


Research Article

Vocal and Facial Behavior During Affect Production in Autism Spectrum Disorder

Hardik Kothare,^a  Vikram Ramanarayanan,^{a,b}  Michael Neumann,^a Jackson Liscombe,^a Vanessa Richter,^a Linnea Lampinen,^b Alison Bai,^b Cristian Preciado,^b Katherine Brogan,^b and Carly Demopoulos^b

^aModality.AI, Inc., San Francisco, CA ^bUniversity of California, San Francisco

ARTICLE INFO

Article History:

Received January 31, 2023

Revision received August 4, 2023

Accepted October 8, 2024

Editor-in-Chief: Cara E. Stepp

Editor: Ben A. M. Maassen

https://doi.org/10.1044/2024_JSLHR-23-00080

ABSTRACT

Purpose: We investigate the extent to which automated audiovisual metrics extracted during an affect production task show statistically significant differences between a cohort of children diagnosed with autism spectrum disorder (ASD) and typically developing controls.

Method: Forty children with ASD and 21 neurotypical controls interacted with a multimodal conversational platform with a virtual agent, Tina, who guided them through tasks prompting facial and vocal communication of four emotions—happy, angry, sad, and afraid—under conditions of high and low verbal and social cognitive task demands.

Results: Individuals with ASD exhibited greater standard deviation of the fundamental frequency of the voice with the minima and maxima of the pitch contour occurring at an earlier time point as compared to controls. The intensity and voice quality of emotional speech were also different between the two cohorts in certain conditions. Additionally, facial metrics capturing the acceleration of the lower lip, lip width, eye opening, and vertical displacement of the eyebrows were also important markers to distinguish between children with ASD and neurotypical controls. Both facial and speech metrics performed well above chance in group classification accuracy.

Conclusion: Speech acoustic and facial metrics associated with affect production were effective in distinguishing between children with ASD and neurotypical controls.

Supplemental Material: <https://doi.org/10.23641/asha.28027796>

Autism spectrum disorder (ASD) is a neurodevelopmental disorder (American Psychiatric Association, 2013) with an estimated overall prevalence of one in 36 children aged 8 years in the United States (Maenner, 2023). A defining feature of ASD, according to the *Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5)* criteria, is impairment in nonverbal communicative behaviors used for social interaction (American Psychiatric Association, 2013) such as facial expression, which

can manifest as absent/minimal, more intense/exaggerated, poorly integrated, or inappropriate to context. Vocal features such as fundamental frequency (F_0) and prosody, which carry emotion-specific information (Nussbaum et al., 2022), have been described as atypical when produced by children with ASD (Nadig & Shaw, 2012). Facial expressions of emotion have also been characterized as less natural and more intense in individuals with ASD (Faso et al., 2015). Prior studies have reported atypical production of vocal and facial affect during emotional speech (Hubbard et al., 2017; Loveland et al., 1994) and poor cross-modal coordination between facial expression and emotional speech production in ASD (Sorensen et al., 2019).

The feasibility and potential clinical utility of speech biomarkers for automated assessment of atypical vocal and facial expression in ASD and other neurodevelopmental disorders have been established by prior work

Correspondence to Hardik Kothare: hardik.kothare@modality.ai.

Disclosure: Hardik Kothare, Vikram Ramanarayanan, Michael Neumann, Jackson Liscombe, and Vanessa Richter are full-time salaried employees of Modality.AI, Inc., and hold ownership interest in the company. Carly Demopoulos serves on the Scientific Advisory Board of Modality.AI, Inc., and holds ownership interest in the company. All other authors have declared that no competing financial or nonfinancial interests existed at the time of publication.

(Ramanarayanan et al., 2022). Acoustic–phonetic and lexical features extracted from short, unstructured conversations accurately identified the diagnostic status of children with ASD 66% of the time and that of typically developing children 86% of the time (Cho et al., 2019). Automated assessment of prelinguistic vocalizations has also been shown to be helpful in predicting future diagnosis of ASD (Pokorny et al., 2017). Atypical facial expressions, defined by facial action unit (FAU) intensities, in ASD can be automatically quantified through the use of computer vision and FAU intensities (Leo et al., 2018). FAUs are the components of facial expressions defined by the movement of a muscle or set of muscles (Ekman & Friesen, 1978). FAU intensities are a way to quantify emotional expressions through temporal and geometric analysis of FAUs. When individuals with ASD were cued to mimic facial expressions that carried either positive or negative valence (expressions that were inferred to carry positive or negative values), the degree of facial movements did not depend on the emotional valence and the movements were fleeting, exaggerated, and jerky (Zane et al., 2019) as compared to a control group. Prior work has demonstrated the feasibility and utility of computer vision in a standalone or multimodal framework in ASD research (Bangerter et al., 2020; Samad et al., 2017; Sorensen et al., 2019). Given that there are no standardized measures of facial or vocal affect production ability currently available, these automated, objective measurements have potential clinical utility in quantifying domain-specific nonverbal communicative behavior.

Our prior work has demonstrated the utility of a cloud-based multimodal conversational platform (Ramanarayanan et al., 2023, 2024; Suendermann-Oeft et al., 2019) that uses a virtual human guide, Tina, to conduct self-driven assessments that elicit speech and facial expressions through a variety of tasks for detection and progress monitoring of various neurological and mental health disorders like amyotrophic lateral sclerosis (Neumann et al., 2021, 2024), depression (Neumann et al., 2020), Parkinson’s disease (Kothare et al., 2022), and schizophrenia (Richter et al., 2022). During an interactive session with Tina, analytic modules extract objective metrics in real time that can be accessed by researchers or clinicians through a user-friendly dashboard. In prior work in ASD (Kothare et al., 2021), we showed that atypical affect production, measured using a novel affect production task (APT), correlates with accuracy in recognition of vocal and facial affect in children with ASD. Furthermore, we identified a positive correlation between jaw kinematic measures and the motor speed of the dominant hand, which supports the hypothesis that there is a coupling between speech motor coordination and fine motor skills in ASD (Talkar et al., 2020).

Building on this foundation, the current work aims to identify facial and vocal markers that show significant differences between children with ASD and neurotypical controls (NTCs). Objective audiovisual metrics of affect production in ASD may be used to quantify expressive aspects of nonverbal communication. Impairment in nonverbal communication is one of the diagnostic criteria for ASD (American Psychiatric Association, 2013). As such, quantification of this symptom domain also has potential clinical utility in tracking clinical presentation over time or in response to interventions. This is particularly salient, as objective measures of symptom presentation in ASD are lacking. As such, clinical trials currently must rely on subjective observation and informant report measures.

We leverage the aforementioned objective multimodal metrics to answer the following research questions in this work:

1. *Effect of cohort:* Which vocal and facial metrics associated with affect production show significant differences between children with ASD and NTCs? Are cohort differences due to actual affective communication differences rather than due to general nonaffective differences in facial and vocal expression?
2. *Effect of production task demands:* Can these differences be captured in shorter utterances to make the task more accessible to individuals with lower verbal ability? Can these differences be captured without provision of additional emotional context (i.e., illustrated narrative describing an emotional situation)?
3. *Effect of produced emotion:* How do these objective metrics vary across emotions produced (happy, sad, angry, afraid)?
4. *Group differences in vocal imitation:* Are children with ASD as sensorimotorically capable of accurately repeating monosyllabic productions to convey emotions as their NTC peers?
5. *Objective classification accuracy:* How effective are the metrics in classifying the two groups—ASD and NTC?

Method

Participants

The study was approved by the institutional review board of the University of California, San Francisco (UCSF IRB Approval 11-05249 and 21-33613). Informed consent from the participants’ guardians and written assent from the participants were obtained prior to enrollment. The study was conducted onsite at the University of

California, San Francisco. Data from 40 participants with ASD (14 female, mean age \pm standard deviation = 12.50 \pm 2.68 years) and 21 NTC participants (11 female, mean age \pm standard deviation = 12.52 \pm 2.88 years) who completed an interactive session on the cloud-based multimodal dialogue platform (see Table 1) between December 2019 and December 2022 were included in the analysis. Inclusion criteria for the NTC group were: no neurological or psychiatric diagnosis and a Social Communication Questionnaire score in the nonclinical range (Rutter et al., 2003). To minimize differences across participants and cohorts, these sessions were conducted in the same controlled environment on the same device (a MacBook Pro with an Intel Core i7 processor) in the presence of a clinical psychology doctoral student who accompanied the participant in the testing room to help with any technical difficulties and provide behavioral support during data collection (e.g., redirecting attention during breaks). Diagnoses in the ASD cohort were confirmed according to *DSM-5* criteria by a licensed clinical psychologist (author C.D.) who established research reliability on the Autism Diagnostic Observation Schedule–Second Edition (ADOS-2; Lord et al., 2000) and the Autism Diagnostic Interview–Revised (ADI-R; Lord et al., 1994). Information obtained from the ADI-R and the ADOS-2 (used as an observational tool only, as scoring was not possible due to deviation from standard administration because of COVID-19 masking mandates) was used to inform diagnostic determinations along with parent report measures of social, emotional, behavioral, and adaptive functioning via the Behavior Assessment System for Children–Third Edition (Reynolds & Kamphaus, 2015), performance-based measures of language skills via the Clinical Evaluation of Language Fundamentals–Fifth Edition (CELF-5;

Wiig et al., 2013), and general intellectual abilities on the Wechsler Intelligence Scale for Children (WISC; Wechsler, 2014) and the Test of Nonverbal Intelligence–Fourth Edition (TONI-4; Brown et al., 2010). Standardized test scores are included in Table 1 and Figure 1.

The ASD cohort had a lower average score (see Table 1 and Figure 1) on the WISC Full-Scale IQ (t test; $t = -3.25$, $p = .0019$), the CELF-5 Expressive Language Index (t test; $t = -2.30$, $p = .0249$), and the CELF-5 Receptive Language Index (t test; $t = -2.83$, $p = .0064$) but not on the TONI-4 Nonverbal IQ (t test; $t = -1.78$, $p = .0805$).

Task

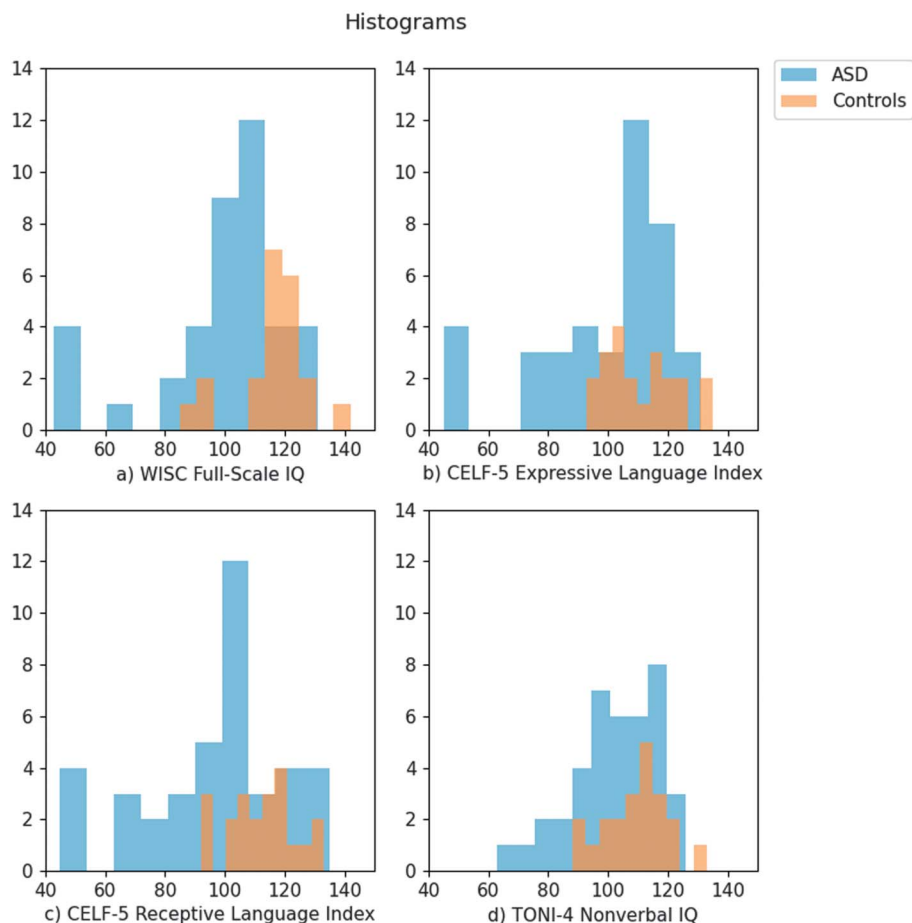
The APTs presented in the interactive session asked the participants to produce one of four emotions—happy, sad, angry, and afraid—through the subtasks listed below. All tasks are performed under directed conditions in which the emotion the participant is expected to communicate is explicitly stated, with the exception of the imitation task in which the emotion is not specified and the participant is simply asked to mimic each stimulus. The session begins with a speaker test, background noise check, and a microphone test. The speaker test determines if the participant is able to hear sounds. The virtual guide, Tina, says a number from zero to nine, and the participant is asked to enter the number in a text field. The background noise measures ambient background noise in decibels while the participant remains silent. During the microphone test, the participant is asked to speak, and it is determined whether the intensity of the participant’s speech is at least 40 dB. The participant is asked to ensure that their face is fully visible with no face coverings or

Table 1. Participant demographics.

Variable	ASD	Controls
Sample size	40	21
Sex at birth, female (%)	14 (35)	11 (52%)
Nonbinary/trans (%)	4 (10%)	1 (4.76%)
Mean age (SD) in years	12.5 (2.68)	12.52 (2.88)
African American (%)	1 (2.5%)	0 (0%)
Asian (%)	5 (12.5%)	3 (14%)
Caucasian (%)	20 (50%)	11 (52%)
Hispanic (%)	3 (7.5%)	1 (5%)
Multiracial (%)	11 (27.5%)	6 (29%)
Wechsler Intelligence Scale for Children Full-Scale IQ ($M \pm SD$; range)	99.25 \pm 21.93; 43–131	116.19 \pm 13.00; 85–142
CELF-5 Expressive Language Index ($M \pm SD$; range)	98.98 \pm 22.75; 45–131	111.24 \pm 12.01; 93–135
CELF-5 Receptive Language Index ($M \pm SD$; range)	96.75 \pm 24.23; 45–135	112.57 \pm 11.27; 92–133
TONI-4 Nonverbal IQ ($M \pm SD$; range)	102.68 \pm 14.36; 63–126	109.10 \pm 11.29; 88–133

Note. ASD = autism spectrum disorder group; CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; TONI-4 = Test of Nonverbal Intelligence–Fourth Edition.

Figure 1. Histograms of scores: (a) WISC Full-Scale IQ, (b) CELF-5 Expressive Language Index, (c) CELF-5 Receptive Language Index, and (d) TONI-4 Nonverbal IQ. WISC = Wechsler Intelligence Scale for Children; CELF-5 = Clinical Evaluation of Language Fundamentals–Fifth Edition; TONI-4 = Test of Nonverbal Intelligence–Fourth Edition; ASD = autism spectrum disorder group.



shadows obscuring their face. Participants are also asked, “How do you identify?” with four options to choose from: a boy, a girl, nonbinary, or other than a boy or a girl. Tina then welcomes the participant to the session.

- Task 1. Noncontextual monosyllabic emotion production (eight prompts, two prompts per emotion): An instruction video recorded by an actor in a neutral voice and facial expression is played for the child. It begins with the actor saying, “This activity measures how you can communicate the way you are feeling by the way your face looks and your voice sounds.” The actor then lists the four emotions they will be asked to communicate. The video continues by explaining that the word they will be asked to say is “oh” (/oo/): “The way you say it, and the way your face looks has to make me understand what feeling you are trying to communicate.” The actor gives an example with the emotion “disgusted,” an affective

expression the participant will not be asked to produce for any of the tested turns. The video ends with the actor reminding the child to use their face and voice to communicate the prompted affect. After each video prompt (e.g., “Use your face and voice to say ‘oh’ in a way that seems happy”), the participant produces the word “oh” in a manner that, to the best of their ability, conveys the emotion specified in the prompt (i.e., happy, sad, angry, afraid). The purpose of this task is to assess the ability of the participant to produce a monosyllabic utterance that conveys a specified affective meaning.

- Task 2. Noncontextual monosyllabic vocal imitation (16 prompts, four prompts per emotion): The virtual agent instructs the participant: “Now I want you to listen to the way the person says ‘oh’ on this recording and try to repeat it in the same way. Try to sound exactly the way it sounds on the recording.” There is no visual stimulus or emotion label

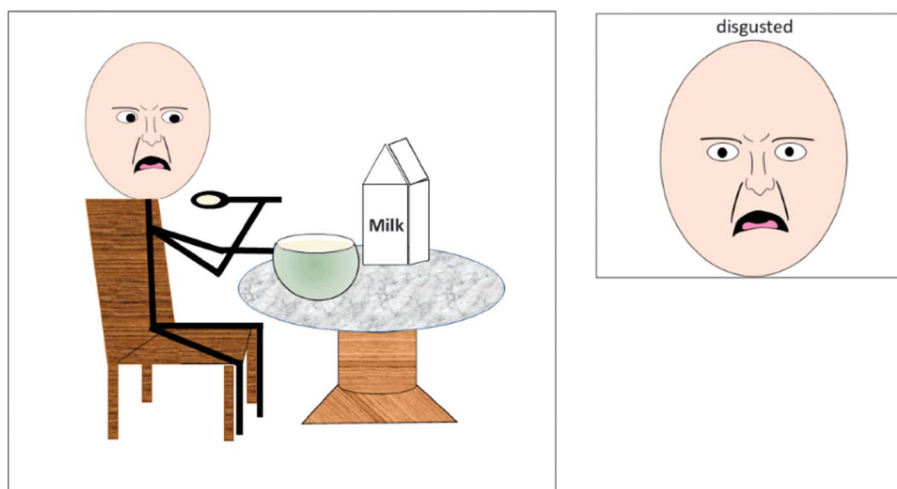
presented in this condition. After each audio recording is played, the participant imitates the recording to the best of their ability. Through this imitation, we can assess whether the participant is sensorimotorically capable of producing exemplary emotional vocal expression, independent of social-cognitive processes needed for emotional communication (e.g., the basic ability to produce specific sounds that convey an emotion).

- Task 3. Contextualized monosyllabic emotion production (16 prompts, four prompts per emotion): The virtual agent introduces the next task by saying, “This next activity is all about Jessie. Jessie identifies as a [child’s identified gender] just like you.” Jessie is gender-matched to the participant’s selection in the beginning of the session (boy, girl, neither boy nor girl, both boy and girl). This is in an effort to make the child see Jessie as relatable and as someone who would have similar emotional reactions as well as to prevent the child from attempting to “impersonate” a child of a different gender. The agent continues, “This activity measures how you can communicate the way you are feeling by the way your face looks and the way you say your words.” The virtual agent then explains that she will be telling stories about Jessie, and in response to each story, the child must say the same word “oh” in a way that conveys the emotion that Jessie is feeling. The agent gives the following example with the emotion “disgusted,” which the child will not be asked to produce: A picture illustrating the narrative is shown on the screen, and the virtual agent says, “Suppose Jessie poured some milk in the cereal bowl but didn’t realize it was old. It smelled awful. Jessie felt disgusted and said ‘oh.’” The word “oh” is

produced to sound as though the agent is disgusted. The agent then prompts the child with a new situation and tells them Jessie is surprised, another emotion the child will not have to produce for tested turns. The agent then prompts the child by saying, “Jessie says . . . ,” and the child responds by saying “oh” in a surprised way. The human examiner administering the task may have the child repeat the condition if more practice is required to understand the task before continuing. The child is then presented with single pictured illustrated narratives for each turn. When Tina prompts the child with, “Jessie says . . . ,” the child produces the monosyllable “oh” to convey the specified, situationally appropriate emotion through facial and vocal expression. See Figure 2 for an example illustration. The purpose of this task condition is to assess the ability to produce specified emotions under a given emotional context to help those who may not understand the concept of named emotions out of context.

- Task 4. Noncontextual sentence-length emotion production (eight prompts, two prompts per emotion): An instruction video-recorded by an actor in a neutral voice and facial expression is played for the child. It begins by saying, “This activity measures how you can communicate the way you are feeling by the way your face looks and the way you say sentences,” and lists the four emotions they will be asked to communicate. The video continues by explaining that the sentence they will be asked to say is, “I’ll be right back. The way you say it, and the way your face looks has to make me understand what feeling you are trying to communicate.” The actor gives an example with the emotion “disgusted,” an emotion the participant will not be asked to produce. The

Figure 2. Example picture stimulus for Task 3 to evoke an emotional production of the monosyllable “oh.”



video ends with the actor reminding the child to use their face and voice to communicate the prompted emotion. After each video prompt (e.g., “Use your face and voice to say ‘I’ll be right back’ in a way that seems happy”), the participant produces the sentence “I’ll be right back” in a manner that, to the best of their ability, conveys the emotion specified in the prompt. The sentence “I’ll be right back” was selected for its emotionally neutral semantic context in an effort to parallel the stimulus used in an analogous affect recognition task, the Diagnostic Analysis of Nonverbal Accuracy–Second Edition (DANVA-2; Nowicki & Duke, 1994). The DANVA-2 vocal affect stimuli use the phrase, “I’m going out of the room right now, but I’ll be back later,” conveying happy, sad, angry, and fearful vocal affect. We chose a shortened, simpler sentence for this task in order to make the test more accessible to individuals who may have lower language levels and speak in less complex sentences.

For all the tasks described above, there is an option to repeat each turn, which the examiner would select if the child was not responding to the prompt, not attending to the task, or otherwise had an unusable turn. See Figure 3 for a schematic of the interactive session.

Data Analysis

Extraction of Metrics

Speech audio data were collected by the platform at a sampling rate of 48 kHz. All speech acoustic metrics were extracted using Praat (Boersma & van Heuven, 2001). These metrics were spectral domain metrics ($F0$ [Hz]; jitter [difference of difference of periods, %]; $F1$, $F2$, and $F3$ formant frequencies [Hz]; $F2$ slope [Hz/s]; cepstral peak prominence [CPP; dB]; harmonics-to-noise ratio

[dB]), signal energy metrics (shimmer [%], signal-to-noise ratio [SNR; dB], intensity [dB]), and duration metrics (speaking duration [s], articulation duration [s], time point of maximum and minimum $F0$).

To extract facial metrics, MediaPipe face detection based on BlazeFace (Bazarevsky et al., 2019) was used to determine framewise x and y coordinates of the face. Facial landmarks were then generated by the MediaPipe face mesh algorithm (Kartynnik et al., 2019), 14 of which are key landmarks in the computation of jaw kinematics, lip aperture, mouth surface area, eyebrow height, and so forth. All facial metrics were normalized by dividing them by the intercaruncular distance (see Figure 4) to account for cross-participant positional variability relative to the camera (Roesler et al., 2022).

See Table 2 for an overview of the metrics and Supplemental Material S1 for a glossary. All metrics went through a two-step automatic outlier detection. Metrics are not excluded on a participant level but on a turn level in this method. First, all metric values beyond 5 SDs , that is, extreme outliers, from the mean metric value were removed. These extreme outliers likely arise from incorrect task performance or noncompliance. Second, the mean of the distribution was recomputed, and any values beyond 3 SDs were flagged as outliers and removed from the analysis in accordance with the three sigma rule (Upton & Cook, 2008).

Research Question 1: Effect of Cohort

Metrics Showing Statistically Significant Differences Between Cohorts

Since each subtask had multiple turns per emotion, metrics were averaged across turns within every emotion.

Figure 3. Schematic of the virtual agent-based multimodal dialogue platform used in the study. Note that the metrics shown on the right are an exemplar set. See Table 2 for a comprehensive list.

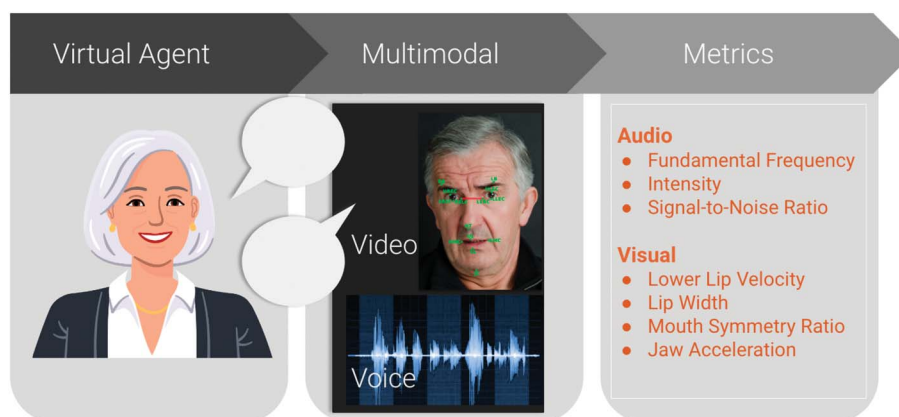
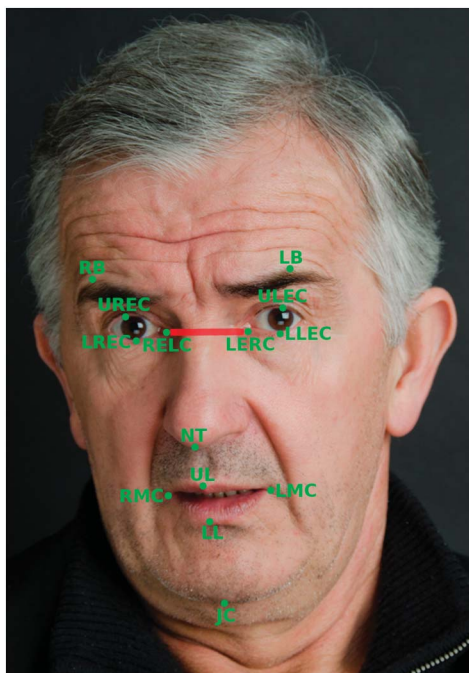


Figure 4. The 14 facial landmarks used to compute facial metrics. The intercaruncular distance is shown in red, the distance between the right eye left corner (RELC) and the left eye right corner (LERC). The other landmarks are as follows: RB/LB = right and left brow; UREC/ULEC/LREC/LLEC = upper/lower right and left eye center; NT = nose tip; UL/LL = upper/lower lip; RMC/LMC = right and left mouth corner; JC = jaw center.



All metrics were z-scored by sex at birth across the entire sample to account for sex-specific differences. To identify which metrics showed statistically significant differences, nonparametric Kruskal–Wallis tests (Kruskal & Wallis, 1952) were run for all metrics. Effect sizes, as measured by Glass’s Δ (Glass et al., 1981), were calculated for all metrics, and only those metrics with an absolute Glass’s Δ of greater than 0.6 or large effect sizes (Panzarella et al., 2021) are reported in this article.

Human Rater Classification Accuracy

To assess if group differences in facial and vocal metrics were associated with differences in affective communication as opposed to general differences in facial and vocal expression, two human raters (with at least average scores on the DANVA-2 Facial and Vocal Affect Recognition subtests; Nowicki & Duke, 1994) classified the emotion (happy, sad, angry, afraid, or neutral) produced by participants in response to each prompt. Raters were blinded to the prompted emotion and group. Facial affect was classified from video responses of each participant in the absence of vocal audio. Likewise, vocal affect was classified in the absence of video. Thus, raters made affective judgments based solely on the facial and speech behavior of the participant, respectively. Percent accuracy values of emotion judgment using video and audio were calculated for each rater and were then averaged across raters. Nonparametric independent-samples Mann–Whitney U tests were then run to identify differences between the two cohorts in percent accuracy of the rater’s perception of facial and vocal affective expression.

Research Question 2: Effect of Production Task Demands

To evaluate whether group differences varied according to task demands (i.e., production length, comprehension of contextual narrative), the above-mentioned Kruskal–Wallis tests were performed on task–metric combinations.

Research Question 3: Effect of Produced Emotion

Testing for an Interaction Effect Between Emotion and Cohort

To test whether an interaction effect between emotion and cohort was present when it came to differences in metric values, we ran a two-step analysis. For this analysis, we aggregated all 30 metrics across the two

Table 2. An overview of the automatically extracted metrics, 472 in total (Tasks 1–3 = 360 and Task 4 = 112).

Modality	Domain	Metrics
Speech	Energy	Shimmer (%), signal-to-noise ratio (dB), intensity (dB), articulation intensity (dB)
	Spectral	Mean; standard deviation; max and min fundamental frequency (F_0 , Hz); jitter (%); formant frequencies F_1 , F_2 , and F_3 (Hz) ^a ; F_2 slope (Hz/s) ^a
	Timing	Speaking duration (s), ^b percentage pause time (%), ^b articulation duration (s), time point of maximum and minimum F_0
	Voice quality	Cepstral peak prominence (dB), harmonics-to-noise ratio (dB)
Facial	Oral/labial	Lip aperture, lip width, mouth surface area, mean symmetry ratio of the mouth surface area of the left half to the right half
	Articulatory movement	Velocity, acceleration, and jerk of the lower lip and jaw center
	Ocular and circumocular	Eye opening, vertical displacement of the eyebrows

^aNot calculated for Task 4. ^bCalculated only for Task 4.

comparable tasks that involved noncontextual and contextual monosyllabic production (Tasks 1 and 3) by averaging them. In Step 1, we ran a one-way repeated-measures analysis of variance (ANOVA; Vallat, 2018) to test for an effect of emotion without controlling for cohort. In Step 2, to test for an interaction effect between cohort and emotion, we ran a mixed-design ANOVA (Murrar & Brauer, 2018) for the metrics that showed a significant effect of emotion in Step 1. The between-subjects factor was cohort and the within-subject factor was emotion to account for repeated measurements. For metrics that showed a significant interaction effect between emotion and cohort, post hoc Wilcoxon signed-rank tests were run for pairwise comparison.

Research Question 4: Group Differences in Vocal Imitations

To evaluate whether the cohorts differed in vocal affect imitation, which assesses the ability to produce vocal sounds that convey emotion without requiring knowledge of how to use vocalization for the purpose of communication emotion, we looked at metrics showing differences in Task 2 (vocal imitation of a noncontextual monosyllable).

Research Question 5: Objective Classification Accuracy

Classification Experiment

For all metrics with Glass's Δ greater than 0.6, a leave-one-out logistic regression classifier model using the Scikit-learn Python module (Pedregosa et al., 2011) was run using speech acoustic metrics alone, facial metrics alone, and the combination of both modalities. Receiver operating characteristic (ROC) curves were plotted for these three models. Area under the curve (AUC; the higher the better) and the Brier score (Brier, 1950) measuring the accuracy of probabilistic predictions (the lower the better) were generated for these models. This process was repeated for each individual emotion to test classification of groups while controlling for emotion.

Results

Research Question 1: Effect of Cohort

The effect sizes of metrics, as measured by Glass's Δ , that showed a statistically significant difference between the ASD cohort and controls and had an absolute value greater than 0.6 are shown in Figure 5. A positive effect size denotes a greater median value for the ASD cohort, and a negative effect denotes a smaller median value for

Figure 5. Effect sizes of speech and facial metrics that show statistically significant differences between ASD and controls at an alpha threshold of .05. Positive effect sizes (dark blue boxes) indicate larger values in the ASD cohort. Negative effect sizes (red boxes) indicate smaller values in the ASD cohort. Task 1: noncontextual monosyllabic emotion production, Task 2: noncontextual monosyllabic vocal imitation, Task 3: contextualized monosyllabic emotion production, and Task 4: noncontextual sentence-length emotion production. ASD = autism spectrum disorder group; F_0 = fundamental frequency; SNR = signal-to-noise ratio; CPP = cepstral peak prominence; Avg = average; Accel = acceleration; Vert = vertical.

Metric	Task	Emotion			
		Happy	Sad	Afraid	Angry
Standard Deviation F_0	Task 1		0.892		
	Task 2		0.884		
	Task 3				
	Task 4				
Max F_0 Time Point	Task 1				
	Task 2	-0.813	-0.676	-0.775	
	Task 3				
	Task 4				
Min F_0 Time Point	Task 1				
	Task 2	-0.767	-0.735		
	Task 3				
	Task 4				
SNR	Task 1			-0.716	
	Task 2			-0.621	
	Task 3				
	Task 4				
Jitter	Task 1			0.674	
	Task 2				
	Task 3				
	Task 4				
CPP	Task 1				
	Task 2			-0.751	
	Task 3				
	Task 4				
Eye Opening Avg	Task 1				-0.620
	Task 2				
	Task 3			-0.793	
	Task 4			-0.740	
Lower Lip Accel Abs Avg	Task 1		0.809		
	Task 2				
	Task 3				
	Task 4				-0.631
Lip Width Avg	Task 1	-0.674			
	Task 2				
	Task 3				
	Task 4				
Eyebrow Vert Displacement Avg	Task 1				
	Task 2				
	Task 3				
	Task 4			-0.792	

the ASD cohort. Note that nonsignificant effects are represented by blank spaces, irrespective of the size or direction of the actual effect.

With regard to vocal metrics, participants with ASD exhibited a larger standard deviation of the F_0 of their voice when conveying sadness during noncontextual monosyllabic imitation and production tasks. Moreover, the time point of the minimum and maximum values of F_0 during the imitation task occurred earlier in the ASD cohort when the emotion to be conveyed was sad or

happy. The maximum value of $F0$ also occurred earlier in the ASD cohort while repeating an afraid “oh.” The SNR during the noncontextual monosyllabic imitation and production tasks conveying fear was lower in the ASD cohort. The ASD cohort also had a greater jitter value during the noncontextual monosyllabic production task and a lower CPP value while imitating an afraid “oh.”

In the case of facial metrics, when participants with ASD produced an angry “oh” for the noncontextual monosyllabic production task, an afraid “oh” for the contextualised monosyllabic production task, and a noncontextual sentential production conveying fear, they had a smaller eye opening than the control group. Acceleration of the lower lip during a sad noncontextual monosyllabic production of “oh” was higher in the ASD group and lower during an angry sentential production. During a happy noncontextual monosyllabic production, children with ASD had a smaller lip width than controls. Also, during a happy sentential production, the ASD group had smaller average eyebrow vertical displacement.

Average human rater accuracy for facial affect recognition was 54% (47% for the ASD cohort and 65% for NTC). Average human rater accuracy for vocal affect recognition was 56% (53% for ASD and 63% for NTC). For all tasks and emotions, human rater accuracy in emotion perception for facial data, defined by the agreement between the rater’s emotion classification and the prompted emotion, was significantly different between the two cohorts ($U = 487.00, p = .002$), indicating that the ASD group was less effective in communicating the prompted emotions than the control group. Lower human rater accuracy for the ASD cohort was seen across tasks: Task 1 ($U = 529.5, p = .007$), Task 2 ($U = 535.00, p = .018$), Task 3 ($U = 436.00, p = .002$), and Task 4 ($U = 420.00, p < .001$). When rater accuracy of video data was split by emotions, lower accuracy for the ASD cohort was observed for three of the four emotions: happy ($U = 562.50, p = .016$), sad ($U = 576.00, p = .023$), afraid ($U = 393.00, p < .001$), and angry ($U = 703.50, p = .334$).

Overall human rater accuracy of emotion perception based on speech data was not significantly different between the two cohorts ($U = 463.50, p = .142$). Such differences were also not seen when data were split by tasks: Task 1 ($U = 474.00, p = .277$), Task 2 ($U = 469.00, p = .201$), Task 3 ($U = 404.00, p = .052$), Task 4 ($U = 511.50, p = .531$). While overall rater accuracy and accuracy for combined emotions for each tasks did not show significant group differences for speech data, when perceptual accuracy of speech data was compared for each emotion across tasks, the ASD group demonstrated significantly poorer vocal communication of happiness ($U = 343.50, p = .005$) and sadness ($U = 411.00, p = .048$). Significant

group differences were not identified for vocal communication of fear ($U = 523.50, p = .535$) and anger ($U = 649.00, p = .373$).

Research Question 2: Effect of Production Task Demands on Facial and Vocal Metrics

From Figure 5, it can be observed that differences in facial and vocal metrics between the two cohorts can be captured even with monosyllabic or shorter utterances. In fact, only three metrics showed differences when the participants produced sentence-length utterances (Task 4). Interestingly, all three metrics were facial metrics (eye opening, lower lip acceleration, and eyebrow displacement).

The prompt during Task 3 included an illustrated narrative providing additional emotional context. Notably, only one facial metric (eye opening during the expression of fear) showed differences between the two cohorts in Task 3. All other differences in objective metrics were captured when additional narrative context was not provided.

Research Question 3: Effect of Produced Emotion

Twenty-seven of the 30 metrics aggregated across Tasks 1 and 3 showed a significant effect of emotion (see Supplemental Material S1). The three metrics that did not show an effect of emotion were: $F2$ slope, time point of minimum $F0$, and mean symmetry ratio of the mouth surface area. Four of the 27 metrics showing a significant effect of emotion also showed a significant interaction effect between cohort and emotion (see Supplemental Material S1). These metrics were average lip width, average velocity of the lower lip, average acceleration of the lower lip, and average jerk of the lower lip. Post hoc pairwise Wilcoxon signed-ranks tests were run to test for which emotion the metrics were significantly different between cohorts. Only one pairwise test was statistically significant (see Table 3); mean velocity of lower lip was significantly higher in the NTC cohort as compared to the ASD cohort when the emotion was happy ($p = .03$, Hedges’s $g = 0.60$).

Research Question 4: Group Differences in Imitation

There were differences only in speech metrics and not facial metrics when participants were asked to repeat a monosyllable after listening to an audio stimulus, which is not surprising given that participants were only asked to imitate the vocal expression. The differences in vocal imitation were related to the standard deviation of $F0$, time points of maximum and minimum values of $F0$, SNR, and CPP.

Table 3. Metric and emotion aggregated across Tasks 1 and 3 showing a significant difference between cohorts.

Metric	Emotion	Post hoc pairwise test (Wilcoxon signed-ranks)	
		p	Hedges's g (measure of effect size) Control > ASD
Lip width average	Happy	.06	0.56
Mean velocity of lower lip	Happy	.03	0.60
Mean acceleration of lower lip	Happy	.06	0.46
Mean jerk of lower lip	Happy	.10	0.38

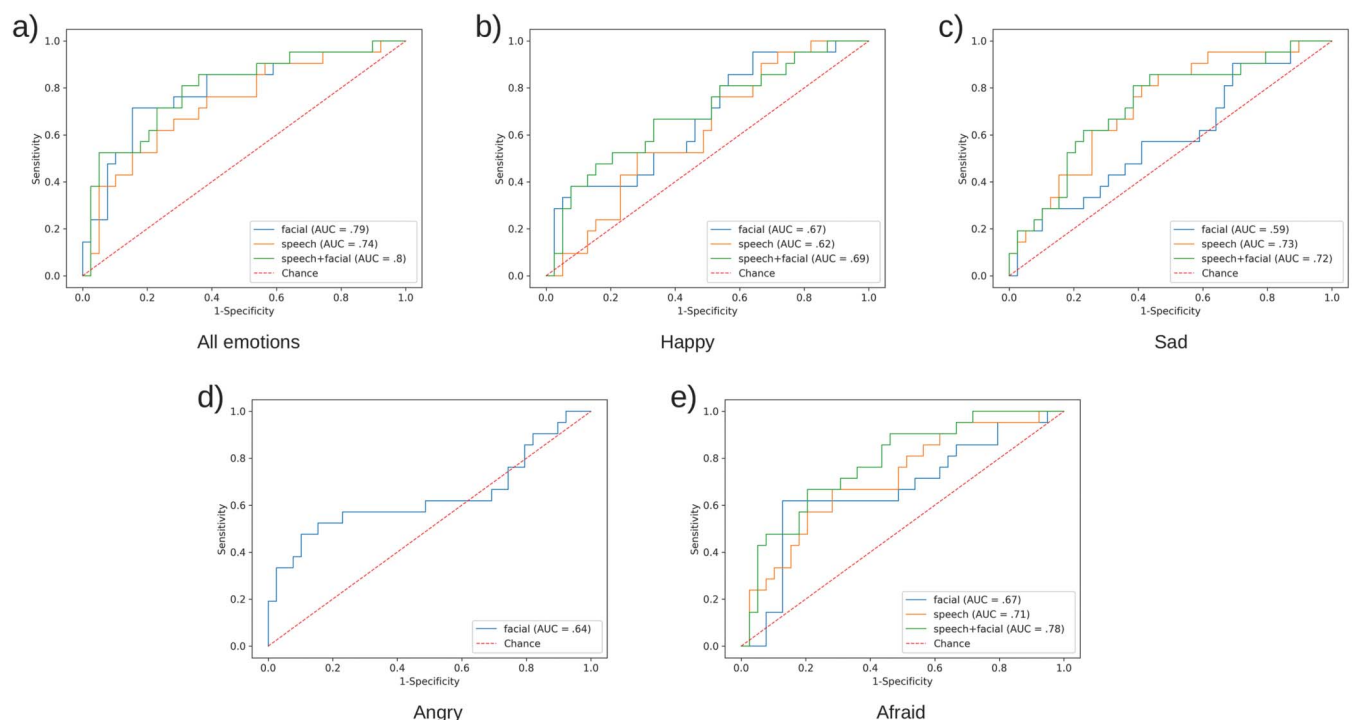
Note. ASD = autism spectrum disorder group.

Research Question 5: Objective Classification Accuracy

ROC curves for the classification experiment between cohorts can be seen in Figure 6. When all metrics, across emotions, with an absolute effect size greater than 0.6 were used as features in the classifier, both facial (AUC = .79, Brier score = 0.17) and speech metrics (AUC = .74, Brier score = 0.20) performed well above chance with the facial metrics outperforming the speech metrics. The performance, as measured by the AUC, of the classifier model was slightly better than individual modalities when metrics from both modalities were considered (AUC = .8,

Brier score = 0.18). When metrics related to happy utterances were considered, facial metrics (AUC = .67, Brier score = 0.21) again outperformed the speech metrics (AUC = .62, Brier score = 0.22) in classification of the two cohorts. A multimodal model, with both speech and facial metrics, was again slightly better than the individual modalities alone (AUC = .69, Brier score = 0.20). For sad utterances, the speech metrics performed much better (AUC = .73, Brier score = 0.20) than the facial metrics (AUC = .59, Brier score = 0.22), and the performance did not improve drastically when a combination of both modalities was used (AUC = .72, Brier score = 0.20). Since there were no differences in speech metrics for angry

Figure 6. (a) ROC curves for speech metrics alone, facial metrics alone, and a combination of speech and facial metrics when all metrics, across emotions, with absolute effect size greater than 0.6 were considered. (b) ROC curves for happy emotion metrics. (c) ROC curves for sad emotion metrics. (d) ROC curves for angry emotion metrics. (e) ROC curves for afraid emotion metrics. ROC = receiver operating characteristic; AUC = area under the curve.



utterances between the two cohorts, a classifier model with the two facial metrics showing a difference between the two cohorts was run (AUC = .64, Brier score = 0.21), and its performance, while not being superlative, was well above chance. For afraid utterances, speech metrics (AUC = 0.71, Brier score = 0.20) were slightly better at classifying the cohorts than the facial metrics (AUC = .67, Brier score = 0.22). A combination of both modalities had a much better performance for afraid utterances (AUC = .78, Brier score = 0.18).

Discussion

In this study, we investigated which audiovisual metrics associated with affect production and imitation showed significant differences between children with ASD and NTCs. We examined effects of task demands and specific emotions on group differences and used a leave-one-out logistic regression classifier model to evaluate the efficacy of facial and vocal metrics in classifying groups.

With regard to Research Question 1, we identified group differences in objective facial and vocal metrics within each emotional category and across-task conditions. We also identified relatively lower human rater accuracy in identifying affect conveyed by the ASD cohort, which is to be expected given that this group is partially defined by deficits in nonverbal communication. These differences in human accuracy, in conjunction with prior work demonstrating prediction of human rater performance from objective metrics (Demopoulos et al., 2024), suggest that the metrics are capturing differences in ability to communicate affect as opposed to nonspecific differences in facial movement and vocal expression. Specifically, in this prior study, we found that the linear combination of objective facial metrics predicted 32%–60% of the variance in human rater accuracy for facial APTs and the linear combination of objective vocal metrics predicted 41%–58% of the variance in human rater accuracy for vocal APTs. This suggests that the automatically extracted metrics are measuring information that human raters are using in making affective judgments.

With regard to Research Question 2, effects of task demands, while objective metrics extracted from all tasks were useful in distinguishing between the two cohorts, there were more metrics that significantly distinguished groups in Tasks 1 and 2 (noncontextual monosyllabic production and imitation) than those from Tasks 3 and 4 (contextualized monosyllabic and noncontextual sentence-length production). This suggests that the task with most minimal expressive and receptive language demands (i.e., brief verbal instructions and requiring production of only a monosyllabic utterance) was equally, if not more,

effective in identifying differences in affect production associated with autism. This also suggests that assessment via the APT can be accessible to individuals who require minimal language demands for valid assessment. Notably, for Tasks 3 and 4, only facial and no vocal metrics showed differences between children with ASD and controls. The nature of the tasks, prompted monosyllabic speech after a narrative/picture stimulus and noncontextual sentence-length productions, may have a role to play in this observation. Specifically, both of these tasks have greater verbal demands in different ways. For example, greater receptive language skills are necessary to understand the narrative, even though only a monosyllabic utterance is required for vocal response to the contextual monosyllabic condition. In contrast, greater speech/expressive language skills are necessary to produce the longer sentence-length utterance, while the semantic content of the sentence is not meaningful to affective vocal production in and of itself.

Regarding Research Question 3, interaction effects between group and prompted emotion, a main effect of emotion was identified across most objective metrics, as expected given that these metrics were selected based on their relevance to communicating affect. An interaction between group and emotion was also identified for several facial metrics related to mouth movements and positions. Taken together with human rater data indicating overall less effective communication of facial affect in the ASD group, these interaction effects suggest that poor affect production in the ASD group may be associated with ineffective use of mouth movements and position during facial expression. Indeed, we observed smaller lip width during monosyllabic production of speech conveying happiness in the ASD cohort. A happy vocalization is often accompanied by smiling where the mouth orifice is widened (Shor, 1978; Tartter, 1980). Smaller lip width in the ASD cohort during happy emotional speech production may indicate the absence of an accompanying smile, therefore indicating an inability to express the emotion successfully.

Several other group differences were identified under specific emotion conditions and specific tasks. For example, when conveying a sad emotion, participants with ASD had a greater standard deviation of the $F0$ of their voice as compared to controls during both noncontextual monosyllabic production and imitation tasks conveying a sad emotion. Relatedly, according to the human rater accuracy data, the ASD cohort demonstrated poorer vocal communication of sadness. Indeed, increased pitch variation during speech production in general and emotional speech production in particular has been observed quite consistently in studies of individuals with autism described as “high functioning” (Diehl et al., 2009; Edelson et al., 2007; Fosnot & Jun, 1999; Nadig & Shaw, 2012). This

increased pitch variability has been shown to be language agnostic and is not associated with the language ability of the ASD participants (Bonneh et al., 2011; Green & Tobin, 2009; Sharda et al., 2010).

We also observed that during emotional speech production, maximum $F0$ and minimum $F0$ time points occurred earlier in the ASD cohort. Atypical prosody has been documented in both receptive and expressive speech in high-functioning autism (McCann & Peppé, 2003; McCann et al., 2007; Peppé et al., 2007). Individuals with ASD are said to experience difficulties with social acceptance due to the atypical prosody of their speech (Paul et al., 2005; Shriberg et al., 2001), underscoring that prosodic differences may functionally impact paralinguistic aspects of vocal communication. Specifically, it has been postulated that extreme pitch variation in ASD could be placed arbitrarily in the utterance, thus rendering the acoustic cues of the speech nonmeaningful to listeners (Nadig & Shaw, 2012). Understanding the root of these prosodic differences could direct novel approaches to improving communication skills via targeting the barriers to effective use of ancillary acoustic cues of vocalization (not only what is said but *how* it is said).

The SNR was lower in the ASD cohort during the noncontextual monosyllabic imitation and production of “afraid.” Furthermore, metrics indicative of voice quality (i.e., jitter and CPP) were higher and lower, respectively, in the ASD cohort for the afraid utterances. Group differences in human rater accuracy of fear were not identified, however, suggesting these may be vocal differences not associated with affective communication.

Additionally, we also observed lower average vertical displacement of the eyebrows during happy noncontextual sentence-length production and a smaller average eye opening during noncontextual monosyllabic productions of fear and anger in the ASD cohort. These results, combined with the lower human rater accuracy for the ASD cohort, suggest that individuals with ASD exhibit reduced expressivity of nonverbal cues of emotion during affect production. The eyes and eyebrows have a significant role to play in the expression of distinct human emotions (Perveen et al., 2012). Furthermore, reduced eye contact is also common in individuals with ASD. The findings of reduced facial expressivity in the eye region of the face may be associated with lack of experience in reading those cues in others or, alternatively, a lack of salience for those movements resulting in a tendency not to produce them or watch for them in others.

With regard to Research Question 4, group differences in affect imitation (Task 2), only speech acoustic metrics showed large differences. No emotional cue was provided in this task apart from the paralinguistic

information in the audio stimulus (a noncontextual monosyllabic vocalization). Furthermore, there was no facial stimulus to imitate and no instructions regarding production of facial expression. Differences in speech metrics but not in facial metrics for this task may suggest that both cohorts had a similar range of facial motoric productions during these vocal imitations, but the ASD cohort differed in the reproduction of the paralinguistic information in the actor’s speech sounds. Interestingly, the average rater accuracy was still relatively lower for facial imitation in the ASD group despite failure to identify differences in objective facial metrics. This may suggest that both groups performed a similar range of facial movements under non-directed conditions, but the typically developing controls group produced more facial expressions that were congruent with the vocal affect of the stimulus prompt and the ASD group produced facial expressions that were less communicative of the emotion being conveyed in the vocal affect of the stimulus prompt. These results are consistent with prior studies reporting reduced multimodal affective communication (Hubbard et al., 2017; Loveland et al., 1994; Sorensen et al., 2019).

Finally, for Research Question 5, we evaluated classifier models to distinguish between ASD and controls. In these analyses, we observed that classifiers performed best when metrics across all emotions were considered. However, there were slight differences in model performance when individual emotions were considered. For utterances that were supposed to convey sadness and fear, speech acoustic metrics performed better than facial metrics. When it came to happy utterances, there was an equal number of speech and facial metrics showing differences between the two cohorts, but the facial metrics had better classification performance. Interestingly, for anger, only facial metrics showed large differences between the two cohorts, and their classification performance was well above chance. For most of the classification models though, the combination of both modalities (speech and facial) was either more effective than individual modalities or equivalent to the performance of the better performing modality (speech metrics for sadness). This observation underscores the importance of a multimodal framework approach in studying a complex disorder like ASD with significant cross-domain atypicalities (C.-P. Chen et al., 2017; J. Chen et al., 2020; Kothare et al., 2021; Samad et al., 2017). The observed AUCs are comparable to prior classification studies with multimodal features (Cho et al., 2019).

The current study comes with a set of limitations. First, while this study focused on measurement of objective features of facial and vocal expression of emotion under cued conditions in order to standardize the emotion intended to be communicated across participants, it is not the same as measuring emotional expression under

spontaneous environmentally provoked conditions. The objective metrics extracted essentially measure the participants' ability to consciously and intentionally produce or act out the emotions. This distinction is important as natural expression of emotion may not communicate the experienced emotion directly but may be influenced by what one intends to communicate (i.e., one may not wear one's heart on one's sleeve for certain reasons and in certain contexts). There may be differences in communicative intentions that would distinguish ASD from other groups in a more naturalistic assessment of expressive emotional communication. Future studies should investigate the differences between objective measures of emotional expression upon cue and emotional expression in the natural environment. Second, the control cohort was smaller than the ASD cohort, and future investigations should consider larger and more equally matched cohorts. Third, it cannot be ascertained that group differences between the cohorts arise solely due to functional deficits. Differences observed could just be a reflection of how the two groups responded to task prompts. It must be noted that the results of this study may only be generalizable under certain contexts and prompts. Lastly, although the doctoral student did not interact with the participants during data collection, their presence in the room may have affected performance differentially between cohorts. Future studies involving remote data collection in natural environments may help answer some of these questions. That being said, it is worth underscoring that the platform described in this study can also be accessed remotely using any device equipped with a webcam and a microphone (Ramanarayanan et al., 2020). This is especially important considering that there is a growing need to remove barriers and expand telehealth services in children with neurodevelopmental disorders (Masi et al., 2021).

In conclusion, we found that audiovisual metrics extracted through a multimodal conversation-based dialogue platform show significant differences between children with ASD and NTCs and may have potential for monitoring behavior in ASD. Both monosyllabic and sentence-length prompts may have their own advantages in evoking emotional expressions. Monosyllabic utterances would make the task more accessible to people who are minimally verbal. Sentence-length utterances may capture nuances of speech acoustics and facial movement due to the greater length of data available per utterance; however, our current findings suggest that a monosyllabic utterance may be sufficient to effectively measure affect production. We also observed that emotional context in the form of a narrative, which requires more skill in receptive language than other tasks, was not necessary to evoke group differences in emotional expressions. Future research examining psychometric performance of these

different task conditions is needed to determine minimum required task demands for sensitive measurement while maximizing inclusivity and accessibility of the task. Emotion-specific and task-specific differences in metrics and model performance were also observed. Furthermore, we found that a multimodal approach is important to classify children with ASD from controls. This is even more important because of the emotion-specific differences in classification performance of the individual modalities.

Data Availability Statement

The de-identified data (speech and facial metrics) generated and analyzed during the current study is available from the corresponding author on reasonable request. Note that the raw audio and video are not shareable as they contain personally identifiable information regarding the participants.

Acknowledgments

This study was supported by National Institutes of Health Grants K23 DC016637 and R01DC019167, Autism Speaks Grant 11637, and UCSF Weill Institute for Neuroscience Weill Award for Clinical Neuroscience Research awarded to Carly Demopoulos.

References

- American Psychiatric Association.** (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). <https://doi.org/10.1176/appi.books.9780890425596>
- Bangerter, A., Chatterjee, M., Manfredonia, J., Manyakov, N. V., Ness, S., Boice, M. A., Skalkin, A., Goodwin, M. S., Dawson, G., Hendren, R., Leventhal, B., Shic, F., & Pandina, G.** (2020). Automated recognition of spontaneous facial expression in individuals with autism spectrum disorder: Parsing response variability. *Molecular Autism*, *11*(1), Article 31. <https://doi.org/10.1186/s13229-020-00327-4>
- Bazarevsky, V., Kartynnik, Y., Vakunov, A., Raveendran, K., & Grundmann, M.** (2019). *BlazeFace: Sub-millisecond neural face detection on mobile GPUs*. arXiv. <https://doi.org/10.48550/ARXIV.1907.05047>
- Boersma, P., & van Heuven, V.** (2001). Speak and unSpeak with PRAAT. *Glott International*, *5*(9/10), 341–347.
- Bonneh, Y. S., Levanon, Y., Dean-Pardo, O., Lossos, L., & Adini, Y.** (2011). Abnormal speech spectrum and increased pitch variability in young autistic children. *Frontiers in Human Neuroscience*, *4*, Article 237. <https://doi.org/10.3389/fnhum.2010.00237>
- Brier, G. W.** (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, *78*(1), 1–3. [https://doi.org/10.1175/1520-0493\(1950\)078<0001:VOFEIT>2.0.CO;2](https://doi.org/10.1175/1520-0493(1950)078<0001:VOFEIT>2.0.CO;2)
- Brown, L., Sherbenou, R. J., & Johnsen, S. K.** (2010). *Test of Nonverbal Intelligence—Fourth Edition: TONI-4*. Pro-Ed.
- Chen, C.-P., Tseng, X.-H., Gau, S. S.-F., & Lee, C.-C.** (2017). Computing multimodal dyadic behaviors during spontaneous

- diagnosis interviews toward automatic categorization of autism spectrum disorder. In *Proceedings of Interspeech 2017* (pp. 2361–2365). <https://doi.org/10.21437/Interspeech.2017-563>
- Chen, J., Liao, M., Wang, G., & Chen, C.** (2020). An intelligent multimodal framework for identifying children with autism spectrum disorder. *International Journal of Applied Mathematics and Computer Science*, 30(3), 435–448. <https://doi.org/10.34768/amcs-2020-0032>
- Cho, S., Liberman, M., Ryant, N., Cola, M., Schultz, R. T., & Parish-Morris, J.** (2019). Automatic detection of autism spectrum disorder in children using acoustic and text features from brief natural conversations. In *Proceedings of Interspeech 2019* (pp. 2513–2517). <https://doi.org/10.21437/Interspeech.2019-1452>
- Demopoulos, C., Lampinen, L., Preciado, C., Kothare, H., & Ramanarayanan, V.** (2024). Preliminary investigation of psychometric properties of a novel multimodal dialog based affect production task in children and adolescents with autism. In *Proceedings of Interspeech 2024* (pp. 5123–5127). <https://doi.org/10.21437/Interspeech.2024-1359>
- Diehl, J. J., Watson, D., Bennetto, L., McDonough, J., & Gunlogson, C.** (2009). An acoustic analysis of prosody in high-functioning autism. *Applied Psycholinguistics*, 30(3), 385–404. <https://doi.org/10.1017/S0142716409090201>
- Edelson, L., Grossman, R., & Tager-Flusberg, H.** (2007, May). *Emotional prosody in children and adolescents with autism* [Poster presentation]. Annual International Meeting for Autism Research, Seattle, WA.
- Ekman, P., & Friesen, W. V.** (1978). *Facial action coding system: Manual*. Consulting Psychologists Press. <https://doi.org/10.1037/t27734-000>
- Faso, D. J., Sasson, N. J., & Pinkham, A. E.** (2015). Evaluating posed and evoked facial expressions of emotion from adults with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, 45(1), 75–89. <https://doi.org/10.1007/s10803-014-2194-7>
- Fosnot, S. M., & Jun, S.** (1999). Prosodic characteristics in children with stuttering or autism during reading and imitation. In *Proceedings of the 14th International Congress of Phonetic Sciences* (pp. 1925–1928).
- Glass, G. V., McGaw, B., & Smith, M. L.** (1981). *Meta-analysis in social research* (Vol. 56). Sage.
- Green, H., & Tobin, Y.** (2009). Prosodic analysis is difficult ... but worth it: A study in high functioning autism. *International Journal of Speech-Language Pathology*, 11(4), 308–315. <https://doi.org/10.1080/17549500903003060>
- Hubbard, D. J., Faso, D. J., Assmann, P. F., & Sasson, N. J.** (2017). Production and perception of emotional prosody by adults with autism spectrum disorder. *Autism Research*, 10(12), 1991–2001. <https://doi.org/10.1002/aur.1847>
- Kartynnik, Y., Ablavatski, A., Grishchenko, I., & Grundmann, M.** (2019). *Real-time facial surface geometry from monocular video on mobile GPUs*. arXiv. <https://doi.org/10.48550/ARXIV.1907.06724>
- Kothare, H., Ramanarayanan, V., Roesler, O., Neumann, M., Liscombe, J., Burke, W., Cornish, A., Habberstad, D., Sakallah, A., Markuson, S., Kansara, S., Faerman, A., Bensidi-Slimane, Y., Fry, L., Portera, S., Suendermann-Oeft, D., Pautler, D., & Demopoulos, C.** (2021). Investigating the interplay between affective, phonatory and motoric subsystems in autism spectrum disorder using a multimodal dialogue agent. In *Proceedings of Interspeech 2021* (pp. 1967–1971). <https://doi.org/10.21437/Interspeech.2021-1796>
- Kothare, H., Roesler, O., Burke, W., Neumann, M., Liscombe, J., Exner, A., Snyder, S., Cornish, A., Habberstad, D., Pautler, D., Suendermann-Oeft, D., Huber, J., & Ramanarayanan, V.** (2022). Speech, facial and fine motor features for conversation-based remote assessment and monitoring of Parkinson's disease. In *Proceedings of the 2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 3464–3467). <https://doi.org/10.1109/EMBC48229.2022.9871375>
- Kruskal, W. H., & Wallis, W. A.** (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 47(260), 583–621. <https://doi.org/10.1080/01621459.1952.10483441>
- Leo, M., Carcagni, P., Distanto, C., Spagnolo, P., Mazzeo, P., Rosato, A., Petrocchi, S., Pellegrino, C., Levante, A., De Lumè, F., & Lecciso, F.** (2018). Computational assessment of facial expression production in ASD children. *Sensors*, 18(11), Article 3993. <https://doi.org/10.3390/s18113993>
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., Pickles, A., & Rutter, M.** (2000). The Autism Diagnostic Observation Schedule–Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of Autism and Developmental Disorders*, 30(3), 205–223. <https://doi.org/10.1023/A:1005592401947>
- Lord, C., Rutter, M., & Le Couteur, A.** (1994). Autism Diagnostic Interview–Revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, 24(5), 659–685. <https://doi.org/10.1007/BF02172145>
- Loveland, K. A., Tunali-Kotoski, B., Pearson, D. A., Brelsford, K. A., Ortegon, J., & Chen, R.** (1994). Imitation and expression of facial affect in autism. *Development and Psychopathology*, 6(3), 433–444. <https://doi.org/10.1017/S0954579400006039>
- Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Ladd-Acosta, C. M., McArthur, D., Pas, E. T., Salinas, A., Vehorn, A., Williams, S., Esler, A., Grzybowski, A., Hall-Lande, J., ... Shaw, K. A.** (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years—Autism and developmental disabilities monitoring network, 11 sites, United States, 2020. *Morbidity and Mortality Weekly Report. Surveillance Summaries*, 72(2), 1–14. <https://doi.org/10.15585/mmwr.mmwr.ss7202a1>
- Masi, A., Mendoza Diaz, A., Tully, L., Azim, S. I., Woolfenden, S., Efron, D., & Eapen, V.** (2021). Impact of the COVID-19 pandemic on the well-being of children with neurodevelopmental disabilities and their parents. *Journal of Paediatrics and Child Health*, 57(5), 631–636. <https://doi.org/10.1111/jpc.15285>
- McCann, J., & Peppé, S.** (2003). Prosody in autism spectrum disorders: A critical review. *International Journal of Language & Communication Disorders*, 38(4), 325–350. <https://doi.org/10.1080/1368282031000154204>
- McCann, J., Peppé, S., Gibbon, F. E., O'Hare, A., & Rutherford, M.** (2007). Prosody and its relationship to language in school-aged children with high-functioning autism. *International Journal of Language & Communication Disorders*, 42(6), 682–702. <https://doi.org/10.1080/13682820601170102>
- Murrar, S., & Brauer, M.** (2018). *The SAGE encyclopedia of educational research, measurement, and evaluation*. SAGE. <https://doi.org/10.4135/9781506326139>
- Nadig, A., & Shaw, H.** (2012). Acoustic and perceptual measurement of expressive prosody in high-functioning autism: Increased pitch range and what it means to listeners. *Journal of Autism and Developmental Disorders*, 42(4), 499–511. <https://doi.org/10.1007/s10803-011-1264-3>
- Neumann, M., Kothare, H., & Ramanarayanan, V.** (2024). Multimodal speech biomarkers for remote monitoring of ALS

- disease progression. *Computers in Biology and Medicine*, 180, Article 108949. <https://doi.org/10.1016/j.compbiomed.2024.108949>
- Neumann, M., Roesler, O., Liscombe, J., Kothare, H., Suendermann-Oeft, D., Pautler, D., Navar, I., Anvar, A., Kumm, J., Norel, R., Fraenkel, E., Sherman, A. V., Berry, J. D., Pattee, G. L., Wang, J., Green, J. R., & Ramanarayanan, V. (2021). Investigating the utility of multimodal conversational technology and audiovisual analytic measures for the assessment and monitoring of amyotrophic lateral sclerosis at scale. In *Proceedings of Interspeech 2021* (pp. 4783–4787). <https://doi.org/10.21437/Interspeech.2021-1801>
- Neumann, M., Roessler, O., Suendermann-Oeft, D., & Ramanarayanan, V. (2020). On the utility of audiovisual dialog technologies and signal analytics for real-time remote monitoring of depression biomarkers. In *Proceedings of the First Workshop on Natural Language Processing for Medical Conversations* (pp. 47–52). <https://doi.org/10.18653/v1/2020.nlpmc-1.7>
- Nowicki, S., & Duke, M. P. (1994). Individual differences in the nonverbal communication of affect: The Diagnostic Analysis of Nonverbal Accuracy Scale. *Journal of Nonverbal Behavior*, 18(1), 9–35. <https://doi.org/10.1007/BF02169077>
- Nussbaum, C., Schirmer, A., & Schweinberger, S. R. (2022). Contributions of fundamental frequency and timbre to vocal emotion perception and their electrophysiological correlates. *Social Cognitive and Affective Neuroscience*, 17(12), 1145–1154. <https://doi.org/10.1093/scan/nsac033>
- Panzarella, E., Beribisky, N., & Cribbie, R. A. (2021). Denouncing the use of field-specific effect size distributions to inform magnitude. *PeerJ*, 9, Article e11383. <https://doi.org/10.7717/peerj.11383>
- Paul, R., Augustyn, A., Klin, A., & Volkmar, F. R. (2005). Perception and production of prosody by speakers with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, 35(2), 205–220. <https://doi.org/10.1007/s10803-004-1999-1>
- Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., & Cournapeau, D. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning in Python*, 12(2011), 2825–2830.
- Peppé, S., McCann, J., Gibbon, F., O'Hare, A., & Rutherford, M. (2007). Receptive and expressive prosodic ability in children with high-functioning autism. *Journal of Speech, Language, and Hearing Research*, 50(4), 1015–1028. [https://doi.org/10.1044/1092-4388\(2007\)071](https://doi.org/10.1044/1092-4388(2007)071)
- Perveen, N., Gupta, S., & Verma, K. (2012). Facial expression recognition using facial characteristic points and Gini index. In *Proceedings of the 2012 Students Conference on Engineering and Systems* (pp. 1–6). <https://doi.org/10.1109/SCES.2012.6199086>
- Pokorny, F. B., Schuller, B., Marschik, P. B., Brueckner, R., Nyström, P., Cummins, N., Bölte, S., Einspieler, C., & Falck-Ytter, T. (2017). Earlier identification of children with autism spectrum disorder: An automatic vocalisation-based approach. In *Proceedings of Interspeech 2017* (pp. 309–313). <https://doi.org/10.21437/Interspeech.2017-1007>
- Ramanarayanan, V. (2024). Multimodal technologies for remote assessment of neurological and mental health. *Journal of Speech, Language, and Hearing Research*, 67(11), 4233–4245. https://doi.org/10.1044/2024_JSLHR-24-00142
- Ramanarayanan, V., Lammert, A. C., Rowe, H. P., Quatieri, T. F., & Green, J. R. (2022). Speech as a biomarker: Opportunities, interpretability, and challenges. *Perspectives of the ASHA Special Interest Groups*, 7(1), 276–283. https://doi.org/10.1044/2021_PERSP-21-00174
- Ramanarayanan, V., Pautler, D., Arbatti, L., Hosamath, A., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Cornish, A., Habberstad, D., Richter, V., Fox, D., Suendermann-Oeft, D., & Shoulson, I. (2023). When words speak just as loudly as actions: Virtual agent based remote health assessment integrating what patients say with what they do. In *Proceedings of Interspeech 2023* (pp. 678–679).
- Ramanarayanan, V., Roesler, O., Neumann, M., Pautler, D., Habberstad, D., Cornish, A., Kothare, H., Murali, V., Liscombe, J., Schnelle-Walka, D., Lange, P., & Suendermann-Oeft, D. (2020). Toward remote patient monitoring of speech, video, cognitive and respiratory biomarkers using multimodal dialog technology. In *Proceedings of Interspeech 2020* (pp. 492–493).
- Reynolds, C., & Kamphaus, R. (2015). *Behavior Assessment System for Children—Third Edition (BASC-3)*. Pearson.
- Richter, V., Neumann, M., Kothare, H., Roesler, O., Liscombe, J., Suendermann-Oeft, D., Prokop, S., Khan, A., Yavorsky, C., Lindenmayer, J.-P., & Ramanarayanan, V. (2022). Towards multimodal dialog-based speech & facial biomarkers of schizophrenia. In *Proceedings of the International Conference on Multimodal Interaction* (pp. 171–176). <https://doi.org/10.1145/3536220.3558075>
- Roesler, O., Kothare, H., Burke, W., Neumann, M., Liscombe, J., Cornish, A., Habberstad, D., Pautler, D., Suendermann-Oeft, D., & Ramanarayanan, V. (2022). Exploring facial metric normalization for within- and between-subject comparisons in a multimodal health monitoring agent. In *Proceedings of the International Conference on Multimodal Interaction* (pp. 160–165). <https://doi.org/10.1145/3536220.3558071>
- Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire: Manual*. Western Psychological Services.
- Samad, M. D., Diawara, N., Bobzien, J. L., Harrington, J. W., Witherow, M. A., & Iftekharuddin, K. M. (2017). A feasibility study of autism behavioral markers in spontaneous facial, visual, and hand movement response data. *IEEE Transactions on Neural Systems and Rehabilitation Engineering*, 26(2), 353–361. <https://doi.org/10.1109/TNSRE.2017.2768482>
- Sharda, M., Subhadra, T. P., Sahay, S., Nagaraja, C., Singh, L., Mishra, R., Sen, A., Singhal, N., Erickson, D., & Singh, N. C. (2010). Sounds of melody—Pitch patterns of speech in autism. *Neuroscience Letters*, 478(1), 42–45. <https://doi.org/10.1016/j.neulet.2010.04.066>
- Shor, R. E. (1978). The production and judgment of smile magnitude. *The Journal of General Psychology*, 98(1), 79–96. <https://doi.org/10.1080/00221309.1978.9920859>
- Shriberg, L. D., Paul, R., McSweeney, J. L., Klin, A., Cohen, D. J., & Volkmar, F. R. (2001). Speech and prosody characteristics of adolescents and adults with high-functioning autism and Asperger syndrome. *Journal of Speech, Language, and Hearing Research*, 44(5), 1097–1115. [https://doi.org/10.1044/1092-4388\(2001\)087](https://doi.org/10.1044/1092-4388(2001)087)
- Sorensen, T., Zane, E., Feng, T., Narayanan, S., & Grossman, R. (2019). Cross-modal coordination of face-directed gaze and emotional speech production in school-aged children and adolescents with ASD. *Scientific Reports*, 9(1), Article 18301. <https://doi.org/10.1038/s41598-019-54587-z>
- Suendermann-Oeft, D., Robinson, A., Cornish, A., Habberstad, D., Pautler, D., Schnelle-Walka, D., Haller, F., Liscombe, J., Neumann, M., Merrill, M., Roesler, O., & Geffarth, R. (2019). NEMSI: A multimodal dialog system for screening of neurological or mental conditions. In *Proceedings of the 19th ACM International Conference on Intelligent Virtual Agents* (pp. 245–247). <https://doi.org/10.1145/3308532.3329415>
- Talkar, T., Williamson, J. R., Hannon, D. J., Rao, H. M., Yuditskaya, S., Claypool, K. T., Sturim, D., Nowinski, L., Saro, H., Stamm, C., Mody, M., McDougle, C. J., & Quatieri,

-
- T. F. (2020). Assessment of speech and fine motor coordination in children with autism spectrum disorder. *IEEE Access*, 8, 127535–127545. <https://doi.org/10.1109/ACCESS.2020.3007348>
- Tartter, V. C. (1980). Happy talk: Perceptual and acoustic effects of smiling on speech. *Perception & Psychophysics*, 27, 24–27. <https://doi.org/10.3758/BF03199901>
- Upton, G., & Cook, I. (2008). *A dictionary of statistics*. Oxford University Press. <https://doi.org/10.1093/acref/9780199541454.001.0001>
- Vallat, R. (2018). Pingouin: statistics in Python. *Journal of Open Source Software*, 3(31), Article 1026. <https://doi.org/10.21105/joss.01026>
- Wechsler, D. (2014). *WISC-V: Technical and interpretive manual*. Pearson.
- Wiig, E. H., Secord, W. A., & Semel, E. (2013). *Clinical Evaluation of Language Fundamentals—Fifth Edition: CELF-5*. Pearson.
- Zane, E., Yang, Z., Pozzan, L., Guha, T., Narayanan, S., & Grossman, R. B. (2019). Motion-capture patterns of voluntarily mimicked dynamic facial expressions in children and adolescents with and without ASD. *Journal of Autism and Developmental Disorders*, 49(3), 1062–1079. <https://doi.org/10.1007/s10803-018-3811-7>