a virtual human., in Proceedings of the 34th annual conference of the Cognitive Science Society, 851–856.

Ponnet, K. S., Roeyers, H., Buysse, A., De Clercq, A., and Van der Heyden, E. (2004). Advanced mind-reading in adults with Asperger syndrome. *Autism* 8, 249–266. doi: 10.1177/1362361304045214

Prieto, P., Puglesi, C., Borràs-Comes, J., Arroyo, E., and Blat, J. (2015). Exploring the contribution of prosody and gesture to the perception of focus using an animated agent☆. *J. Phon.* 49, 41–54. doi: 10.1016/j.wocn.2014.10.005

R Core Team (2019). R: A language and environment for statistical computing. Available at: https://www.R-project.org (Accessed May 30, 2019).

R Core Team (2023). R: A Language and Environment for Statistical Computing. Available at: https://www.R-project.org/ (Accessed April 10, 2023).

Remington, A., and Fairnie, J. (2017). A sound advantage: increased auditory capacity in autism. *Cognition* 166, 459–465. doi: 10.1016/j.cognition.2017.04.002

Riby, D. M., Hancock, P. J., Jones, N., and Hanley, M. (2013). Spontaneous and cued gaze-following in autism and Williams syndrome. *J. Neurodevelop. Disord.* 5:13. doi: 10.1186/1866-1955-5-13

Riddiford, J. A., Enticott, P. G., Lavale, A., and Gurvich, C. (2022). Gaze and social functioning associations in autism spectrum disorder: a systematic review and meta-analysis. *Autism Res.* 15, 1380–1446. doi: 10.1002/aur.2729

Ring, M., Gaigg, S. B., and Bowler, D. M. (2015). Object-location memory in adults with autism spectrum disorder. *Autism Res.* 8, 609–619. doi: 10.1002/aur.1478

Ristic, J., Mottron, L., Friesen, C. K., Iarocci, G., Burack, J. A., and Kingstone, A. (2005). Eyes are special but not for everyone: the case of autism. *Cogn. Brain Res.* 24, 715–718. doi: 10.1016/j.cogbrainres.2005.02.007

Rosenblau, G., Kliemann, D., Dziobek, I., and Heekeren, H. R. (2017). Emotional prosody processing in autism spectrum disorder. *Soc. Cogn. Affect. Neurosci.* 2, 224–239. doi: 10.1093/scan/nsw118

Rossion, B., and Pourtois, G. (2004). Revisiting Snodgrass and Vanderwart's object pictorial set: the role of surface detail in basic-level object recognition. *Perception* 33, 217–236. doi: 10.1068/p5117

Roy, J., Cole, J., and Mahrt, T. (2017). Individual differences and patterns of convergence in prosody perception. *Lab. Phonol.* 8:22. doi: 10.5334/labphon.108

RStudio Team (2016). RStudio: Integrated Development for R. Available at: http://www.rstudio.com (Accessed April 25, 2019).

Rutherford, M. D., Baron-Cohen, S., and Wheelwright, S. (2002). Reading the mind in the voice: a study with normal adults and adults with Asperger syndrome and high functioning autism. *J. Autism Dev. Disord.* 32, 189–194. doi: 10.1023/a:1015497629971

Scheeren, A. M., Rosnay, M.De, Koot, H. M., and Begeer, S. (2013). Rethinking theory of mind in high-functioning autism spectrum disorder. *J. Child Psychol. Psychiatry* 54, 628–635. doi: 10.1111/jcpp.12007

Schelinski, S., Roswandowitz, C., and von Kriegstein, K. (2017). Voice identity processing in autism spectrum disorder. *Autism Res.* 10, 155–168. doi: 10.1002/aur.1639

Schelinski, S., and von Kriegstein, K. (2019). The relation between vocal pitch and vocal emotion recognition abilities in people with autism Spectrum disorder and typical development. *J. Autism Dev. Disord.* 49, 68–82. doi: 10.1007/s10803-018-3681-z

Schneider, D., Slaughter, V. P., Bayliss, A. P., and Dux, P. E. (2013). A temporally sustained implicit theory of mind deficit in autism spectrum disorders. *Cognition* 129, 410–417. doi: 10.1016/j.cognition.2013.08.004

Schuwerk, T., Vuori, M., and Sodian, B. (2015). Implicit and explicit theory of mind reasoning in autism spectrum disorders: the impact of experience. *Autism* 19, 459–468. doi: 10.1177/1362361314526004

Senju, A., Southgate, V., White, S., and Frith, U. (2009). Mindblind eyes: an absence of spontaneous theory of mind in Asperger syndrome. *Science* 325, 883–885. doi: 10.1126/science.1176170

Setien-Ramos, I., Lugo-Marín, J., Gisbert-Gustemps, L., Díez-Villoria, E., Magán-Maganto, M., Canal-Bedia, R., et al. (2022). Eye-tracking studies in adults with autism Spectrum disorder: a systematic review and Meta-analysis. *J. Autism Dev. Disord.* 53, 2430–2443. doi: 10.1007/s10803-022-05524-z

Shimojo, S., Simion, C., Shimojo, E., and Scheier, C. (2003). Gaze bias both reflects and influences preference. *Nat. Neurosci.* 6, 1317–1322. doi: 10.1038/nn1150

Stewart, M. E., McAdam, C., Ota, M., Peppé, S., and Cleland, J. (2013). Emotional recognition in autism spectrum conditions from voices and faces. *Autism* 17, 6–14. doi: 10.1177/1362361311424572

Swerts, M., and Krahmer, E. (2008). Facial expression and prosodic prominence: effects of modality and facial area. *J. Phon.* 36, 219–238. doi: 10.1016/j.wocn.2007.05.001

Theeuwes, J., Belopolsky, A., and Olivers, C. N. L. (2009). Interactions between working memory, attention and eye movements. *Acta Psychol.* 132, 106–114. doi: 10.1016/j.actpsy.2009.01.005

Tottenham, N., Hertzig, M. E., Gillespie-Lynch, K., Gilhooly, T., Millner, A. J., and Casey, B. J. (2014). Elevated amygdala response to faces and gaze aversion in autism spectrum disorder. *Soc. Cogn. Affect. Neurosci.* 9, 106–117. doi: 10.1093/scan/nst050

Wahl, S., Marinović, V., and Träuble, B. (2019). Gaze cues of isolated eyes facilitate the encoding and further processing of objects in 4-month-old infants. *Dev. Cogn. Neurosci.* 36:100621. doi: 10.1016/j.dcn.2019.100621

Wang, S., Jiang, M., Duchesne, X. M., Laugeson, E. A., Kennedy, D. P., Adolphs, R., et al. (2015). Atypical visual saliency in autism Spectrum disorder quantified through model-based eye tracking. *Neuron* 88, 604–616. doi: 10.1016/j.neuron.2015.09.042

Wang, L., Pfordresher, P. Q., Jiang, C., and Liu, F. (2021). Individuals with autism spectrum disorder are impaired in absolute but not relative pitch and duration matching in speech and song imitation. *Autism Res.* 14, 2355–2372. doi: 10.1002/aur.2569

Watson, D. G., Tanenhaus, M. K., and Gunlogson, C. A. (2008). Interpreting pitch accents in online comprehension: H* vs. L+H*. *Cogn. Sci.* 32, 1232–1244. doi: 10.1080/03640210802138755

Weber, A., Braun, B., and Crocker, M. W. (2006). Finding referents in time: eye-tracking evidence for the role of contrastive accents. *Lang. Speech* 49, 367–392. doi: 10.1177/00238309060490030301

White, S. J., Coniston, D., Rogers, R., and Frith, U. (2011). Developing the Frith-Happé animations: a quick and objective test of theory of mind for adults with autism. *Autism Res.* 4, 149–154. doi: 10.1002/aur.174

Winn, M. (2014). Fade in, Fade out [Praat script]. Available at: http://www.mattwinn.com/praat/RampOnsetAndOrOffset.txt (Accessed April 13, 2020).

World Medical Association (2013). World medical association declaration of Helsinki: ethical principles for medical research involving human subjects. *JAMA* 310, 2191–2194. doi: 10.1001/jama.2013.281053

Yarbus, A. L. (1967). "Eye movements during perception of complex objects" in Eye movements and vision. ed. A. L. Yarbus (Boston, MA: Springer US), 171–211. doi: 10.1007/978-1-4899-5379-7_8

Zhang, M., Xu, S., Chen, Y., Lin, Y., Ding, H., and Zhang, Y. (2022). Recognition of affective prosody in autism spectrum conditions: a systematic review and meta-analysis. *Autism* 26, 798–813. doi: 10.1177/1362361321995725

Zimmermann, J. T., Wehrle, S., Cangemi, F., Grice, M., and Vogeley, K. (2020). Listeners and lookers: using pitch height and gaze duration for inferring mental states., in Proceedings of the 10th International Conference on Speech Prosody 2020, Tokyo, Japan.