Test-Retest Reliability of Automated Social Communication Algorithms to Support Autism Characterization

Camille Brune, PhD[1]; Khaleel Kamal[2]; Edith V. Ocampo, MS[1]; Latha V. Soorya, PhD[1]; Sarely Licona[3], BS; Zoltán Tősér[4]; Luke Hermle, BS[1] & Anna Hirshman, BA[1]

[1]Department of Psychiatry, Rush University Medical Center, Chicago, IL; [2]Argus Cognitive Inc., Budapest, Hungary; [3]Department of Counseling, Clinical, & School Psychology, University of California, Santa Barbara; [4]Argus Cognitive Inc., Lebanon, NH

BACKGROUND

- Computational behavioral assessment has potential to facilitate rapid assessment and improve accessibility, affordability, and objectivity in behavioral health conditions.
- Argus-MDS is a non-intrusive, automated platform utilizing computer vision and speech analysis to map key social-emotional behaviors in autism, PTSD, and other behavioral health conditions. The system is used to support diagnostic and biomarker evaluations in research applications including clinical trials and natural history studies.

AIMS AND OBJECTIVE

The current study develops psychometric properties of biometric social communication metrics in autistic children through the following aims:

1. Evaluating goal metrics (e.g., precision) of quality control for machine learning
2. Confirming test-retest reliability of biometric output
3. Developing communication systems between engineers and clinicians to examine ground truth (see Figure 4)

PARTICIPANTS AND METHOD

- Participants (Table 1, Figure 1, Figure 2)
- Test-retest window: M=11.93, SD= 10.78 (range: 6-45 days)
- **Visit 1**: NIH medical history, ACE family medical history, Social Responsiveness Scale-2, Pervasive Developmental Disorders Behavior Inventory-Screening Version, Vineland-3, IQ and standardized Autism Diagnostic Observation Schedule (ADOS-2)[1,2] (Figure 3)
- **Visit 2**:standardized ADOS-2 (re-test). See ICCs for ADOS-2 scores (Table 1).

Table 1
Participant Characteristics

	Visit 1		Visit 2		ICC
Measure	Mean (SD)	Range	Mean (SD)	Range	
Age, years	4.27 (3.20)	1-11			
IQ	67.67 (21.54)	49-105			
ADOS SA	13.20 (5.25)	6-21	13.93 (4.71)	7-21	0.86
ADOS RRB	5.87 (2.03)	1-8	5.80 (1.97)	3-8	0.79
ADOS Total	19.07 (5.98)	9-26	19.73 (5.76)	10-28	0.87
ADOS CS	8.00 (1.55)	6-10	7.82 (1.66)	6-10	0.86

Note. SA, Social Affect; RRB, Restricted and Repetitive Behavior; CS, Comparison Score (not calculated for Toddler Module)

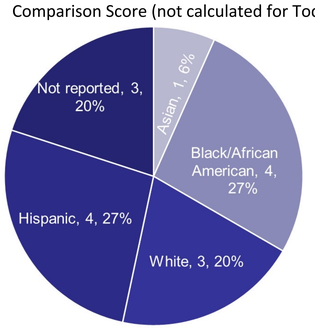

Figure 1. Participant self-identified ethnicity

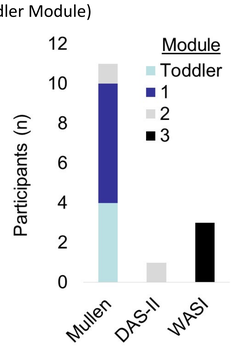

Figure 2. IQ and ADOS-2 Module Administration

DETECTED BEHAVIORAL METRICS

Facial Expressions

- Detected using facial landmarks (Figure 3.1) from Ekman's Facial Action Units[4]

Gaze Detection

- Participant's gaze estimated from Tobii externally mounted camera and room camera with an angular root mean square error of 3.5˚incorporating proximity of the participant's eyes to the camera[3]
- Clinician's gaze from Tobii glasses' sensors

Vocalizations/Speech

- Vocabulary size excludes stop word (e.g., "to") defined by the Natural Language Toolkit
- Emotion words identified in the Opinion Lexicon dictionary (https://www.cs.uic.edu/)
- Word repetitions (i.e., word spoken more than once within a segment)

Social Initiation

- Number of exchanges initiated by the participant or clinician making eye contact immediately followed by a conversational episode or vice versa

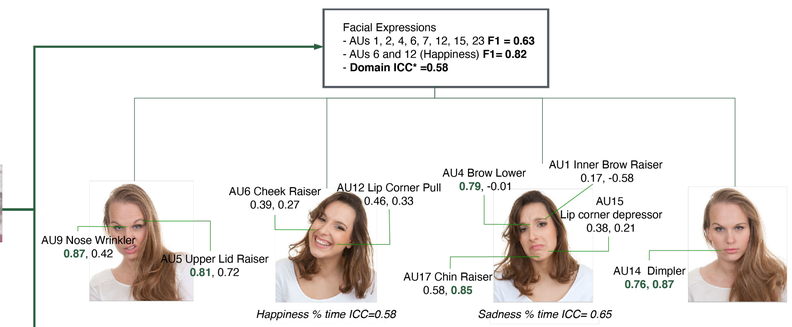

Figure 3.1. Ekman FACS[5]

Figure 3. ARGUS-MDS-ASD Protocol
**Domain ICC represents the average ICC over the domain metrics*
AU ICC scores are respectively for metrics: % of time, average duration

QUALITY CONTROL METRICS

- ***IOU,*** Intersection Over Union: see Figure 4
- ***Precision***: rate of detecting only positive cases, as opposed to false positives.
- ***Recall***: rate of detecting all the positive cases, as opposed to false negative.
- ***F1***: combines the Precision and Recall metrics into a single metric, representing both.
- ***ICC***, Intraclass Correlation Coefficient: a reliability index in test-retest analyses.

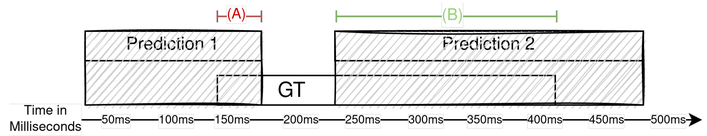

Figure 4. IOU: the temporal ratio of intersection *(A, B)* to union (*sketched area*) between prediction and GT (Ground Truth). In the case of IOU=0.75, Prediction 2 meets the threshold and is hence a positive detection.

Table 2
Test-retest ICCs for measures in the Speech Domain

Measure	Item	ICC
Unusual vocalizations	Mean pitch (Mel scale)	0.53
	Speech pauses, silence for 300-500 ms (#)	0.69
Abnormal prosody	Speech pause (variance of length)	0.57
Conversational exchanges	Vocal interchanges, minimum 3.5 seconds	0.63
	Interchanges (≥2 turns, % time, average duration)	0.57, 0.50
Communicative abilities (# unique words)	Mean length of utterances (MLU)	0.73
	Morphemes (#)	0.64
	Vocabulary size	0.69
Emotion specific vocabulary	**Positive sentiment words (#)**	**0.80**
	Negative sentiment words (#)	0.70
	Emotion specific words (#)	**0.79**
Word repetitions	Repetitions per utterance (#)	0.38

DISCUSSION AND FUTURE DIRECTIONS

- ARGUS-MDS-ASD analyzes real time social communication behaviors during semi-structured clinical evaluations
- Excellent quality and reliability were found for metrics of gaze and social initiations
- We analyzed facial action units over the whole recording, including segments where the participant spoke or displayed other vocal activity. Thus, we implicitly capture characterizations of vocal activities and facial expressions in a composite score. Given our sample diversity (>60% non-white participants), we are examining culprits of potential bias in algorithms detecting facial expressions in non-white faces. One contributing factor is lighting contrast. Because the program learns from input, revised algorithms and additional observations should continue to improve these Quality Control metrics.
- Next steps include assessing convergent validity of *biometric* data with scores on standardized screening and diagnostic instruments; and expanding recruitment to other neurodevelopmental disorders, typically developing individuals and non-English speakers to better differentiate ASD, provide norms for biometric data, and increase inclusivity.

REFERENCES
[1]C. Lord, M. Rutter, P. C. DiLavore, S. Risi, K. Gotham & S. Bishop Autism Diagnostic Observation Schedule, Second Edition. Torrance, CA: Western Psychological Services, 2012.
[2]C. Lord, R. J. Luyster, K. Gotham & W. Guthrie (2012). Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part II): Toddler Module. Torrance, CA: Western Psychological Services, 2012.Autism Diagnostic Observation Schedule, Second Edition.
[3]Tősér, Z., Rill, R. A., Faragó, K., Jeni, L. A., & Lőrincz, A. (2016, September). Personalization of gaze direction estimation with deep learning. In Joint German/Austrian Conference on Artificial Intelligence (Künstliche Intelligenz) (pp. 200-207). Springer, Cham.
[4]Ekman, P., Friesen, W. V., and Hager, J. C. (2002). *The Facial Action Coding System: A Technique for the Measurement of Facial Movement*. San Francisco, CA: Consulting Psychologists Press.
[5]Ekman's Facial Action Coding Units courtesy of imotions.com