ARTICLE **Communicated by Terrence Sejnowski**

Hidden Aspects of the Research ADOS Are Bound to Affect Autism Science

Elizabeth B. Torres
ebtorres@psych.rutgers.edu
Psychology Department; Computer Science, Center for Biomedical Imagining and Modeling; and Rutgers University Center for Cognitive Science, Rutgers University, Piscataway, NJ 08854, U.S.A.

Richa Rai
richarai9@gmail.com
Psychology Department, Rutgers University, Piscataway, NJ 08854, U.S.A.

Sejal Mistry
sejal.mistry@hsc.utah.edu
Mathematics Department, Rutgers University, Piscataway, NJ 08854, U.S.A.

Brenda Gupta
brendap.patel@gmail.com
Montclair State University, Montclair, NJ 07043, U.S.A.

The research-grade Autism Diagnostic Observational Schedule (ADOS) is a broadly used instrument that informs and steers much of the science of autism. Despite its broad use, little is known about the empirical variability inherently present in the scores of the ADOS scale or their appropriateness to define change and its rate, to repeatedly use this test to characterize neurodevelopmental trajectories. Here we examine the empirical distributions of research-grade ADOS scores from 1324 records in a cross-section of the population comprising participants with autism between five and 65 years of age. We find that these empirical distributions violate the theoretical requirements of normality and homogeneous variance, essential for independence between bias and sensitivity. Further, we assess a subset of 52 typical controls versus those with autism and find a lack of proper elements to characterize neurodevelopmental trajectories in a coping nervous system changing at nonuniform, nonlinear rates. Repeating the assessments over four visits in a subset of the participants with autism for whom verbal criteria retained the same appropriate ADOS modules over the time span of the four visits reveals that switching the clinician changes the cutoff scores and consequently influences the diagnosis, despite maintaining fidelity in the same test's modules, room conditions, and tasks' fluidity per visit. Given the changes in

Neural Computation **32, 515–561** (2020)
https://doi.org/10.1162/neco_a_01263

probability distribution shape and dispersion of these ADOS scores, the lack of appropriate metric spaces to define similarity measures to characterize change and the impact that these elements have on sensitivity-bias codependencies and on longitudinal tracking of autism, we invite a discussion on readjusting the use of this test for scientific purposes.

1 Introduction

Autism is an umbrella term that groups a highly heterogeneous set of conditions, ranging from problems with abstract thinking within a social context to profound somatic sensory motor differences. Any random draw of the population with this diagnosis may have extremely different phenotypes. More important, it may have very different genotypes that go on to receive a similar diagnosis of autism (see Figure 1). This heterogeneity poses a problem to science because it becomes challenging to do basic research aimed at developing treatments that target the person's needs while leveraging the person's capabilities and predispositions to learn and adapt within natural and social environments. Inherent to neurodevelopment is the ability of the nascent human nervous systems to develop overcompensatory strategies to cope with a disorder, yet in the current diagnoses of autism, there is no room to extract what those coping capabilities are or how to foster them while treating the condition.

The diagnosis criteria of the DSM-5 from the American Psychiatric Association (APA, 2013) have broadened to include attention deficit hyperactivity disorder and sensory issues, while one of several psychological counterparts, the ADOS (Lord et al., 2000) can now include toddlers. With younger children receiving the diagnosis and broader criteria to diagnose, there are no appropriate medical interventions today that target the coping capacity of the nervous systems and identify in a personalized manner the best route to initiate treatment. Whether psychotropic drugs recommended by psychiatrists or behavioral treatments recommended by psychologists, the broad spectrum of autism today has no treatments that capitalize on what the nervous system already does well. There is a one-size-fits-all model to intervene from a very early age, informed and driven by a behavioral (observational definition) but no physical outcome measures of treatment effectiveness. Indeed, a recent report to the U.S. Senate[1] on the progress of the Autism Collaboration, Accountability, Research, Education, and Support (CARES) Act (2014)

[1] Report to the Committees on Armed Services of the Senate and House of Representatives, Department of Defense Comprehensive Autism Care Demonstration June 2018, *Report on Efforts Being Conducted by the Department of Defense on Applied Behavior Analysis Services*, requested by: Senate Report 114-49, p. 157, accompanying S. 1376, the National Defense Authorization Act for Fiscal Year 2016.

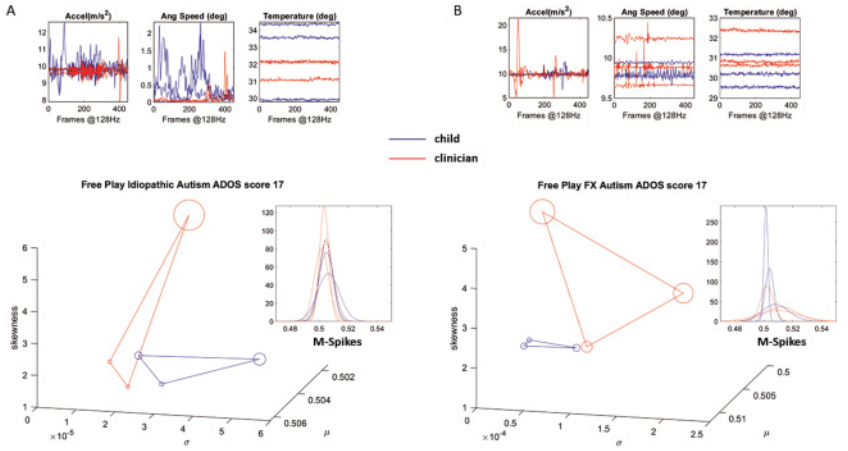


Figure 1: Impossible-to-stratify autism subtypes for research purposes with the Autism Diagnostic Observational Schedule, Module 2 (ADOS-2). Two participants with different phenotypes and different genotypes (A-idiopathic versus B-fragile X) received the same autism diagnosis from the ADOS-2 (score of 17 denoting autism) by the same clinician. Wearable sensors capturing 3.5 s of acceleration, body orientation rotations, and temperature show very different raw data waveforms from the child and clinician during the task Free Play of module 1 in the ADOS-2 test. Stochastic signatures derived from the fluctuations in bodily acceleration are shown as a map of empirically estimated gamma moments. These were derived from the fluctuations in the acceleration amplitude (i.e., the spike peaks) normalized to account for allometric effects due to anatomical differences. These normalized peaks' fluctuations converted to unitless micromovement spikes (M-spikes) are from the right and left wrists and torso of the child and the clinician. The raw data are from synchronously registered motions of their upper body as they socially interact during this task. The disparate signatures for these participant-clinician dyads are shown using empirically estimated mean (*x*-axis), variance (*y*-axis), skewness (*z*-axis), and kurtosis (proportional to the size of the marker) of the empirically estimated continuous gamma family (PDF insets). Notice that scales are different (for visualization purposes) due to large differences in data range.

signed into law by President Obama and extended on September 30, 2019, by the current administration, reveals that behaviorally defined interventions to treat behaviors such as those defined by the ADOS-2 instrument do not rise to the standards of the American Medical Association. As such, medical insurance providers will limit medical insurance coverage unless these behaviorally based approaches, as defined by these behavioral instruments, are medically relevant. The need to improve the medical research and the resulting medical treatments for autism is now more evident than ever before owing to the large aging adult autistic

population in need of medical support and corresponding medical insurance coverage in the United States.

Because the statistical assumptions underlying the behaviorally defined detection criteria that guide and inform the scientific research are not based on empirical data from physical measurements, but rather on assumed expectations defined by subjective observation, it is difficult to uncover and define medical target treatments tailored to the person's specific phenotypic and genotypic characteristics and aimed at treating the medical issues. This is so because this important capacity for adaptation in the autistic nervous systems remains hidden to the naked eye of the observer trained to catch pre-set expected aspects of social responses to specific social presses. Social behavior is much too complex and dynamic to compartmentalize in such ways. In so doing, one risks a gross loss of data that is relevant to these pressing medical issues. Such information can also be of use in stratifying the many subtypes of autism that we now see in our labs. We see and quantify (e.g., using advanced wearable instruments) a variety of medical conditions in children who have identical autism scores (e.g., as in Figure 1)—for example, dysautonomia (dysregulated heart rhythms, food aspiration owing to swallowing issues, peristalsis dysfunction, sphincter dysfunction, gut autonomy dysfunction, delayed reflexes, and seizures, among others), excess tolerance to pain, temperature dysregulation, metabolic dysregulation, altered microbiota, and an overall profound lack of autonomous neuromotor control (e.g., frequent falls, vestibular dysfunction, abnormal vestibulo-cochlear and vestibulo-occular reflexes, balance issues, gait abnormalities). While these were rendered "comorbid" conditions by the subjective psychological and psychiatric instruments that behaviorally define autism, their prevalence among the population has now alerted medical insurance providers in the United States to the urgent need to address these medical issues in both basic and translational research. Indeed, the Autism CARES Act approved in 2019 budget of $3.87 billion to that end.

In recent years, the scientific community has raised the need to stratify autism into various autism subtypes to facilitate the path toward more appropriate, personalized treatments. But reliance on instruments that have no physical measurements and rely exclusively on behavioral observation and behavioral criteria impede progress toward a personalized approach. Indeed, several debates have been published surrounding controversial reliance on the adoption of the research-grade ADOS (Lord et al., 2000) for use in scientific autism research (Constantino & Charman, 2016). The field seems to have reached a point where some autism researchers advocate the importance of tracking and quantifying the biorhythms of the nervous systems (Constantino et al., 2017; Klin, 2008; Klin, Jones, Schultz, Volkmar, & Cohen, 2002a, 2002b; Tordjman et al., 2015; Torres, Brincker, et al., 2013; Torres & Denisova, 2016; Torres, Isenhower, et al., 2016) and the use of biophysical data to adapt tenets of precision medicine (Hawgood, Hook-Barnard,

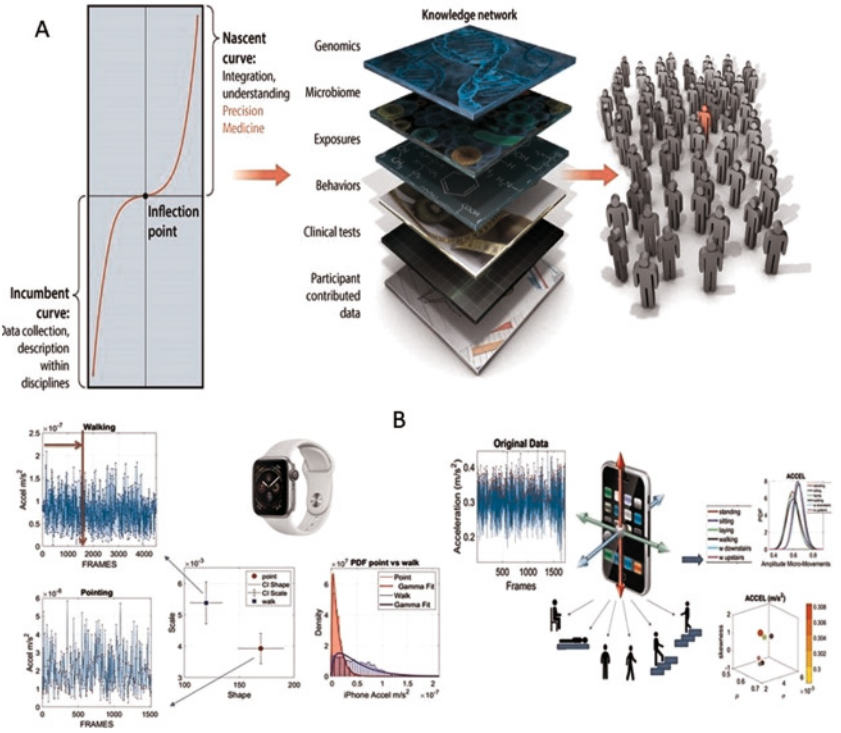


Figure 2: The precision Medicine paradigm connecting different layers of the network of knowledge to develop personalized target treatments. When translating this model to psychiatry and psychology, observation of behavior can be complemented by digital data from wearable biosensors. (Panel A from Hawgood, Hook-Barnard, O'Brien, & Yamamoto, 2015. Reprinted with permission from AAAS.) (B) Commercially available wearable biosensors affording the level of precision of research-grade sensors to, for example, separate activities of daily living that involve voluntary versus involuntary motions. Panel B shows the output of the Apple watch distinguishing walking from pointing activities, along with the categorization of activities of daily life from accelerometer data collected with a smart phone. (Panel B from Torres, Vero, & Rai, 2018.)

O'Brien, & Yamamoto, 2015; Insel, 2014) to a nascent field of precision psychiatry (Friston, Redish, & Gordon, 2017; Torres, Isenhower, et al., 2016).

The precision medicine paradigm (see Figure 2A) aims for the development of personalized targeted drug treatments that successfully combine different layers of the knowledge network spanning from patients' self-reports to genomic data. In the fields of cancer research, the tenets of precision medicine have been implemented, and several target treatments

have been successfully developed. Yet adoption of this paradigm has lagged in fields that deal with neurological and neuropsychiatric disorders. These fields rely heavily on observation of behavior and behavioral definitions of clinical phenotypes. As a result, methods that describe behavioral phenotypes through patients' self-reports or through scoring reports by accredited professionals remain disconnected from basic scientific methods that assess various underpinnings of behavior through objective physical means with the purpose of defining medical conditions (e.g., neurological, immunological, endocrine/metabolic). This is perhaps the case because research-grade instrumentation used to perform such measurements in the laboratory settings had not been widely available to the clinical field until the very recent revolution in wearable biosensors. The advent of wearables that nonintrusively record biorhythmic activities self-generated by the nervous systems, has made such level of precision to measure behaviors commercially available to all (e.g., see Figure 2B).

Other fields have adopted the digital technology and begun the path of integration with clinically validated inventories. For example, recent work in Parkinson's disease (PD) has digitized the Universal Parkinson's Disease Rating Scale (UPDRS) and used it in research settings to help stratify different subtypes of PD (Ryu, Vero, Dobkin, & Torres, 2019). Similarly, the digitized ADOS (Whyatt & Torres, 2017) is amenable to integrating clinical behavioral criteria defining and detecting autism with digital biomarkers of behavior to help stratify different subtypes of autism (see appendix Figures 12 and 13) and redefine the behavioral phenotype with far higher underlying precision.

As autism is a lifelong condition, those who were diagnosed in the 1970s through behavioral criteria are aging autistic adults today. Yet with the prevalence since then of behaviorally defined therapies devoid of neurological criteria, their nervous systems never received neuroprotective therapies to scaffold autonomy and promote agency. It has been reported that among the autistic aging adults' over 40 years of age, symptoms of motor disorders and Parkinsonism are far more prevalent (20%) than among those over 65 years of age without autism (0.9%) (Starkstein, Gellar, Parlier, Payne, & Piven, 2015). These findings extend to excess involuntary movements in autistics 5 to 40 years of age, even for autistic individuals who did not take psychotropic medications with motor side-effects (Torres & Denisova, 2016). The current results point to an accelerated path toward neurodegeneration that calls for medically based research, defined according to medical and physiological standards, now absent from the (behaviorally defined) detection criteria.

The motivation behind our work has been to extend the monologue style of ADOS scoring highlighting social deficits to a dialogue dyadic style that aims at identifying and scoring inherent capacity for social exchange, already present in the child's nervous systems—that is, those escaping observation. Such hidden capabilities in autistic individuals are not immediately

obvious to the naked eye of a clinician trained to expect certain aspects of the social behavior to be present during the exchange and coding only those expectations. The potentially overcompensatory biological coping strategies are capturable in the continuous digital data of the dyad. In this sense, such data could give us an entry point into the nervous systems' self-generated biorhythms to help create support and augment sensory systems for the affected person. As a result of this more holistic approach to autism, we could help treat the child's somatic-sensory-motor systems, scaffolding all aspects of neurodevelopmental readiness for social exchange and supporting the person with age-dependent accommodations across the life span.

In the process of testing the robustness of the biometrics of social exchange that we derived from the digitized ADOS scoring across different testing contexts, we had different clinicians test the same participant within the same room layout and same ADOS module. To our surprise, we captured dramatic differences in the bodily responses of the same child to different clinicians administering the same module and testing the child under similar room conditions (digital results reported elsewhere). Because these changes in the digitized behavioral responses of the child are not included in the ADOS criteria and because very likely these activities are largely beneath the clinician's awareness, away from naked eye detection capacity, we decided to examine the scores that the two clinicians conferred to the child's responses. Here, we ask whether such subtle differences in the child's behavioral outputs would also affect clinicians' scoring.

As the child's bodily and facial micromotions were different for each clinician, it may also be the case that the type of visual feedback that these subtle motions from the child offer to the clinician change the diagnosis for the same child in a clinician-dependent manner. Did the subtle changes in the micromotions of the child rise to the level of detection such that the scoring changed? We found large differences in clinicians' ratings of the same child, for the same module of the test administered under similar room conditions. This discrepancy motivated us to further examine these phenomena more systematically in our lab and to study the ADOS scores more generally in large, open-access data sets. What are the statistical features of these empirical scores reported by research-reliable ADOS testers? And what can these scores tell us about the form of autism that we know today in the research arena?

In this study, we first assess clinicians' ADOS scoring of a modest cohort of autistic children, using the ADOS as an experimental protocol, where we vary the clinician and the module type for the same child. We then, for the first time, compare the performance of the ADOS in neurotypical children of different ages. And finally, we examine the statistical features of these scores taken for a large cohort, across different research sites in the United States and abroad. These scores are openly accessible in the Autism Brain Imaging Data Exchange (ABIDE) repository. We discuss our results considering the use of this test in basic scientific research to guide and inform

genetic research in autism aimed at developing target treatments within the tenets of precision medicine.

2 Methods

The Rutgers University Institutional Review Board (IRB) committee approved all protocols used in this study. Parental and legal guardian consent was obtained for all participants. All procedures were performed in compliance with the Helsinki Act under IRB approval. For the consent process, we read and explained to the parents the IRB-approved protocol and the child assent form. After the informed-consent process, all families agreed to participate and signed written parental permission and assent forms. The participants' names and identifiers were removed to maintain confidentiality across the entire analyses. Further, deidentified data were used in all sections of this article, including the tables.

2.1 The ADOS Assessment. The ADOS assessment is a tool to aid the diagnosis of autism. It is often used in combination with other tools such as the *Diagnostic and Statistical Manual* (DSM-5; American Psychological Association, 2013) and the ADI-R screening tool. ADOS had at some point four modules (now five modules and training for toddlers' and adults' levels are more common). Each module is designed to provide the most appropriate test for an individual at a certain language level. There is a newly created calibrated severity score (CSS) in ADOS-2. It is based on the person's age as it converts an individual's total ADOS-2 score in comparison to other individuals with ASD at the same age and language level (for each individual module). For the actual ADOS-2 administration, age does not determine the module but may determine algorithm items within that module. We note, however, that two children of the same age will likely have very different neuromotor control age, as assessed by objective physiological metrics. In general, age has no real meaning in neurodevelopmental disorders, where the coping nervous systems of different individuals born at similar dates evolve at very different rates (Torres, Brincker, et al., 2013).

In this experimental setup, the most appropriate module was determined by two experienced clinicians: a developmental pediatrician and a developmental clinical psychologist. Two clinically certified raters independently videotaped and discussed the sessions to ensure module administration fidelity. The clinicians administering the test were not involved in any aspect of the design of this study.

The two clinicians who administered the ADOS to the participants were research reliable: they are trained professionals who have undergone ADOS-2 clinical training. ADOS-2 clinical training is an introductory workshop with instructional methods that include lectures, videos, demonstrations of administration and scoring, and discussions. This workshop serves

as a prerequisite for more thorough training required to obtain research reliability needed to use the ADOS-2 for research purposes.

"Research reliable" means that prospective trainers have submitted ADOS-2 administration videos to fellow research-reliable examiners and have received at least an agreement of 80% of similar scores on the ADOS-2 administration as displayed through the submitted video. The prospective research-reliable trainee and the already research-reliable trainer have at least 80% similar scoring on the ADOS-2 administration video submitted by the trainee. Three videos must be submitted, and a minimum of 80% similar scores must be achieved. Once they passed the requirement, they became research-reliable ADOS raters. The information about four of the modules that we used is taken from the manual and briefly summarized below but are not meant to be complete (see the supplementary table) and when in doubt, consult the ADOS manual:

- Module 1: This is designed for individuals who do not have consistent verbal communication skills. The tasks use completely nonverbal scenarios for scoring.
- Module 2: This is designed for individuals who have minimal verbal communication skills, including young children at age-appropriate skill levels, whereby tasks require moving around the room and interacting with objects. All objects are standard and come in a standardized kit.
- Module 3: This is designed for individuals who are verbally fluent. These participants are also capable of playing with age-appropriate toys. The test is conducted largely at a desk or table. In our setup, the table was always the same, and it was positioned in the room in the same configuration. The room where the test took place was not changed from session to session and participant to participant. Researchers and ADOS-certified personnel in the study (four total) made sure that the conditions were identical in each session, task, child, rater, and module.
- Module 4: This is designed for individuals who are verbally fluent but no longer at an age to play with toys. This module incorporates some module 3 elements, yet it is more conversational regarding daily living experiences.

Often the examiner chooses a module and then realizes that the participant's functional abilities anticipated by that module do not match the rater's expectation. Then the tester chooses another module. This is a common practice. Therefore, our experiment manipulated the module type to probe participants' responses to the same module administered by two different raters or to determine the adequacy of the modules.

The modules involving playing with toys or objects have the tester present standardized scenarios to evoke responses and rate the child's performance. Some elements of the game (e.g., a puzzle) are left out on purpose

to evoke the child's need to ask for other pieces. The examiner uses several strategies to evoke different responses and assess the child's reaction. This mode of testing behaviors inherently makes certain assumptions about social expectations that do not necessarily transfer from culture to culture. Further, the ADOS states that no sensorimotor issues are present without ascertaining the intactness of the child's peripheral nervous systems before administering the test. A child response may not be voluntary. When the nervous systems are damaged, there is an inevitable component of the response that is never evaluated during the test because of the test's assumption that no significant sensory and motor issues are present. As such, it is never possible to assess causality. For all these reasons and because our lab receives a highly diverse population from many different cultures and neuromotor developmental stages, the experimental protocol manipulated the variable representing the rater while maintaining constant all other conditions (e.g., context, module, room, layout of all objects in the room) and videotaping the sessions by other two independent ADOS-certified raters. As it is required by the ADOS, raters flexibly adapted each session to the child's responses to the flow of the tasks while maintaining fidelity to the tasks employed in each module of choice.

The response of the child determines the score. Likewise, the way in which the rater evokes the response influences the child's choice of actions that are consequential to the rater's provoking actions. To probe the extent to which a change in the tester influenced the scoring, our experiment manipulated the rater as a parameter while holding all other conditions constant in two visits. Participants had four visits to the lab. For each participant, two modules were selected, and research-reliable testers were employed. One module was rendered the more adequate one, while the other was rendered the feasible one. By "more adequate," we mean that the module was at the child's verbal and developmental levels; "feasible" means that the child could perform the entire module, but it would not be the adequate one to perform a diagnosis or aid a clinician in performing a diagnosis of autism. We note that previous research indicates that inappropriate ADOS module use invalidates the assessment and the scores do not accurately reflect the child's performance on the assessment. Nevertheless, since this study is not about diagnosing autism but about evaluating the use of this ADOS test in basic science, specifically assessing the variability that using different raters may add to the scores, in addition to changing the rater, we are also manipulating the use of the modules across visits.

Each module took between 40 and 60 minutes to complete. Both the rater and the participant were recorded by two video cameras from different angles and by smart sensors that they wore embedded in the clothing, as watches on the wrists and on the ankles. (The digital data will be the subject of a different publication. Additional information about the ADOS test can be found in the supplementary table.)

2.2 Assessment of the ADOS Scores in the Lab. We measured the outcome of the ADOS-2 test in 52 individuals: 26 controls (ages 7 to 66 years old) and 26 suspected to have (and then diagnosed with) autism spectrum disorders (ages 4 to 20 years old). We used the baseline visit 1 to characterize the participants diagnosed with ASD and those who were neurotypical. This was done to ascertain the extent to which the scores' range from typical control participants deviate from the 0 scores denoting the absence of behavior otherwise present and contributing to the overall cutoff number. We were motivated by the critical need to create a similarity metric for autism research measuring departure from normative data. Given the adoption of the research-grade ADOS for research and the fact that this observational inventory is not a norm-reference test (Lord et al., 2000), we ascertained the spread of scores obtained from neurotypicals. This gave us a sense of departure from typical ranges.

The absence of neurotypical assessment using the ADOS is not well known in the community that adopts the test for research. We quote from the ADOS-G paper: "Replication of psychometric data with additional samples including more homogeneous non-Autistic populations and more individuals with pervasive developmental disorders who do not meet Autism criteria, establishing concurrent validity with other instruments, evaluation of whether treatment effects can be measured adequately, and determining its usefulness for clinicians are all pieces of information that will add to our understanding of its most appropriate use." Table 1 shows the 52 participants' scores and ages at the baseline visit, when the most appropriate ADOS-2 module was selected for each child.

2.3 Repeated Measurements of Scores (But Not as Longitudinal Assessment of Change). In 14 of the individuals with ASD (mean 9.3 years old $\pm$ 3.0), we reassessed them across four visits taking place within 1.3 years on average ($\pm$6 months) to determine the extent to which switching the clinician or the ADOS module (or both) would change the outcome of the test for the same child. To that end, for each child, in the first two visits, we used the same clinician but employed two different ADOS modules. According to each assessment in each visit, the raters determined the modules that were the most appropriate and feasible. From these assessments, the 14 individuals described are those for whom the second round of visits (visits 3 and 4) retained the same appropriate and feasible modules despite the passage of time.

The first module (visit 1) determined the most appropriate module at baseline for the given child. The second module (visit 2) was feasible (the child could do it) but not appropriate. For example, if the most appropriate module in visit 1 was module 3, we would choose module 2 for visit 2. Then the same clinician would give these two modules whenever the participant retained them. That is, the clinician determined that the module for visits 3 and 4 was the same as the module for visits 1 and 2. Then the set of modules from visits 1 and 2, according to the raters' evaluation, was administered in

Table 1: Baseline ADOS-2 with 26 Participants with ASD and 26 Controls.

ID ASD	Age	V1 Mod	V1 SA	V1 RRB	V1 Total	ID Ctrl	Age	V1 Mod	V1 SA	V1 RRB	V1 Total
1	4	3	7	3	10	1	8	3	0	0	0
2	8	3	6	3	9	2	10	3	0	0	0
3	10	3	14	3	17	3	9	3	1	1	2
4	13	3	6	1	7	4	12	3	2	0	2
5	6	3	6	3	9	5	7	3	0	0	0
6	6	3	6	3	9	6	7	3	2	0	2
7	11	3	10	2	12	7	10	3	4	5	9
8	5	1	15	3	18	8	9	3	0	0	0
9	9	1	11	2	13	9	7	3	1	0	1
10	6	3	14	3	17	10	11	3	1	0	1
11	14	1	7	1	8	11	7	3	2	0	2
12	10	1	11	6	17	12	15	4	1	1	2
13	4	3	9	2	11	13	11	4	1	0	1
14	11	1	13	3	16	14	13	4	0	0	0
15	9	3	6	4	10	15	31	4	1	0	1
16	7	3	6	2	8	16	49	4	2	0	2
17	7	3	9	2	11	17	48	4	0	0	0
18	11	3	10	1	11	18	38	4	0	0	0
19	4	1	20	6	26	19	29	4	0	0	0
20	8	3	13	5	18	20	30	4	0	0	0
21	10	1	16	8	24	21	32	4	0	1	1
22	13	1	11	8	19	22	22	4	1	2	3
23	10	2	15	8	23	23	48	4	0	0	0
24	4	2	2	4	6	24	20	4	7	1	8
25	18	3	15	6	21	25	66	4	0	0	0
26	20	4	6	1	7	26	43	4	2	0	2

those subsequent visits (by a different rater): module 3 in visit 3 and module 2 in visit 4 in this example (see Table 2).

We switched clinicians and maintained module fidelity and identical room setup. In this way, each child had a chance to become familiar with the two modules by the time that we switched the clinician. Those two same modules would then be fluidly administered by the new clinician to give us the opportunity to probe the influences of the clinician on the child's response. The flexibility in task administration according to the child's responses was respected to ensure fluid responses.

We hypothesized that this switching of clinicians (despite the use of the same modules' and tasks' order in each administration) would have a substantial effect on the ADOS subscores, thus significantly affecting the reliability of the total score and the cutoff for the diagnosis given to the child by the rater (clinician). To test this hypothesis, we used nonparametric statistics whereby we do not assume any distribution a priori. Table 2 shows the 14 scores across the four visits.

Table 2: Repeated Assessment of 14 Participants across 4 Visits, 2 Different ADOS-2 Modules, and 2 Raters (Aut Autism, S Autism Spectrum).

ID	Age	V1 Mod	V1 SA	V1 RRB	V1 Tot	V1 Dx	V2 Mod	Age	V2 SA	V2 RRB	V2 Tot	V2 Dx	V3 Mod	Age	V3 SA	V3 RRB	V3 Tot	V3 Dx	V4 Mod	Age	V4 SA	V4 RRB	V4 Tot	V4 Dx
1	4.3	3	7	3	10	S	2	5.2	5	2	7	S	3	5.9	4	5	9	S	2	7.3	4	4	8	S
2	8.9	3	6	3	9	Aut	2	9.2	8	0	8	S	3	10.5	7	4	11	Aut	2	10.8	8	1	8	S
3	10.1	3	14	3	17	Aut	2	10.4	10	3	13	Aut	3	10.9	16	8	24	Aut	2	11.3	14	7	21	Aut
4	12.5	3	6	1	7	S	4	13.9	2	6	8	S	3	14.1	9	2	11	S	4	14.6	2	6	8	S
5	6.6	3	6	3	9	Aut	2	6.8	6	1	7	S	3	7.2	5	3	8	S	2	7.4	4	3	7	S
6	12.1	3	10	2	12	Aut	2	12.5	8	1	9	S	3	12.1	12	6	18	S	2	13.2	8	5	13	Aut
7	14	1	7	1	8	S	2	14.3	8	2	10	S	1	14.7	9	5	14	S	2	15.1	9	6	15	S
8	10	1	11	6	17	Aut	1	10.4	10	5	15	S	1	10.7	17	9	26	Aut	1	11.3	14	7	21	S
9	4	3	9	2	11	Aut	2	4.5	8	3	11	Aut	3	5	15	7	22	S	2	5.4	13	7	20	S
10	11.6	1	13	3	16	Aut	1	11.7	16	2	18	Aut	1	12.1	11	7	18	Aut	1	12.4	11	8	19	Aut
11	9.2	3	6	4	10	Aut	2	9.6	10	1	11	Aut	3	9.1	10	6	16	Aut	2	10.2	4	6	10	Aut
12	8.2	3	6	2	8	S	2	8.7	8	0	8	S	3	9.2	6	2	8	S	2	9.8	7	3	10	Aut
13	7	3	9	2	11	Aut	2	7.4	9	2	11	Aut	3	7.9	11	6	17	Aut	2	8.1	12	4	16	Aut
14	11.9	3	10	1	11	Aut	2	12	8	1	9	Aut	3	12.4	6	2	8	Aut	2	12.1	7	4	11	Aut

Finally, since the passage of time inevitably affects any longitudinal study and since development occurs in these children under unexpected or hidden coping mechanisms of their physical development, we also normalized by age at the time of each visit. In addition to the use of the absolute scores above (to test for absolute value effects), we use relative scores (to test for derivative effects) assessing the rate of change, in scores per age quantity. The latter denote dynamic changes expected to occur in childhood owing to the nonlinear accelerated and irregular nature of neurodevelopment (Torres, Smith, Mistry, Brincker, & Whyatt, 2016). As in the previous analyses, we here employed nonparametric tests.

2.4 Assessment of the ADOS Scores in the Open Access Autism Brain Imaging Data Exchange Repository (ABIDE). The ABIDE records contain ADOS-2 and ADOS-G scores that we extracted to plot the distributions of these tests' subscores across 1324 participants, ranging between 5 and 65 years of age. We present the distribution of participants in Figure 3 comprising clinical records from ABIDE I and II. Preliminary results from this work involving involuntary head motions during resting state in fMRI studies were published elsewhere (Caballero, Mistry, Vero, & Torres, 2018; Torres & Denisova, 2016; Torres, Mistry, Caballero, & Whyatt, 2017). Here we focus on the nature of the distributions of the ADOS scores, assumed to be symmetric by the ADOS manual and by reliability and validity tests commonly used in clinical psychology to validate this diagnostic tool driving basic scientific research (Havdahl et al., 2017, 2016).

Clinical tools for diagnoses use the tenets of signal detection theory (SDT; Swets, 1996; Swets & Pickett, 1982), albeit in a black box approach that has yet to verify the implicit assumptions made by the statistical packages that such papers report. It has been suggested that nonparametric methods to correct for nonnormality may be inadequate in a diagnostics test (Witt, Taylor, Sugovic, & Wixted, 2015). This may be even more problematic in a test that has not mapped out the noise levels inherent in the person's behaviors. For example, in the ADOS, as in many other clinical tests, the clinician plays the dual role of being the stimulus (via the prompts to evoke social overtures or primed responses, or both) and, at the same time, the observer, scoring the participant's response. The participant's responses contain motor noise with specific age-dependent signatures (Torres, Brincker, et al., 2013) that are not currently considered in the ADOS scores. Nevertheless, they are a subliminal source of information (i.e., through visual feedback) that the raters may unconsciously use to determine the scores that eventually help detect autism.

The detection step is hindered by the very outcome of this test, which is aimed at measuring social interactions between two people. The process of scoring is heavily one-sided. The outcomes depend on the clinician's observation. In the absence of the required assumptions of normality and variance homogeneity, the sensitivity (d') and response bias (β) do not

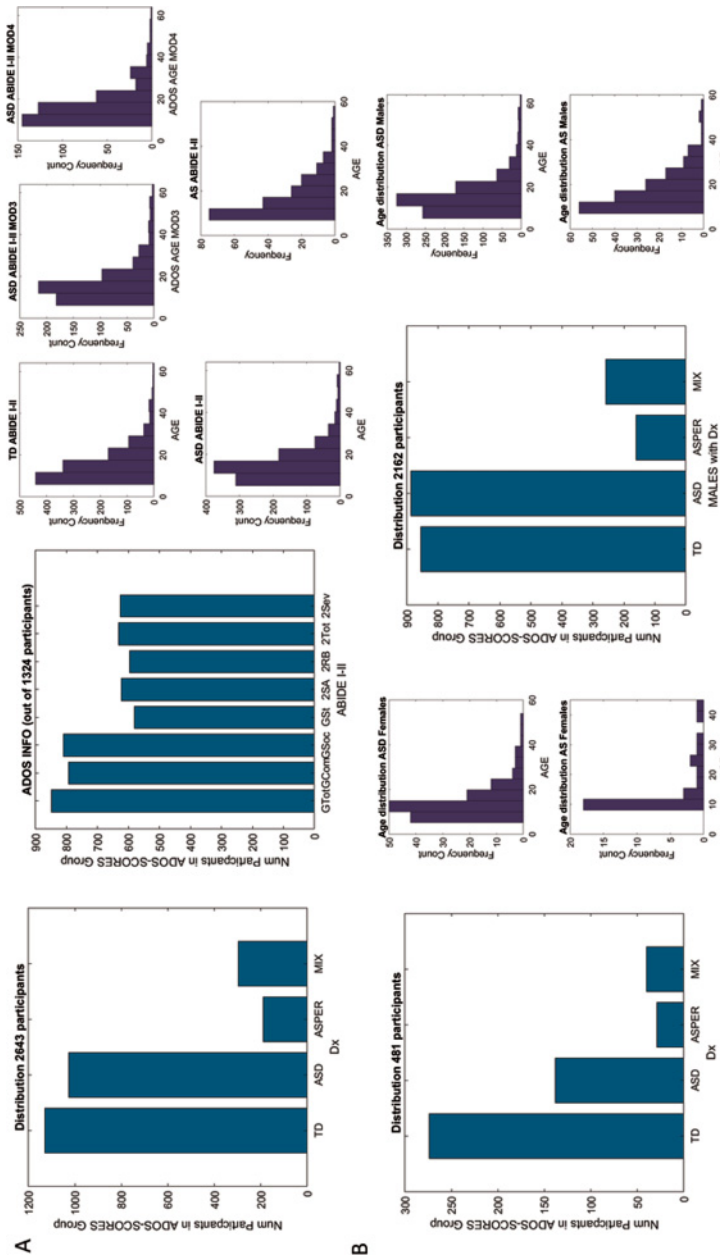


Figure 3: ABIDE data sample. (A) Participants (2643) according to the diagnosis and ADOS information from 1324 participants in the spectrum of autism divided by sex: 481 female and 2162 male participants (B). Histograms are also divided by age and sex for TD, ASD, AS, and distributions of participants who performed module 3 or module 4.

represent accurate and independent measures (Swets & Pickett, 1982; Witt et al., 2015). There are codependencies between them. When we add subtle micromotions present in the bodily biorhythms of the child, noise that comes from them in autistic individuals may also have an impact on the visual feedback that the clinician receives. How this inherent motor noise affects scoring is unknown (via visual feedback and unconscious mirroring) but may be a contributing factor in the scoring of the child's responses.

Because of the broad use of this test in autism research and the possibility that the assumptions required for research validity may not be met, it becomes crucial to verify the normality assumption of the distributions of scores and subscores, along with the assumption that the rater's bias and response are independent. To that end, we take the unique opportunity that ABIDE offers with thousands of records providing the ADOS-G and ADOS-2 versions of the test for DSM-IV and DSM-5 criteria, and we examine the distributions of the ADOS scores. We gather the scores in frequency histograms built using various binning procedures (Freedman & Diaconis, 1981; Scott, 1979, 2015). Then we fit the normal distribution and other distributions (e.g., log normal, exponential, gamma, Weibull) using maximum likelihood estimation (MLE) methods in Matlab.

In addition to the MLE tests, we use the Lilliefors test (Lilliefors, 1967) and the Kolmogorov-Smirnov test to compare theoretical distributions to the empirical one we get from the ABIDE data sets.

Besides testing the distribution of ADOS-G and ADOS-2 scores across the full set of ABIDE, we also separate the clinical data by module 3 versus module 4. Further, we separate the data from the females and the males in the ABIDE repository, which contains an unusually large number of females (absent in any random draw of the population). Typical females, females with ASD, and females with Asperger's syndrome can be physiologically stratified and separated from the males using involuntary micromovements (Torres, Isenhower, et al., 2013; Torres et al., 2017) and voluntary movements (Torres, Isenhower, et al., 2013). Here we ask if the ADOS scores of ABIDE separate them too or if, unlike physiological criteria, the clinical ADOS scores confound males and females. Note that in any regular study, this question cannot be asked owing to the near five-to-one autistic male-to-autistic female ratio in the population. A random draw of the autistic population would not give us enough power to assess the ADOS scores in males versus females.

All studies in the ABIDE data repository were performed under IRB approval in accordance with the Helsinki Act.

3 Results

3.1 Normative Data Spread Significantly Departs from Lowest-Bound 0-Score. The analyses of the in-person visit to assess the ADOS-2 in 52

participants—26 suspected (and confirmed) as having autism or ASD and 26 typical controls—yielded significantly different distributions of scores between the two age- and sex-matched groups. Figure 4A shows the frequency histograms of the ASD (top) and typical control (bottom) groups. Figure 4B shows the outcome of the nonparametric Kruskal-Wallis (one-way ANOVA) test highlighting the statistical significance of differences captured by the comparison. Furthermore, the right panel of Figure 4B highlights the output of this nonparametric test for the comparison of the 0-score, denoting the absence of behavioral symptoms. For typical controls, the scores from the empirical data spanned a distribution of nonzero values, evident at the $p < .01$ level.

3.2 ADOS-2 RRB Outcome Is Significantly Affected by Clinician. The data from the repeated measurements across four visits enabled us to examine the influences of the clinician in 14 children who returned to the lab to perform the ADOS-2 test. The scores from all the children were pooled, and the total score was compared across visits using the nonparametric Kruskal-Wallis test followed by the multicompare test. The total score comparison revealed no significance (chi square, 7.21; *p*-value, 0.06). Yet given the borderline value close to the 0.05 significance level, we examined the social affect score and the ritualistic repetitive behavior (RRB) score making up the total. We found no differences across visits in the social affect score. However, the RRB score significantly changed across visits (chi square, 21.01; *p*-value, 0.0001) with major differences when switching clinician in visits 3 and 4. Despite the use of the same modules, room setup, and task fluidity for each child, the differences in ADOS-2 scores for RRB were marked as systematically different by the post hoc multicompare test. These outcomes can be appreciated in Figure 5 for total (panel A), social affect (panel B), and RRB (panel C).

We further tested all the scores for each clinician by pooling across all children and score type to examine the types of distributions best fitting their frequency histograms. Figure 6A shows this analysis for each clinician, while Figure 6B shows the output of the nonparametric Kruskal-Wallis test, which revealed statistically significant differences. Figure 6C shows the failure of normality for scores by clinician 1, and Figure 6D shows that for clinician 2. The use of MLE to ascertain the fit of several probability distribution functions confirmed that the normal is not a good fit for either (see Figure 6E). Further, the gamma distribution was used as per the MLE outcome to fit the data and compare the scores of the two clinicians for the same children, same modules, same module order or visit, and same order of tasks. Figure 6F shows the fits of the normal distribution (left-hand side) and the gamma distribution (right-hand side). The gamma distribution fit was best for clinician 2 and poor for clinician 1. The normal was poor for both. The log normal and exponential were also poor fits.

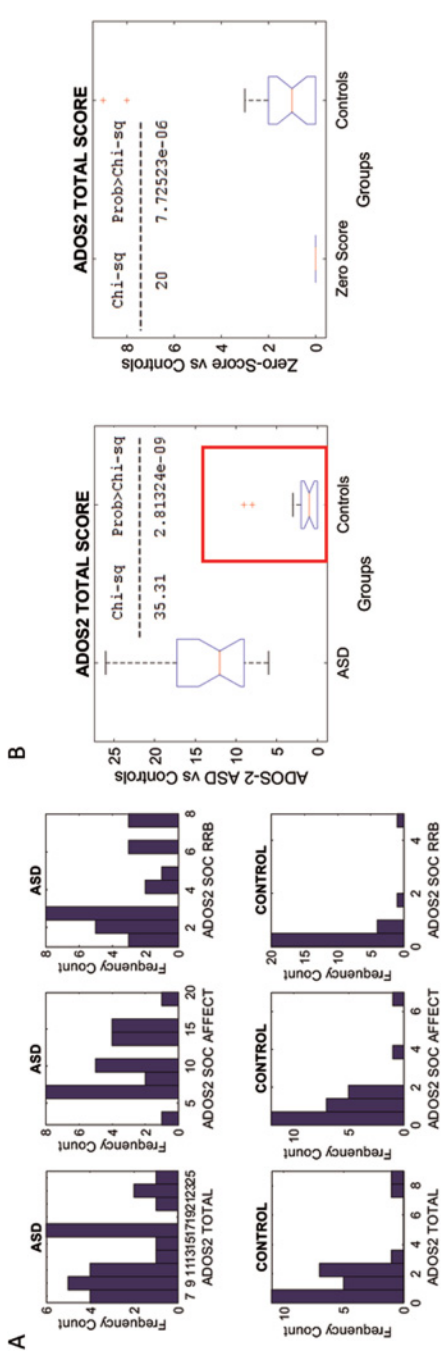


Figure 4: The need for normative data to build a true similarity metric for autism research that adopts the ADOS test. (A) Frequency histograms of scores from the ADOS-2 rated scores for individuals with ASD and autism cutoffs versus those of typical controls. (B) Differences with statistical significance between the two groups' total scores and (right panel) a distribution of scores for typical controls calling for a reassessment of the ADOS scoring system to convert it from a criterion-referenced to a norm-referenced system and build an appropriate metric space, amenable to measure change and derive developmental trajectories in basic scientific research.

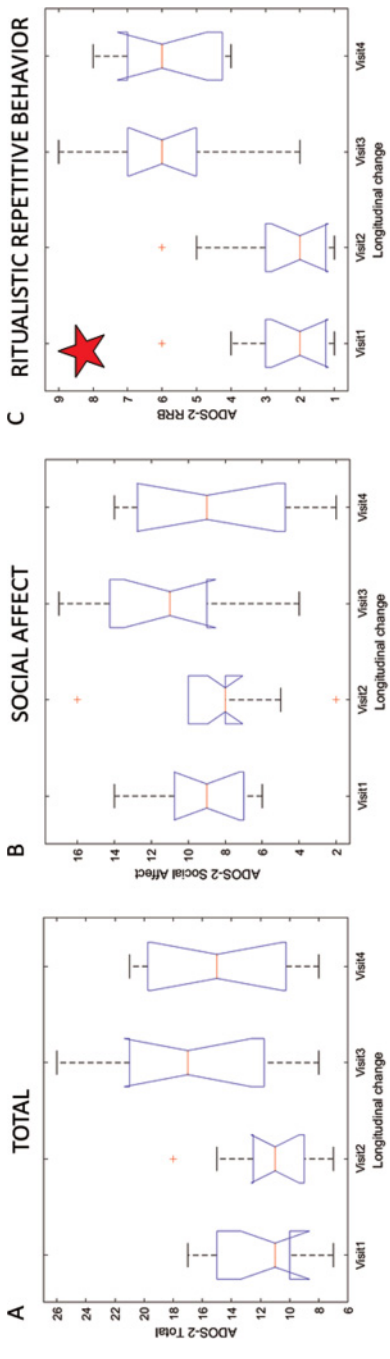


Figure 5: Susceptibility of the RRB ADOS-2 subscore to the rater clinician. (A) Output of the Kruskal-Wallis nonparametric one-way ANOVA using box-and-whiskers format shows differences in total score between clinician 1 in visits 1 and 2 and clinician 2 in visits 3 and 4 but these differences do not reach statistical significance ($p < 0.06$). Notice the broader ranges in the score distribution of the clinician 2 and the presence of higher scores overall. (B) Social affect subscores were not significantly different between clinicians. (C) RRB scores were significantly different for the same children, same modules, and same task order. Clinician 1 rated the children significantly lower, with overall ranges contributing to a bias toward autism spectrum. In contrast, the higher values of the rates by clinician 2 contribute to a bias toward autism diagnosis for the same children.

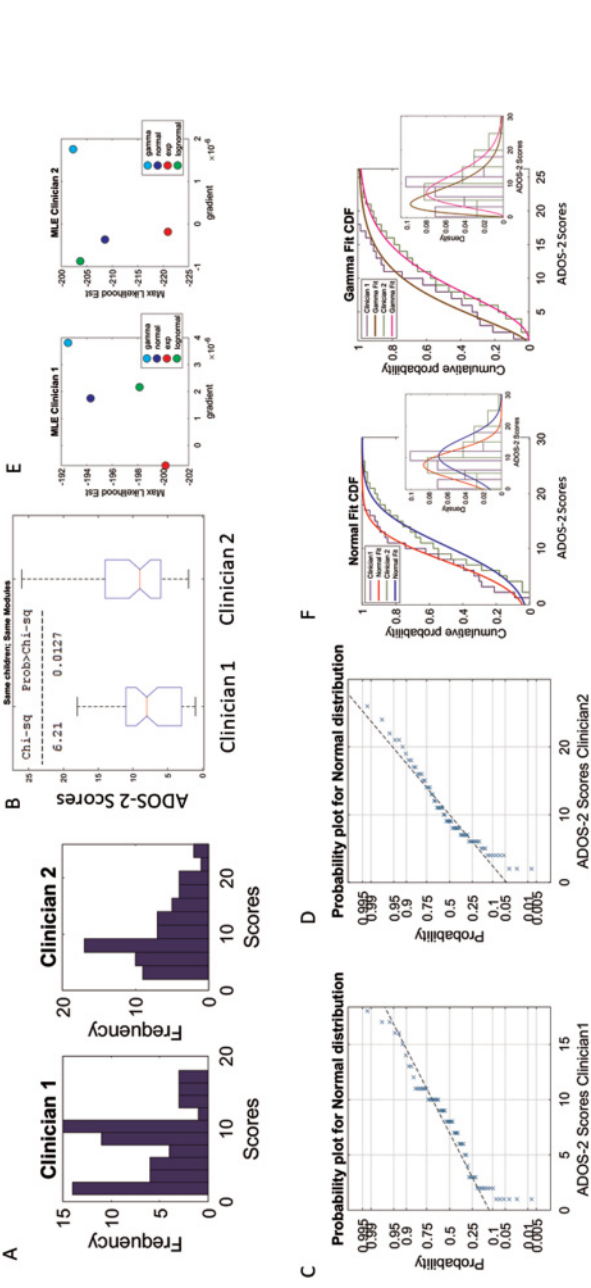


Figure 6: Nonnormality of overall scores from each rater clinician. (A) Differences in rater clinician styles can be appreciated in the frequency histograms of the scores that each one obtained from the same cohort. (B) Differences in scores for the same cohort reach statistical significance (despite the relatively small number of measurements: 14 children × 4 visits × 2 subscores, 112 scores). (C,D) Nonnormality of the score samples from each rater clinician. (E) MLE output renders the normal distribution inappropriate in each rater clinician case. (F) Fitting the normal versus the gamma family for each score set empirical CDF underscores the fundamentally different stochastic signatures derived from the variability of their rating of the same cohort of children with the same ADOS-2 modules.

3.3 Age-Corrected ADOS-2 Derivative Scores Confirm the RRB Score as the Most Affected by Change in Rater Clinician. The relative changes in score and age (with the age measured in years, months) were obtained for each of the 14 participants we tracked over four visits. When examining these derivative data, we found significant differences in the RRB ADOS-2 scores. Consistent with the effect that the change of clinician in visit 3 had revealed for the size data given by absolute ADOS-2 scores, here, the derivative data considering the age change of the participant from visit to visit also reveal significant changes in the RRB scores. The same individuals were rated significantly different by the clinicians, thus yielding different scores for the same module and tasks. The Kruskal-Wallis test for the comparison of the ADOS-2 RRB scores across visits yielded significant differences across visits (chi square, 13.45; *p*-value, 0.003).

Figure 7A shows the evolution of the clinician' diagnostic criteria over time. Visits 1 and 2 with clinician 1 show 10 of 14 with autism versus 4 of 14 with ASD. This pattern changes in visit 2 for this rater clinician to 6 of 14 with autism and 8 of 14 with ASD. The second clinician rather differently scores the same cohort of children performing the same tasks under the same room setup, from the same modules, compared to the rater clinician 1 in visits 1 and 2. In visits 3 and 4, rater clinician 2 scores these same children as 50-50 autism-ASD. The individual evolution for each child is seen in Figure 7B, whereby the different clinicians' styles of scoring can be seen. There is no interrater reliability rendering these ADOS-2 criteria robust. For the same child and same ADOS-2 module, we see changes in the classification of autism versus ASD. These differences in perception biases for this lab cohort will be examined next in relation to the distributions of scores derived from the large cross-sectional population data from ABIDE.

3.4 ADOS-G and ADOS-2 Scores Do Not Distribute Normally. The ADOS-G and the ADOS-2 scores for each of the criteria comprising the total, communication, social, and stereotypical behaviors across ages in ADOS-G and the total, severity, social affect and repetitive ritualistic behaviors in ADOS-2 were not normally distributed. Figure 8A shows the frequency histograms taken across ABIDE I and II scores for each of the above-mentioned criteria of the ADOS-G, while Figures 8B and 8C break down the scores for modules 3 and 4 (note that ABIDE does not report module 1, and module 2 is sparsely used). Figure 9 shows the corresponding frequency histograms for ADOS-G. The empirical cumulative distribution functions (CDFs) for each criterion are shown in Figure 8D for ADOS-2 and Figure 9D for ADOS-G. They were tested separately because they are different ADOS versions, as were each subscore distribution. Notice here the similarities across CDFs for modules 3 and 4 and the similarity of each of these empirically estimated CDFs with the CDF corresponding to all scores. We note that pooling all scores is not valid, yet we do it here

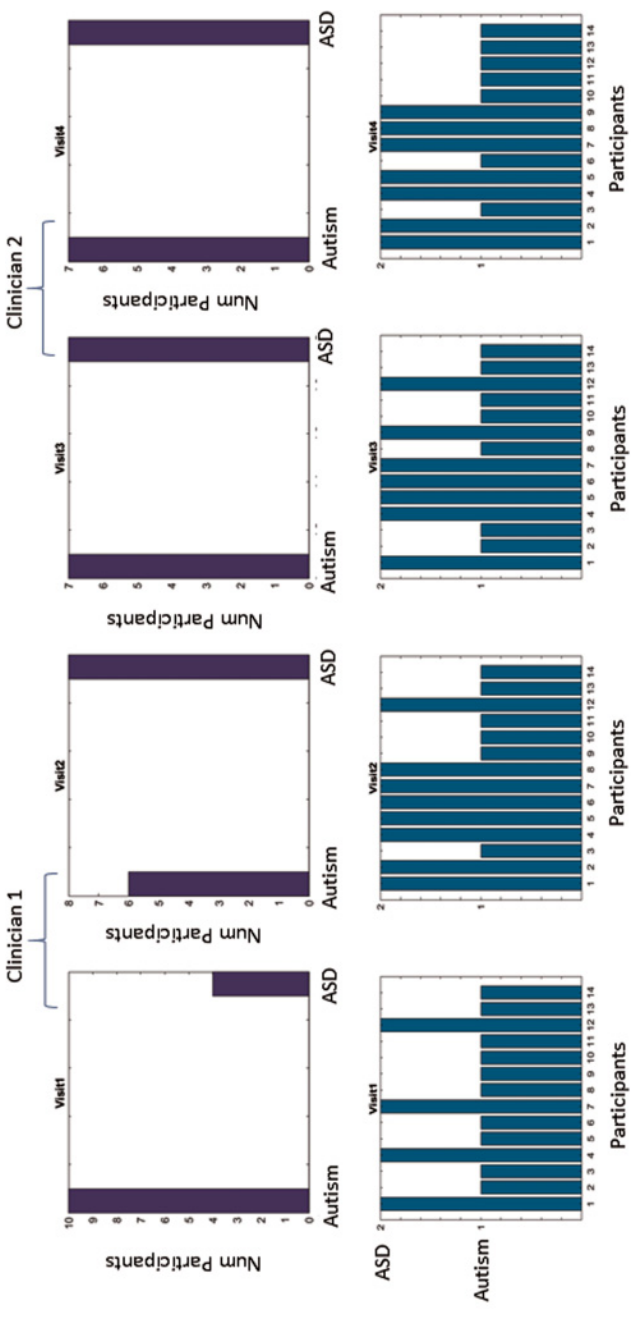


Figure 7: Nonreliable diagnostics for the same cohort of children under the same ADOS-2 modules and tasks. (A) Bias of rater clinician 1 versus rater clinician 2 reveals different styles bound to affect the outcome. (B) Variability of diagnostic classification (autism versus autism spectrum) for each of the 14 participating children in the four visits spanning 1.3 years on average $\pm$ 6 months.

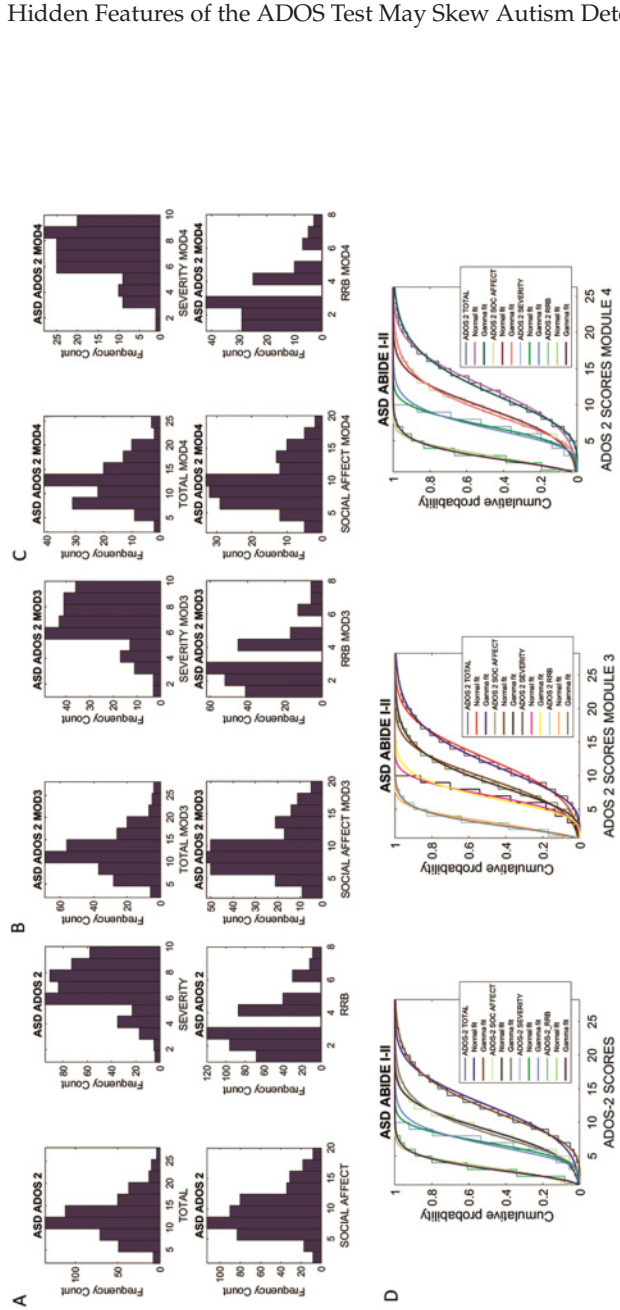


Figure 8: Nonnormality of the ADOS-2 total and each subscore distribution in the ABIDE I and II sites. (A) Empirically obtained frequency histograms for each of the ADOS-2 total and subscores from all ABIDE scores. (B) Same as in panel A but restricted to module 3 scores. (C) Same as in panel A but restricted to module 4 scores. (D) Empirically estimated CDFs for each subscore for the normal and gamma distributions' fit to the data. All cases failed the test of normality (see the text), but no statistically significant differences were found between the empirical CDFs generated by modules 3 and 4. No differences were found between each of the modules and the pooled data in panel A.

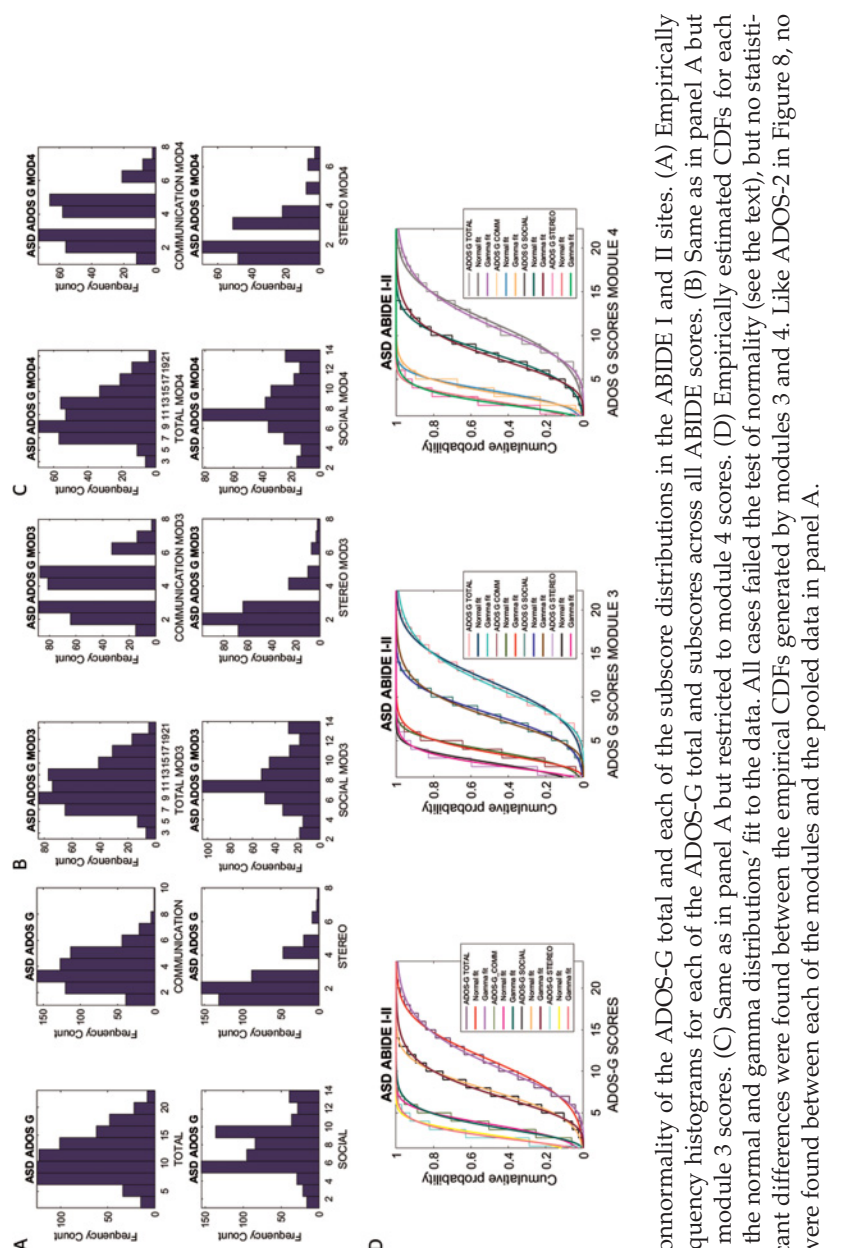


Figure 9: Nonnormality of the ADOS-G total and each of the subscore distributions in the ABIDE I and II sites. (A) Empirically obtained frequency histograms for each of the ADOS-G total and subscores across all ABIDE scores. (B) Same as in panel A but restricted to module 3 scores. (C) Same as in panel A but restricted to module 4 scores. (D) Empirically estimated CDFs for each subscore for the normal and gamma distributions' fit to the data. All cases failed the test of normality (see the text), but no statistically significant differences were found between the empirical CDFs generated by modules 3 and 4. Like ADOS-2 in Figure 8, no differences were found between each of the modules and the pooled data in panel A.

to show that the shape and dispersion of these histograms are quite similar despite the differences in module tasks.

The Lilliefors test failed the normality test with $p < 0.01$ for each subscore of each ADOS test version. More important, we used maximum likelihood estimation (MLE) and fit several distributions to assess the best fit with 95% confidence. The MLE revealed that the continuous gamma family of PDFs had a better fit than the normal PDF. The results from the normal and gamma fits are shown in Figures 8D and 9D for each of the corresponding subscores CDFs of the ADOS-2 and ADOS-G, respectively.

3.5 Females and Males Are Indistinguishable According to ADOS Scores. The ADOS score data from ABIDE were divided into those for the male and female participants to ascertain if (1) if the distributions of the total and subscores were symmetric (test for normality) and (2) they had statistically different overall scores comprising social, communication, and stereotypical repetitive motions that could distinguish the two phenotypes. The motivation for this comparison emerged from the ABIDE repository involuntary head motion data accompanying these ADOS scores (Caballero et al., 2018; Torres et al., 2017) and from distinction based on motor patterns derived from natural voluntary behaviors (Torres, Isenhower, et al., 2013; Torres, Nguyen, et al., 2016). These patterns can also automatically and blindly separate females/males with ASD from females/males with AS (from the subset of ABIDE with a DSM-IV classification) (Torres et al., 2017).

We reasoned that given that the ADOS tests are used to drive the science of autism (i.e., to correlate physiological data with it), it may be important to ascertain whether the variability in rater scoring from these versions of the ADOS matched the ability to distinguish the male from the female phenotype using involuntary motor noise. This is important because such motor noise is inherent in the autistic phenotype and provides visual feedback to the rater that could influence the rater's criterion currently missing the females in ADOS-driven detection.

Figures 10AB and 11AB show the distributions of scores for males (10A and 10B) and females (11A and 11B). In all cases, the normality test failed according to the Lilliefors test ($p < 0.001$). Recall that this test returns a decision for the null hypothesis that the data come from a distribution in the normal family, against the alternative that it does not come from such a distribution. In all cases, the test rejected the null hypothesis at the 5% significance level. As with the pooled data in Figures 7A and 8A and those from the breakdown into scores from modules 3 or 4 in Figures 7B and 7C and 8B and 8C, we used MLE to evaluate the fit of different distributions, which we show in the center panel of Figures 10A and 10B for the normal and the gamma family of distributions fit to the empirical CDFs of the total and subscores (the left side is ADOS-G and the right side ADOS-2).

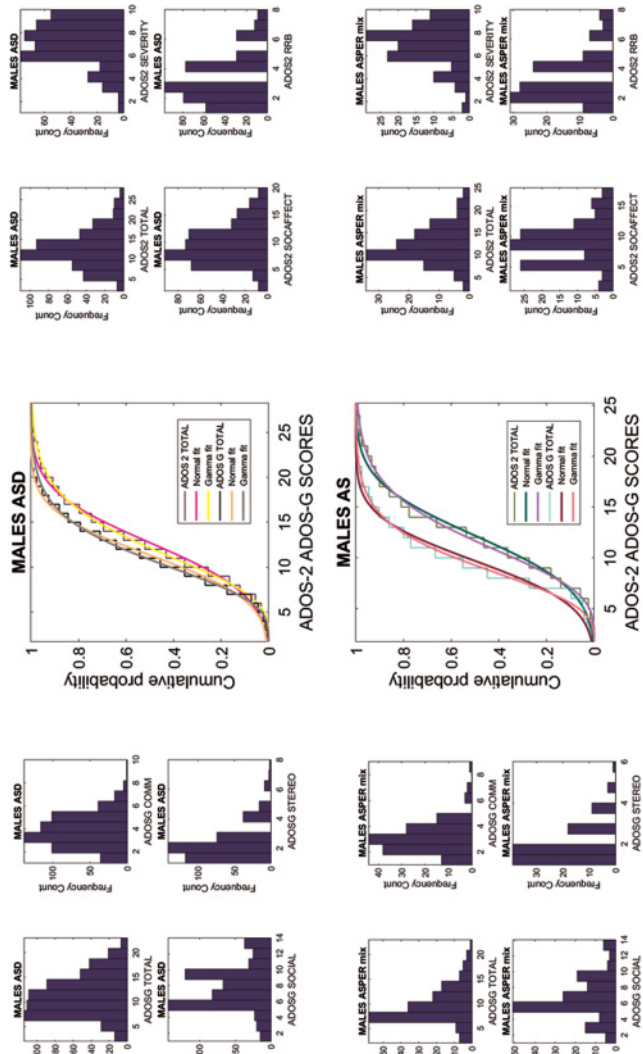


Figure 10: Nonnormality of the distributions spanned by the males with ASD and those with Asperger's syndrome according to the DSM-IV column of ABIDE. (A) The left panel shows the frequency histograms for each ADOS-G subscore; the central panel shows the empirical CDF's fit by the theoretical normal and gamma family CDFs for the total score; and the right panel shows similar plots for the ADOS-2. (B) Same format as in panel A for the case of Asperger's syndrome. Notice the separation of CDFs between ADOS-G and ADOS-2 contrasting with those of the ASD cases in panel A.

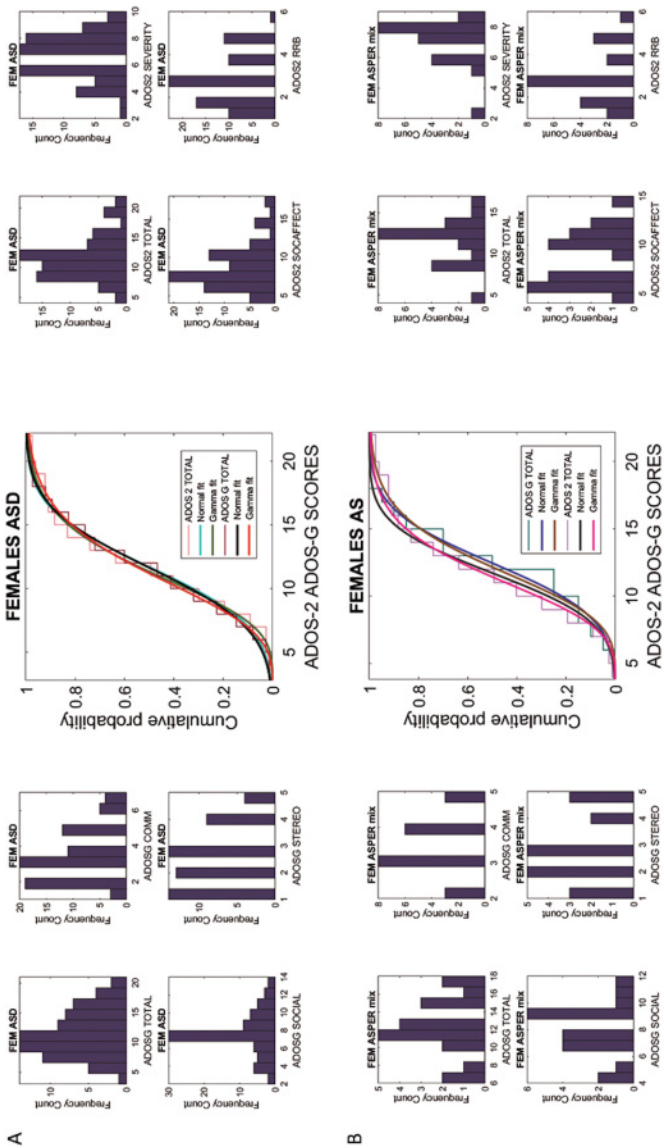


Figure 11: Nonnormality spanned by the distributions of females with ASD and AS in the ABIDE repository. (A) Frequency histograms of the ADOS-G total scores and subscores (left panel) and ADOS-2 (right panel), with the center panel showing the empirical CDF fit by the normal and gamma distributions for the ADOS-G and ADOS-2 scores. (B) Same as in panel A for females with a DSM-IV diagnosis of Asperger's syndrome.

Despite the well-established physiological and neurobiological distinctions between males and females in the spectrum of autism, we could not find any statistically significant difference between the ADOS-G (or the ADOS-2) total scores to automatically separate these two distinct phenotypes. We could not find either statistically significant differences between the subscores of each ADOS test when comparing males and females using the Wilcoxon rank-sum test, as all p-values were above 0.5. This is important, as motor noise (whether visible to the rater or occurring beneath the rater's awareness) would add to the observation criteria via visual feedback unless it is purposefully discarded as irrelevant to the diagnosis.

However, we did notice a significant difference in the distributions of the total scores for males when comparing those of the ADOS-G versus those of the ADOS-2 (see the center panel in Figure 10B). This difference was assessed using the Kolmogorov-Smirnov test for two empirical distributions and yielded a p-value of 0.0015. In contrast, this statistically significant difference was absent in the ADOS-G versus ADOS-2 comparison of total scores from the females' data with a p-value of 0.8 (see Figure 11A, center). Note here that these comparisons do not have meaningful clinical value. They are exclusively taken within a research framework to learn whether two versions of the same test always yield consistent separation or if in some cases they do not.

Given the visible separation of ADOS-G and ADOS-2 scores for the males, we proceeded to further interrogate the cohort of participants with AS according to the DSM-IV classification. This is possible as ABIDE provides the information on DSM-IV versus DSM-5 on two separate columns of the data matrix. We then asked if the males with AS were also separable from the males with ASD, according to the ADOS scores from the two versions of the same test. Notice here the relevance of this question, as the ADOS-G and ADOS-2 are indistinctly used in autism research (as instantiated by the ABIDE repository), and no differentiation is ever made by peer-reviewed papers that use one or the other to inform and guide the results from their physiology- and neurobiology-based research.

3.6 Incongruent Results between ADOS-G and ADOS-2 When Comparing Males with ASD and AS. We expected that despite their subtle differences, the variants of the same test would provide consistent results for participants with AS and for those with ASD. In the case of ADOS-G, we found a statistically significant difference between the scores of males with ASD and those with AS using the Wilcoxon rank-sum test, which yielded a p-value less than 0.001 (3.7×10^{-7}), significant at the 0.001 level. Further, the Kolmogorov-Smirnov test comparing two empirical distributions yielded statistical significance at the 0.001 level (p-value of 2.4×10^{-6}). These results rendered the two distributions for AS and ASD statistically different

at the 0.001 level (i.e., rejecting the null hypothesis that the two sets came from a similar distribution).

In contrast to the ADOS-G, the ADOS-2 total score comparison (i.e., within the same test) between the males with ASD and AS, using the Wilcoxon rank-sum test, yielded a p-value of 0.31, nonsignificant at the 0.05 level. The Kolmogorov-Smirnov test comparing two empirical distributions yielded a p-value of 0.5, with no significant difference and failing to reject the null hypothesis that the two data sets came from a similar distribution. Figure 10 shows the frequency histograms and CDF fits to the empirical data for males using the left-hand panel for the ADOS-G and the right-hand panel for the ADOS-2. Notice that the comparisons were made within each test, as they are different versions of the ADOS.

Comparisons of the females with ASD and AS are shown in Figure 11, using the total scores from the ADOS-G and ADOS-2, respectively. These outputs were congruent for the two versions of the test in that neither ADOS version distinguished the two groups. The Wilcoxon rank-sum test yielded a p-value of 0.30 for ADOS-G and 0.54 for ADOS-2. The Kolmogorov-Smirnov test comparing two empirical distributions yielded a p-value of 0.42 for ADOS-G and 0.96 for ADOS-2. Notice that in both males and females, breaking the data into module 3 or 4 scores made no difference. As in the case of Figures 8 and 9, the distributions retained the shape and dispersion across the modules, suggesting that blind classification of participants is not possible using these tests. In other words, given the scores of a module 3 or a module 4 version containing different tasks, it is not possible to know if they came from module 3 or 4, as they span identical distributions. The numbers are statistically indistinguishable, so any machine learning algorithm attempting to classify participants based on these scores would fail, despite their coming from entirely different sets of tasks and being aimed at assessing individuals with disparate language or communication levels.

We note that given the differences in sample size and the nonnormality in the distributions of scores, we used the Wilcoxon rank-sum test in all the above comparisons. This is a nonparametric test for two populations, used when samples are independent and of different sizes. It is equivalent to the nonparametric Mann-Whitney U-test for equality of population medians. The test statistic that rank-sum returns is the rank sum of the first sample (Hollander & Pena, 2004).

4 Discussion

In this work, we studied the statistical properties of the scores generated by the ADOS test in two different contexts. In one study, we assessed the scoring of research-reliable testers in a laboratory environment, using the ADOS as an experimental protocol. In the other study, we examined the ADOS scores reported by research-reliable raters in the open access ABIDE

repository (Di Martino et al., 2014), including 1324 clinical records of participants with ASD.

We examined the extent to which the rater's style and inherent biases could change the autism-versus-ASD diagnosis for a given child under similar laboratory conditions. We further asked in this laboratory context which of the scores would be the most affected by the change in rater. We found significant differences in diagnoses that depended on the rater. We also found that the RRB scores were the most susceptible to the nuances of the rater's style and inherent biases. However, given the small sample set of ADOS scores for the study in the lab, we decided to examine the statistical features of the ADOS scores in a much larger data set. To that end, we used the open-access ABIDE repository and interrogated the reported scores of ADOS-G versus ADOS-2 in several cohorts. These included scores from older studies that reported scores based on the DSM-IV criteria separating AS and ASD. We also took advantage of the rare opportunity that ABIDE offers with respect to females with AS and ASD. This repository has large numbers of females affected by these conditions. These are difficult to find in any random draw of the population, given the disparate ratio of approximately five males for every female in the spectrum. Given the comparable numbers in ABIDE, we could contrast the statistical features of their scores to those of corresponding males.

We found that the distributions of scores from the ADOS tests do not follow the a priori assumption of normality that researchers who are adopters of the test often make. Given these findings, it may be important to reconsider adopting this test to inform basic scientific research in neurodevelopment. Across a multitude of research papers, discrete scores that do not have a properly defined norm are systematically forced to be (linearly) correlated with continuous physical data. Yet the lack of normality in the distributions of the scores, along with the lack of independence between raters' bias and sensitivity, pose a problem for research validity, according to SDT and ROC area-under-the-curve types of analyses (Somoza & Mossman, 1991). One now wonders how many false positives we may have in research studies.

The sensitivity and reliability tests that researchers adopting this clinical scoring system assume to assess its validity are built under specific statistical assumptions of normality (Jang, Wixted, & Huber, 2009; Kroll, Yonelinas, Dobbins, & Frederick, 2002; White & Wixted, 2010; Witt et al., 2015). Under such assumptions, the shape and dispersion of the variability from the scores' probability distributions are critical to maintain independence between the inherent bias of the observer and the sensitivity of the test.

What does it really mean to have autism or to be on the autism spectrum? And how can that distinction be made relative to normative data from typically developing controls when no such data exist? Here we see, even in a rather modest cohort of neurotypical participants, a significant spread of scores for typical neurodevelopment, thus indicating the presence of

behavioral symptoms in the neurotypical population. Given such fluctuations, how can we build a proper metric for neurodevelopmental research?

The ADOS range of scores is based on positive integer values, with 0-value at the lower bound, that is, the behavior is not present as specified. Behaviors coded on the ADOS are assumed to be those that occur in nonspectrum individuals (e.g., eye contact, pointing, shared enjoyment), as well as behaviors that could occur in ASD (e.g., stereotyped or idiosyncratic language, complex mannerisms). We note that those motions present in nonautistic individuals have not been experimentally assessed under ADOS conditions, as noted in the classical paper by Lord et al. (2000) in the last sentence of the paper. However, motions inherently present in natural behaviors that scaffold social interactions, such as eye contact, pointing movements, and face-processing micromotions, are routinely studied in the neuromotor control developmental literature. These studies use objective means that characterize several types of motions (Klin, 2000; Klin et al., 2002a, 2002b; Klin, Lin, Gorrindo, Ramsay, & Jones, 2009; Torres, Brincker, et al., 2013; Torres, Isenhower, et al., 2013; Torres, Yanovich, & Metaxas, 2013), and would be amenable to combine with validated clinical scores of subjective observation-based inventories. Combining such complementary data would bring a level of precision that we now lack in behavioral analyses that is purely observational and does not reflect the neurological aspects of this condition.

As demonstrated by our analyses, the human eye has a limited capacity to detect motor noise at various layers of motor control inherently present in social interactions, communicative language, and repetitive ritualistic motions of the kinds that the ADOS assesses. Relying on such data exclusively poses a challenge to basic research when guiding and informing our scientific quest in the search for adequate medical treatments. Clinical tools in other fields are now routinely combined with digital data and commercially available biosensors to help patients and clinicians obtain feedback along more than one data channel, beyond the limits of the naked eye. In neurodevelopment, and particularly in disorders of the nervous systems that go on to receive a diagnosis of autism today, it may be important to combine multiple sources of the various layers of the knowledge network (as in Figure 2A) to inform personalized approaches of new, more effective ways to develop target therapies. Along those lines, the lack of normative data seems to be an obstacle to the precision medicine paradigm, which could benefit from stratifying subtypes in heterogeneous presentations under an umbrella term like *autism*.

Other fields could help autism researchers establish a proper metric to quantify the type of accelerated rates of change that one sees in early neurodevelopment by characterizing normative states and departure from them at expected neurodevelopmental milestones. For example, in pediatrics, the growth charts from the Centers for Disease Control and the World Health Organization serve the purpose of establishing normative criteria to

measure departure from typical development as the child physically grows day by day. In autism, there is nothing comparable to the growth charts, so there are no metrics of similarity to detect and track change and its rate. Indeed, it has now been established that observational behavioral criteria grounded on psychological, behaviorally defined (subjective) constructs drive the scientific quest in autism and supersede physiological (objective) criteria (Torres & Whyatt, 2018; Whyatt & Torres, 2018). The time is ripe to build a new standard similarity metric that combines validated clinical scores with the type of motor noise data that we can easily harness today with wearables from the fast-growing and rapidly developing nervous systems of young infants and young children.

One challenge ahead to combine digital and clinical data, as other fields have done for personalized assessments under the tenets of precision medicine, is the absence of a proper age-dependent statistical framework to measure relative changes away from typical levels. In this sense, although the ADOS modules were designed to account for possible disparities in cognitive and verbal capacity, there are no age-dependent physical criteria in the research-adopted version to ascertain the physical rate of change in nervous systems that develop asynchronously across the population. In autism two children of the same physical age may have entirely different profiles of motor noise in the different levels of neuromotor control that contribute to the emergence of autonomous social interactions and communicative language. Since behaviors are made up of motions at the observable macro- and the hidden microlevels, we need new metrics that consider the interactions between these disparate levels of inquiry combining different time, spatial, and frequency scales. We also need methods that reveal inherent capacity for entrainment and synchronous dyadic exchange during social interactions.

Many of the children with autism are capable of social exchange when properly supported. Yet, the support depends on our knowledge of what the coping nervous systems of the child can already do at the voluntary, spontaneous, automatic, involuntary, reflexive, and autonomic levels of function. If we had a better sense of the inherent capacity for social exchange, we could work with that child rather than inavertedly stress the nervous systems with more uncertainty—for example, by prompting the child to perform expected (socially appropriate) behaviors. In the absence of a test that can reveal these features through proper metric scales, research to develop such accommodations may be impeded. The lack of normality accompanied by the lack of methodology to properly reveal change and its rate in a rapidly developing system poses a challenge for researchers who aim at developing personalized targeted therapies for stratified autism subtypes. It is difficult to reveal the target for medical treatment in the face of one-size-fits-all methods that assume a single statistical distribution across the population. This assumption also bears the potential for false positives and lower reliability than has been assumed up to now.

Even the so-called standardized ADOS severity scores do not address these needs; they were developed for criteria-referenced clinical use, so the scale was not built using typical controls as a norm-referenced test would be (Gotham, Pickles, & Lord, 2009; Hus, Gotham, & Lord, 2014; Hus & Lord, 2014). The score is designed to compare an individual with ASD to other individuals with ASD of the same age and language level. It also has a range of ages 6 to 10 for individuals with ASD and is not meant to represent ASD on a range of 1 to 10. Autism is not only highly heterogeneous. It also has neurodevelopmental asynchronies in a group of the same age, meaning that two individuals may be 10 years old but one may have the signatures of neuromotor control from normative 3-year-old children (Torres, Brincker, et al., 2013). See Figure 14 in the appendix. Thus, aging with autism is different from typically aging. What is the normative range of scores that reflect age-appropriate typical social interactions?

Because of the prevalent influence that the research-grade ADOS test has on basic science at all levels, it is imperative to reexamine the inherent theoretical assumptions that adopters of this test have made and verify that the outcome of this test, as administered in research settings, empirically matches the theoretical assumptions of the users.

Data-driven approaches tend to preserve empirically assessed features of phenomena. They do not throw away important variability and offer the possibility of capturing the true nature of change in a coping neurobiological system that is developing at atypical rates. Although this article uses the ADOS test as the example to illustrate the potential problems that blindly adopting such tests for scientific research may create, the same tenets apply to any other clinical test used in basic research of neurodevelopmental disorders. These disorders reflect in great part problems with the nervous systems, and since nervous systems are adaptable we would be missing self-correcting mechanisms by imposing theoretical models without empirically informing those models.

Open access repositories now make it possible to examine (for the first time by researchers from different fields with different skillsets) the validity of the use of this instrument in research drawing on the large number of records now available. These provide enough statistical power, where the high cost of running these studies often prevents labs from deploying them. For example, recent work using the National Database for Autism Research (NDAR) and the Simons Simplex collection demonstrated that it is possible to shorten the administration time of the ADOS while preserving cutoff criteria. While this earlier work already highlighted the nonnormality of the distributions of the scores reported in those data repositories (Duda, Kosmicki, & Wall, 2014), the work presented here takes a deeper look at the assumptions that the research-grade test makes. Here, we further raise several relevant questions about the need for normative assessment across the human life span to truly formulate a standard test. Such a test would have an age-appropriate metric for research use in neurodevelopment, that is, to

enable derivation of developmental trajectories across different parameters scaffolding the emergence of social behaviors.

4.1 Positive Aspects of the ADOS Adoption in Clinical Settings. Perhaps the use of the ADOS in the clinical arena is less damning than the use of its research-grade version in basic scientific research. In the United States, the autism label ensures coverage for certain treatments that many children could not otherwise have access to, particularly in underrepresented minorities and poor rural areas. Although ADOS testing does not produce an official diagnosis, its use in research labs helps refine the criteria from the DSM-5 and lends more specificity to the symptoms than other criteria. In the United States, this research-grade version could serve as a flag to send parents to federally certified clinics that offer services upon multiprone criteria involving other tests. However, owing to copyright issues, Western Psychological Services (WPS) does not allow researchers to copy, reproduce, or share with the families the ADOS booklet with important details of the outcome. In other words, the children who come to our labs, receive the research-grade ADOS, and pass the cutoff scores are labeled autistic by this test. However, as researchers, we are not allowed to share details with their parents. This obstructs their ability to go to a proper clinic, show these research-grade results, and pursue the diagnosis that will give them access to early intervention programs or individualized education programs for children of school age. If the WPS and the trainers of the ADOS allowed this, the test adopted by researchers would serve as a warning to parents that some aspects of the child's neurodevelopment may be offtrack and accelerate the official diagnosis. In our own experience, in poor regions of the United States, it can take two years before an official diagnosis grants services to families upon suspicion that the infant is offtrack.

Unlike the statistical confounds that the ADOS total scores and subscores surely bring to research, in its current form at the clinic, the ADOS provides psychological comfort to adults who had never been previously diagnosed and could not understand their place in the social scene. Many adults, newly diagnosed at the clinic, express a sense of relief on learning that they are on the autism spectrum and have social interaction differences. Further, the ADOS adds important information to the coarser DSM diagnosis. Thus, the clinical value of this instrument is highly appreciated. When used in laboratory settings as an experimental protocol to study social interactions, the structured ADOS inventory can facilitate the development of new models to study the neuromotor control underpinnings of social exchange. This combination of the clinical protocol and the digital data from wearables has offered new ways to assess complex dyadic behaviors, amenable to extend to real situations. However, continued reliance on its coarse scores while plagued with false positives will prevent us from automatically (in the blind) stratifying the different forms of autism and opening

new avenues to researchers in genetics and other important emerging fields of metabolomics and the human microbiome.

4.2 Somatic Sensory Motor Issues in Neurodevelopmental Disorders on a Spectrum. There is room for transformative change to improve the use of such clinical tests in scientific research studies in general. After years without recognizing the importance of sensory motor issues in mental health (Bernard & Mittal, 2015), the Research Domain Criteria (RDoC) matrix created by the National Institute of Mental Health (Cuthbert & Insel, 2013; Insel, 2014) finally included in January 2019 the entry for sensorimotor issues. This new development, paired with the admission by the DSM-5 that there are sensory issues in autism, will provide a new foundation to explore human social behaviors from a new angle that can include movements and their kinesthetic sensation (Torres & Whyatt, 2018), while offering the opportunity to uncover the inherent capabilities and predispositions that a coping nervous system develops (Brincker & Torres, 2013).

The ADOS test nevertheless still fails to recognize sensorimotor issues (Lord et al., 2000), as we quote the following caveat when choosing a module from the manual:

> Note that the ADOS-2 was developed for and standardized using populations of children and adults *without significant sensory and motor impairments* [emphasis added]. Standardized use of any ADOS-2 module presumes that the individual can walk independently and is free of visual or hearing impairments that could potentially interfere with use of the materials or participation in specific tasks. (Catherine Lord, Rutter, DiLavore, Risi, & Western Psychological Services Firm.)[2]

Interestingly, despite this caveat for the use of the ADOS test stated in the manual, to the best of our knowledge, the makers of the ADOS test have never reported scientific studies of individuals with autism where it is objectively established that individuals in the spectrum of autism have no significant sensory and motor impairments. Yet when we test the children in basic scientific research labs and use high-grade instruments, we invariably find visual, hearing, olfactory, and touch impairments that would surely interfere with the use of the materials in this test. More important yet, such motor noise, involuntary micromotions, dysautonomia and other sensorimotor differences contribute to the disparity in scoring styles because inevitably, not all raters perceive these problems through a unifying lens. Some let it enter into their criteria for scoring, while others filter them out altogether. Figure 7 provides a clear example of it, and so do the inconsistencies with the RRB scores.

These highly quantifiable problems with their somatic and sensorimotor systems are the tip of the iceberg, as deeper problems are present with their

[2] https://www.wpspublish.com/ados-autism-diagnostic-observation-schedule

enteric nervous systems (the gut) and microbiome (Fattorusso, Di Genova, Dell'Isola, Mencaroni, & Esposito, 2019; Heiss & Olofsson, 2019; Hughes, Rose, & Ashwood, 2018; Ng et al., 2019; Pulikkan, Mazumder, & Grace, 2019). Many suffer from pain and temperature dysregulation as their overall sense of touch, vestibular issues with balance, and multi-sensory integration overwhelm them in ways that we can now precisely quantify in a personalized manner. All of these issues deeply affect the ability to develop social interactions. Perhaps the new NIH-RDoC sensorimotor criteria will help WPS redefine ADOS for research and encourage the use of new objective criteria grounded on biophysical metrics assessing the nervous systems' functions.

There is mounting evidence that somatic-sensory-motor issues do exist across the many phenotypes that go on to receive this diagnosis today under the DSM-5 broader criteria community (Behere, Shahani, Noggle, & Dean, 2012; Campione, Piazza, Villa, & Molteni, 2016; Donnellan & Leary, 1995; Eigsti, Rosset, Col Cozzari, da Fonseca, & Deruelle, 2015; Hannant, Tavassoli, & Cassidy, 2016; Jasmin et al., 2009; Kushki, Chau, & Anagnostou, 2011; Mandelbaum et al., 2006; Minshew, Sung, Jones, & Furman, 2004; Mosconi et al., 2013; Mosconi, Mohanty, et al., 2015; Mosconi & Sweeney, 2015; Mosconi, Wang, Schmitt, Tsai, & Sweeney, 2015; Ornitz, 1974; Perry, Minassian, Lopez, Maron, & Lincoln, 2007; Siaperas et al., 2012; Torres, Brincker, et al., 2013; Torres, Isenhower, et al., 2016; Torres, Yanovich, et al., 2013; Troyb et al., 2016; Whyatt & Craig, 2013) and many more.

The neurological differences in autism have been well documented since as far back as the 1970s and solidified as a model by the 1980s (Maurer & Damasio, 1979, 1982; Vilensky, Damasio, & Maurer, 1981). A neurological model was championed by Damasio and Maurer (1978) and developed in a model of support and accommodations (Donnellan & Leary, 1995). Behavioral criteria for detection and treatments took over the field of autism research, and this neurological model was abandoned in favor of behavioral modification methods like those developed by Skinner in the 1950s. Those models from the behaviorist school of thought were developed to condition animals in the laboratory. But the methods to condition lab animals were transferred to human children by Lovaas, Schreibman, and Koegel (1974). The Lovaas methods, along with behaviorally based detection criteria from the same school of thought, remain the gold standard of diagnosis and treatments of autism today. These are not subject to the level of accountability that scientific research is. That is, they do not require any type of institutional review board and do not have to be compliant with the Helsinki Act. This is indeed highly puzzling given their pervasive presence in science.

Today, the DSM-5 allows ADHD and sensory issues in the criteria for autism, so autism is no longer a narrowly, well-defined behavioral disorder (perhaps it never was) despite insistence on defining it by criteria that exclude the underlying physiological conditions that these individuals develop from birth onward. It is a view that strictly separates the description

of behavior from the physiological underpinnings—as if one could produce behaviors without a nervous system. An emergent field aimed at uncovering multiple digital biomarkers to characterize and automatically stratify various aspects of development for research purposes, may be the answer to the start of a new era in scientific research on autism aimed at a physiological characterization of the condition for medical use.

In the United States, the field may not have a choice at this point, given the current demands by medical insurance providers on a higher standard for treatments to conform to the American Medical Association criteria. Why we waited so long and affected generations to adopt a medical criterion is puzzling, when a proper neurological model existed in the 1970s. Even more puzzling, we see that support systems to accommodate the neurological issues were already being successfully adopted back then by affected families. Yet the evolution of those who are adults today and received intense behavioral modification (referred to as conversion therapy in some circles) has shown us the gross error made several decades ago. It is up to a new generation of scientists now to correct that error, but it will not be without resistance from organizations and a powerful lobby advocating a purely behaviorist model of detection and treatment recommendations that exclude physiology. Tricare, one of the major insurance providers in the United States, reported $261.9 million for 13,390 beneficiaries of applied behavioral analyses (ABA) in contrast to $38.2 million for occupational, physical, and speech therapy combined. Prescription medication expenditures were marked at $15.1 million. This Tricare report pertained to families in the military, subject to lower rates in comparison to those by other providers. The high cost preventing broad access to treatment paired with their poor and uncertain medical outcome has raised concerns in the autism CARES report mentioned in section 1 (publicly available online). Because of the lack of autonomous living and agency in the aging autistic adults who received such treatments intensively since early life, a new medical model is imperative.

In our study, it was evident that whether using absolute scores or derivative, age-dependent data accounting for longitudinal dynamic changes from visit to visit, the RRB scores of the ADOS implicitly reflecting sensory motor issues picked up best the switching of the clinician. If we were to combine this structured social test with wearable biosensors, we could automatically stratify autism spectrum disorders, unveil different subtypes of idiopathic autism, and map them to known genetic phenotypes. Such clinically informed digital data could provide objective criteria useful to the community doing basic scientific work (e.g., geneticists, electrophysiologists, neuroimaging) blind to biases and false positives that inadequate statistics introduce under existing approaches. The label of autism marginalizing the affected person within society would no longer be necessary, and treatments and services would be driven and covered by physiologically informed medical criteria. New hybrid methods could also help

physicians treat the medical issues in a more personalized manner. Some collaborative work along those lines has been done between clinicians and researchers, but more research is needed to fully validate and replicate the use of digital ADOS within the smart-mobile and personalized health concepts.

5 Conclusion

We invite reader to consider that science in autism needs to retake the path of independence and reclaim its agency to be able to conduct proper scientific research away from profitable models. This can be done by building an interdisciplinary consortium of scientists from diverse disciplines with complementary skill sets to derive proper metrics and true standardized methods for personalized medicine. Combining open access data sets and interdisciplinary collaborations can lead to empirical verification of our theoretical models and alert us when they fall short of enforced expectations.

Figure Appendix

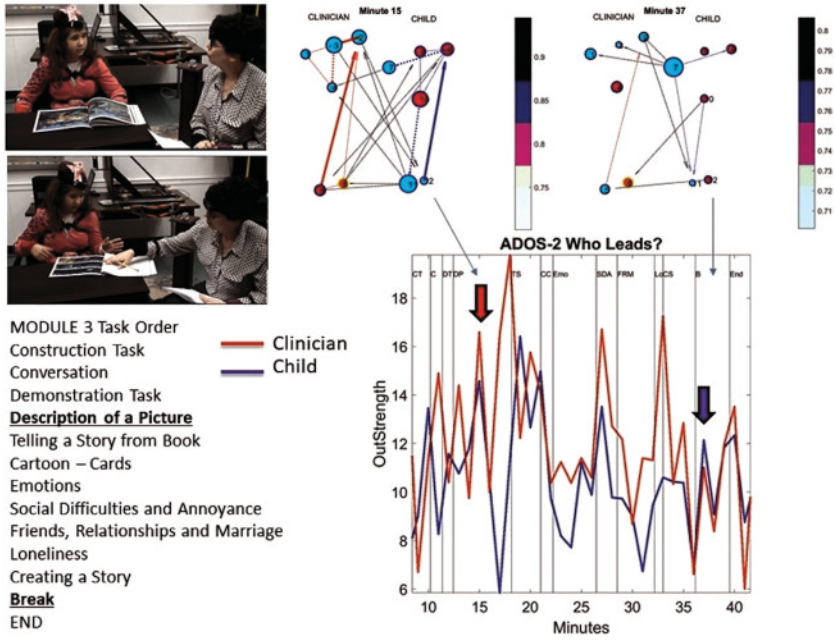


Figure 12: The use of the ADOS test as an experimental protocol to digitize the dyadic interaction that takes place between the child and the clinician (Whyatt & Torres, 2017). Using weighted, directed graphs and network connectivity analyses derived from motor output fluctuations registered by wearables, we can automatically determine who leads and lags in each task and which tasks are socially favorable to the child's physical abilities to entrain with the tester and express joint attention and other forms of synchronization and social cohesiveness spontaneously emerging during the tasks. Nodes of the network represent body parts (wrists, thorax, lumbar, and ankles of child and clinician) with the edge color denoting cross-coherence levels (taken minute by minute at 128 Hz) and represented in the color bar. The size of the nodes represents incoming links, while color gives self-emerging modules (subnetworks with maximal inner connectivity and minimal outer connectivity). Arrows denote out-degree level, and the thickness of the arrow denotes node i leading node j by a positive phase shift (from cross-coherence spectral analyses.) We highlight two tasks—one in which the clinician leads and one in which the child leads.

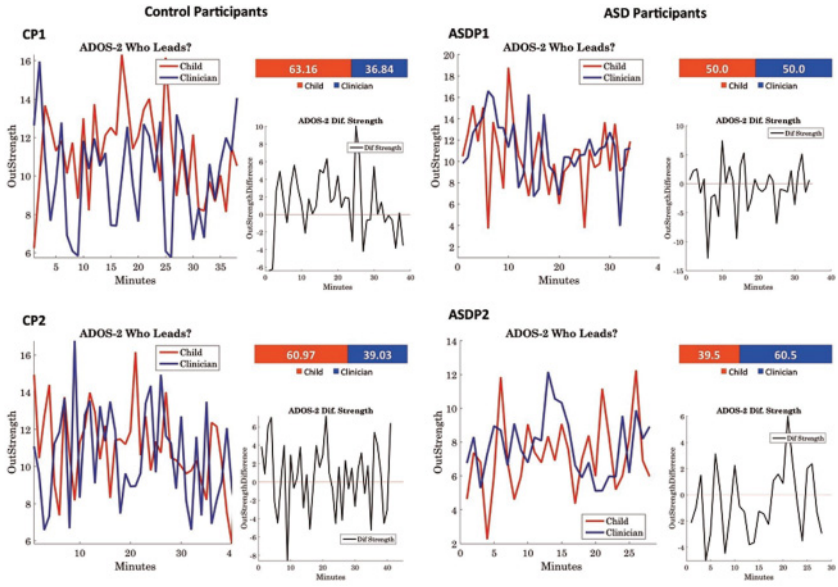
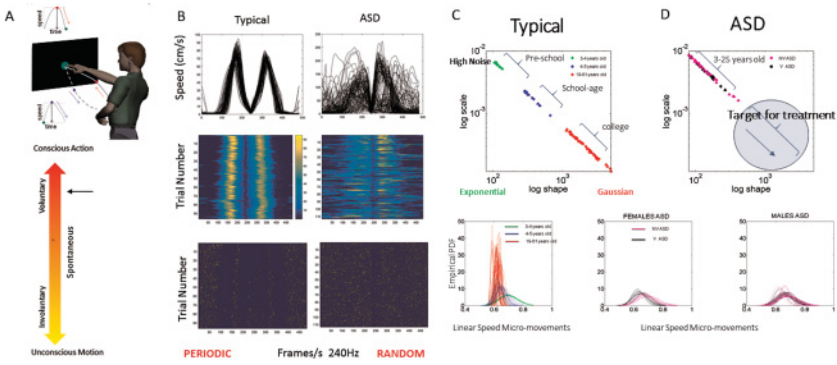


Figure 13: Sample children (two TD and two ASD) with full ADOS-2 tasks and lead-lag profile capturing the child-clinician interactions during the test. Metrics were derived from weighted directed graphs with outstrength measure at each node tallied to quantify which body parts lead the interaction and who leads the social exchange overall, minute by minute. Notice that TD children lead most of the time, while ASD children tend to lag in the interactions. The weighted, directed graphs were derived from cross-coherence analyses of synchronously coregistered motions from inertial measurement units (IMUs) and gyroscopes in a grid of wireless wearable sensors across the body (Whyatt & Torres, 2017).



Author Contributions

Conceptualization, E.B.T.; methodology, E.B.T.; software, E.B.T., S.M.; validation, E.B.T., R.R., S.M. and B.G.; formal analysis, E.B.Y.; investigation,

Figure 14: Quantifying noise levels in the self-generated kinesthetic reafferent feedback derived from motor output variability of goal-directed visuo-motor actions. (A) Pointing task to assess signatures of motor noise across neurodevelopment. The arrow indicates the proposed taxonomy of neuromotor control across different levels of function in the nervous systems. (B) Typical development (TD) signatures of voluntary (instructed) and spontaneous (uninstructed) hand-retracting movements during pointing. Speed profiles are derived from hand-pointing trajectories in panel A forward to the target (first bell-shaped) and away from the target toward a resting position during decision making in a cognitive match-to-sample task. The resting hand raises speed continuously until it reaches a peak, then decelerates to touch the target on a touch monitor that registers the touch. The hand leaves the monitor and accelerates away from the target, toward the body, to reach the resting state again. Midway along the retracting hand trajectory, the hand reaches its second peak and then decelerates again to rest. Trials are recorded at the child's own pace. The target appears in trial 1 and disappears once the child touches it, then appears again, implicitly instructing the child to touch it again. There are 100 trials presented as a heat map with global peaks highlighted in yellow. They are stacked in the order of occurrence and aligned to the touch. The typical behavior of the TD children is highly periodic and well structured. The third plot at the bottom shows the small (local) speed peaks of the resting state, followed by the first ballistic phase of the motion with well-aligned peaks of the maximal speed of the hand on its way to touch the target, then the ballistic phase toward the target, followed by small speed peaks while resting briefly at the target-stratifying ASD subtypes in Wu, Jose, Nurnberger, and Torres (2018). The ballistic phase returning the hand toward the body follows (with no peaks); then the peaks of the speed maxima automatically align again from trial to trial. Another ballistic phase retracting the hand ensues, and the hand lands to a resting state, with the presence once again of the small speed peaks at rest. (B) The participant with ASD (similar age and sex to TD child) depicts very different patterns, consisting of highly disorganized motions, with (involuntary, unintended) random noise output by his system while performing these goal-directed reaches. Notice the absence of a pattern in the forward phase of the pointing to the target and the emergence of some structure in the return motions. Further, notice the presence of random noise (also modeled and empirically characterized in Wu et al., 2018, using a Poisson process. (C) The micromovement spikes reveal states of maturation in TD from 3 years of age to college age, in contrast to its absence in ASD 3 to 25 years old (Torres, Brincker, et al., 2013), suggesting the importance of tracking the shifts in probability space with age and treating autism as a lifelong condition. This contrasts with assuming a theoretical distribution under the one-size-fits-all model of autism ADOS-driven research today.

E.B.T., R.R., S.M., B.G.; resources, E.B.T.; data curation, E.B.T., R.R., S.M., B.G.; writing—original draft preparation, E.B.T.; writing—review and editing, R.R., S.M., B.G.; visualization, E.B.T.; supervision, E.B.T.; project administration, E.B.T.; funding acquisition, E.B.T.

Acknowledgments

We thank the children and families who participated in this study. We thank the anonymous clinicians who helped us with administering the ADOS, scoring it, and reliability assessment. We thank the past and current members of the Rutgers Sensory Motor Integration Lab who helped with data collection and study coordination. We thank the Computational Neurobiology Lab of the Salk Institute of Biological Studies for hosting E.B.T. during her sabbatical, while writing this article. This research was funded by the New Jersey Governor's Council for the Medical Research and Treatments of Autism, grant number CAUT14APL018 and by the Nancy Lurie Marks Family Foundation.

Conflicts of Interest

We declare no conflict of interest. The funders had no role in the design of the study; the collection, analyses, or interpretation of data; the writing of the manuscript; or the decision to publish the results.

References

American Psychological Association. (2013). *Diagnostic and statistical manual of mental disorders: Fifth edition.* Arlington, VA: APA.

Behere, A., Shahani, L., Noggle, C. A., & Dean, R. (2012). Motor functioning in autistic spectrum disorders: A preliminary analysis. *J. Neuropsychiatry Clin. Neurosci.*, *24*(1), 87–94. doi:10.1176/appi.neuropsych.11050105

Bernard, J. A., & Mittal, V. A. (2015). Updating the research domain criteria: The utility of a motor dimension. *Psychol. Med.*, *45*(13), 2685–2689. doi:10.1017/S0033291715000872

Brincker, M., & Torres, E. B. (2013). Noise from the periphery in autism. *Front. Integr. Neurosci.*, *7*, 34. doi:10.3389/fnint.2013.00034

Caballero, C., Mistry, S., Vero, J., & Torres, E. B. (2018). Characterization of noise signatures of involuntary head motion in the Autism Brain Imaging Data Exchange Repository. *Front. Integr. Neurosci.*, *12*, 7. doi:10.3389/fnint.2018.00007

Campione, G. C., Piazza, C., Villa, L., & Molteni, M. (2016). Three-dimensional kinematic analysis of prehension movements in young children with autism spectrum disorder: New insights on motor impairment. *J. Autism. Dev. Disord.*, *46*(6), 1985–1999. doi:10.1007/s10803-016-2732-6

Constantino, J. N., & Charman, T. (2016). Diagnosis of autism spectrum disorder: Reconciling the syndrome, its diverse origins, and variation in expression. *Lancet Neurol.*, *15*(3), 279–291. doi:10.1016/S1474-4422(15)00151-9

Constantino, J. N., Kennon-McGill, S., Weichselbaum, C., Marrus, N., Haider, A., Glowinski, A. L., . . . Jones, W. (2017). Infant viewing of social scenes is under genetic control and is atypical in autism. *Nature*, *547*(7663), 340–344. doi:10.1038/nature22999

Cuthbert, B. N., & Insel, T. R. (2013). Toward the future of psychiatric diagnosis: The seven pillars of RDoC. *BMC Med.*, *11*, 126. doi:10.1186/1741-7015-11-126

Damasio, A. R., & Maurer, R. G. (1978). A neurological model for childhood autism. *Arch. Neurol.*, *35*(12), 777–786. doi:10.1001/archneur.1978.00500360001001

Di Martino, A., Yan, C. G., Li, Q., Denio, E., Castellanos, F. X., Alaerts, K., . . . Milham, M. P. (2014). The Autism Brain Imaging Data Exchange: Towards a large-scale evaluation of the intrinsic brain architecture in autism. *Mol. Psychiatry.*, *19*(6), 659–667. doi:10.1038/mp.2013.78

Donnellan, A. M., & Leary, M. R. (1995). *Movement Differences and Diversity in Autism/Mental Retardation*. Madison, WI: DRI Press.

Duda, M., Kosmicki, J. A., & Wall, D. P. (2014). Testing the accuracy of an observation-based classifier for rapid detection of autism risk. *Transl. Psychiatry.*, *4*, e424. doi:10.1038/tp.2014.65

Eigsti, I. M., Rosset, D., Col Cozzari, G., da Fonseca, D., & Deruelle, C. (2015). Effects of motor action on affective preferences in autism spectrum disorders: Different influences of embodiment. *Dev. Sci.*, *18*(6), 1044–1053. doi:10.1111/desc.12278

Fattorusso, A., Di Genova, L., Dell'Isola, G. B., Mencaroni, E., & Esposito, S. (2019). Autism spectrum disorders and the gut microbiota. *Nutrients*, *11*(3). doi:10.3390/nu11030521

Freedman, D., & Diaconis, P. (1981). On the histogram as a density estimator: L2 theory. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete*, *57*(4), 453–476.

Friston, K. J., Redish, A. D., & Gordon, J. A. (2017). Computational nosology and precision psychiatry. *Comput. Psychiatr., 1*, 2–23. doi:10.1162/CPSY_a_00001

Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J. Autism. Dev. Disord.*, *39*(5), 693–705. doi:10.1007/s10803-008-0674-3

Hannant, P., Tavassoli, T., & Cassidy, S. (2016). The role of sensorimotor difficulties in autism spectrum conditions. *Front. Neurol.*, *7*, 124. doi:10.3389/fneur.2016.00124

Havdahl, K. A., Bal, V. H., Huerta, M., Pickles, A., Oyen, A. S., Stoltenberg, C., . . . Bishop, S. L. (2017). Dr. Havdahl et al. reply. *J. Am. Acad. Child Adolesc. Psychiatry*, *56*(7), 619–620. doi:10.1016/j.jaac.2017.05.010

Havdahl, K. A., Hus Bal, V., Huerta, M., Pickles, A., Oyen, A. S., Stoltenberg, C., . . . Bishop, S. L. (2016). Multidimensional influences on autism symptom measures: Implications for use in etiological research. *J. Am. Acad. Child. Adolesc. Psychiatry*, *55*(12), 1054–1063 e1053. doi:10.1016/j.jaac.2016.09.490

Hawgood, S., Hook-Barnard, I. G., O'Brien, T. C., & Yamamoto, K. R. (2015). Precision medicine: Beyond the inflection point. *Sci. Transl. Med.*, *7*(300), 300ps317. doi:10.1126/scitranslmed.aaa9970

Heiss, C. N., & Olofsson, L. E. (2019). The role of the gut microbiota in development, function and disorders of the central nervous system and the enteric nervous system. *J. Neuroendocrinol.*, *31*(5), e12684. doi:10.1111/jne.12684

Hollander, M., & Pena, E. A. (2004). Nonparametric methods in reliability. *Stat. Sci.*, *19*(4), 644–651. doi:10.1214/088342304000000521

Hughes, H. K., Rose, D., & Ashwood, P. (2018). The gut microbiota and dysbiosis in autism spectrum disorders. *Curr. Neurol. Neurosci. Rep.*, *18*(11), 81. doi:10.1007/s11910-018-0887-6

Hus, V., Gotham, K., & Lord, C. (2014). Standardizing ADOS domain scores: Separating severity of social affect and restricted and repetitive behaviors. *J. Autism. Dev. Disord.*, *44*(10), 2400–2412. doi:10.1007/s10803-012-1719-1

Hus, V., & Lord, C. (2014). The Autism Diagnostic Observation Schedule, module 4: Revised algorithm and standardized severity scores. *J. Autism. Dev. Disord.*, *44*(8), 1996–2012. doi:10.1007/s10803-014-2080-3

Insel, T. R. (2014). The NIMH Research Domain Criteria (RDoC) Project: Precision medicine for psychiatry. *Am. J. Psychiatry*, *171*(4), 395–397. doi:10.1176/appi.ajp.2014.14020138

Jang, Y., Wixted, J. T., & Huber, D. E. (2009). Testing signal-detection models of yes/no and two-alternative forced-choice recognition memory. *J. Exp. Psychol. Gen.*, *138*(2), 291–306. doi:10.1037/a0015525

Jasmin, E., Couture, M., McKinley, P., Reid, G., Fombonne, E., & Gisel, E. (2009). Sensori-motor and daily living skills of preschool children with autism spectrum disorders. *J. Autism Dev. Disord.*, *39*(2), 231–241. doi:10.1007/s10803-008-0617-z

Klin, A. (2000). Attributing social meaning to ambiguous visual stimuli in higher-functioning autism and Asperger syndrome: The Social Attribution Task. *J. Child Psychol. Psychiatry.*, *41*(7), 831–846.

Klin, A. (2008). In the eye of the beholden: Tracking developmental psychopathology. *J. Am. Acad. Child Adolesc. Psychiatry*, *47*(4), 362–363. doi:10.1097/CHI.0b013e3181648dd1

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002a). Defining and quantifying the social phenotype in autism. *Am. J. Psychiatry*, *159*(6), 895–908. doi:10.1176/appi.ajp.159.6.895

Klin, A., Jones, W., Schultz, R., Volkmar, F., & Cohen, D. (2002b). Visual fixation patterns during viewing of naturalistic social situations as predictors of social competence in individuals with autism. *Arch. Gen. Psychiatry*, *59*(9), 809–816.

Klin, A., Lin, D. J., Gorrindo, P., Ramsay, G., & Jones, W. (2009). Two-year-olds with autism orient to non-social contingencies rather than biological motion. *Nature*, *459*(7244), 257–261. doi:10.1038/nature07868

Kroll, N. E., Yonelinas, A. P., Dobbins, I. G., & Frederick, C. M. (2002). Separating sensitivity from response bias: Implications of comparisons of yes-no and forced-choice tests for models and measures of recognition memory. *J. Exp. Psychol. Gen.*, *131*(2), 241–254.

Kushki, A., Chau, T., & Anagnostou, E. (2011). Handwriting difficulties in children with autism spectrum disorders: A scoping review. *J. Autism Dev. Disord.*, *41*(12), 1706–1716. doi:10.1007/s10803-011-1206-0

Lilliefors, H. W. (1967). On the Kolmogorov-Smirnov test for normality with mean and variance unknown. *Journal of the American Statistical Association*, *62*, 399–402.

Lord, C., Risi, S., Lambrecht, L., Cook, E. H. Jr., Leventhal, B. L., DiLavore, P. C., . . . Rutter, M. (2000). The Autism Diagnostic Observation Schedule–Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *J. Autism Dev. Disord.*, *30*(3), 205–223.

Lovaas, O. I., Schreibman, L., & Koegel, R. L. (1974). A behavior modification approach to the treatment of autistic children. *J. Autism Child Schizophr.*, *4*(2), 111–129. doi:10.1007/bf02105365

Mandelbaum, D. E., Stevens, M., Rosenberg, E., Wiznitzer, M., Steinschneider, M., Filipek, P., & Rapin, I. (2006). Sensorimotor performance in school-age children with autism, developmental language disorder, or low IQ. *Dev. Med. Child Neurol.*, *48*(1), 33–39. doi:10.1017/S0012162206000089

Maurer, R. G., & Damasio, A. R. (1979). Vestibular dysfunction in autistic children. *Dev. Med. Child. Neurol.*, *21*(5), 656–659. doi:10.1111/j.1469-8749.1979.tb01682.x

Maurer, R. G., & Damasio, A. R. (1982). Childhood autism from the point of view of behavioral neurology. *J. Autism Dev. Disord.*, *12*(2), 195–205. doi:10.1007/bf01531309

Minshew, N. J., Sung, K., Jones, B. L., & Furman, J. M. (2004). Underdevelopment of the postural control system in autism. *Neurology*, *63*(11), 2056–2061.

Mosconi, M. W., Luna, B., Kay-Stacey, M., Nowinski, C. V., Rubin, L. H., Scudder, C., . . . Sweeney, J. A. (2013). Saccade adaptation abnormalities implicate dysfunction of cerebellar-dependent learning mechanisms in autism spectrum disorders (ASD). *PLoS One*, *8*(5), e63709. doi:10.1371/journal.pone.0063709

Mosconi, M. W., Mohanty, S., Greene, R. K., Cook, E. H., Vaillancourt, D. E., & Sweeney, J. A. (2015). Feedforward and feedback motor control abnormalities implicate cerebellar dysfunctions in autism spectrum disorder. *J. Neurosci.*, *35*(5), 2015–2025. doi:10.1523/JNEUROSCI.2731-14.2015

Mosconi, M. W., & Sweeney, J. A. (2015). Sensorimotor dysfunctions as primary features of autism spectrum disorders. *Sci. China Life. Sci.*, *58*(10), 1016–1023. doi:10.1007/s11427-015-4894-4

Mosconi, M. W., Wang, Z., Schmitt, L. M., Tsai, P., & Sweeney, J. A. (2015). The role of cerebellar circuitry alterations in the pathophysiology of autism spectrum disorders. *Front. Neurosci.*, 9, 296. doi:10.3389/fnins.2015.00296

Ng, Q. X., Loke, W., Venkatanarayanan, N., Lim, D. Y., Soh, A. Y. S., & Yeo, W. S. (2019). A systematic review of the role of prebiotics and probiotics in autism spectrum disorders. *Medicina (Kaunas)*, *55*(5). doi:10.3390/medicina55050129

Ornitz, E. M. (1974). The modulation of sensory input and motor output in autistic children. *J. Autism Child. Schizophr.*, *4*(3), 197–215.

Perry, W., Minassian, A., Lopez, B., Maron, L., & Lincoln, A. (2007). Sensorimotor gating deficits in adults with autism. *Biol. Psychiatry*, *61*(4), 482–486. doi:10.1016/j.biopsych.2005.09.025

Pulikkan, J., Mazumder, A., & Grace, T. (2019). Role of the gut microbiome in autism spectrum disorders. *Adv. Exp. Med. Biol., 1118*, 253–269. doi:10.1007/978-3-030-05542-4_13

Ryu, J., Vero, J., Dobkin, R. D., & Torres, E. B. (2019). Dynamic digital biomarkers of motor and cognitive function in Parkinson's disease. *J. Vis. Exp.*, *149*. doi:10.3791/59827

Scott, D. W. (1979). On optimal and data-based histograms. *Biometrika*, *66*(3), 605–610.

Scott, D. W. (2015). *Multivariate density estimation: Theory, practice, and visualization* (2nd ed.). Hoboken, NJ: Wiley.

Siaperas, P., Ring, H. A., McAllister, C. J., Henderson, S., Barnett, A., Watson, P., & Holland, A. J. (2012). Atypical movement performance and sensory integration in Asperger's syndrome. *J. Autism Dev. Disord.*, *42*(5), 718–725. doi:10.1007/s10803-011-1301-2

Somoza, E., & Mossman, D. (1991). ROC curves and the binormal assumption. *J. Neuropsychiatry Clin. Neurosci.*, *3*(4), 436–439. doi:10.1176/jnp.3.4.436

Starkstein, S., Gellar, S., Parlier, M., Payne, L., & Piven, J. (2015). High rates of Parkinsonism in adults with autism. *J. Neurodev. Disord.*, *7*(1), 29. doi:10.1186/s11689-015-9125-6

Swets, J. A. (1996). *Signal detection theory and ROC analysis in psychology and diagnostics: Collected papers*. Mahwah, NJ: Erlbaum.

Swets, J. A., & Pickett, R. M. (1982). *Evaluation of diagnostic systems: Methods from signal detection theory*. New York: Academic Press.

Tordjman, S., Davlantis, K. S., Georgieff, N., Geoffray, M. M., Speranza, M., Anderson, G. M., . . . Dawson, G. (2015). Autism as a disorder of biological and behavioral rhythms: Toward new therapeutic perspectives. *Front. Pediatr.*, *3*, 1. doi:10.3389/fped.2015.00001

Torres, E. B., Brincker, M., Isenhower, R. W., Yanovich, P., Stigler, K. A., Nurnberger, J. I., . . . Jose, J. V. (2013). Autism: The micro-movement perspective. *Front. Integr. Neurosci.*, *7*, 32. doi:10.3389/fnint.2013.00032

Torres, E. B., & Denisova, K. (2016). Motor noise is rich signal in autism research and pharmacological treatments. *Sci. Rep.*, *6*, 37422. doi:10.1038/srep37422

Torres, E. B., Isenhower, R. W., Nguyen, J., Whyatt, C., Nurnberger, J. I., Jose, J. V., . . . Cole, J. (2016). Toward precision psychiatry: Statistical platform for the personalized characterization of natural behaviors. *Front. Neurol.*, *7*, 8. doi:10.3389/fneur.2016.00008

Torres, E. B., Isenhower, R. W., Yanovich, P., Rehrig, G., Stigler, K., Nurnberger, J., & Jose, J. V. (2013). Strategies to develop putative biomarkers to characterize the female phenotype with autism spectrum disorders. *J. Neurophysiol.*, *110*(7), 1646–1662. doi:10.1152/jn.00059.2013

Torres, E. B., Mistry, S., Caballero, C., & Whyatt, C. P. (2017). Stochastic signatures of involuntary head micro-movements can be used to classify females of ABIDE into different subtypes of neurodevelopmental disorders. *Front. Integr. Neurosci.*, *11*, 10. doi:10.3389/fnint.2017.00010

Torres, E. B., Nguyen, J., Mistry, S., Whyatt, C., Kalampratsidou, V., & Kolevzon, A. (2016). Characterization of the statistical signatures of micro-movements underlying natural gait patterns in children with Phelan mcdermid syndrome: Towards precision-phenotyping of behavior in ASD. *Front. Integr. Neurosci.*, *10*, 22. doi:10.3389/fnint.2016.00022

Torres, E. B., Smith, B., Mistry, S., Brincker, M., & Whyatt, C. (2016). Neonatal diagnostics: Toward dynamic growth charts of neuromotor control. *Front. Pediatr.*, *4*, 121. doi:10.3389/fped.2016.00121

Torres, E. B., Vero, J., & Rai, R. (2018). Statistical platform for individualized behavioral analyses using biophysical micro-movement spikes. *Sensors (Basel)*, *18*(4). doi:10.3390/s18041025

Torres, E. B., & Whyatt, C. (2018). *Autism: The movement sensing perspective*. Boca Raton, FL: CRC Press/Taylor & Francis.

Torres, E. B., Yanovich, P., & Metaxas, D. N. (2013). Give spontaneity and self-discovery a chance in ASD: Spontaneous peripheral limb variability as a proxy to evoke centrally driven intentional acts. *Front. Integr. Neurosci.*, *7*, 46. doi:10.3389/fnint.2013.00046

Troyb, E., Knoch, K., Herlihy, L., Stevens, M. C., Chen, C. M., Barton, M., . . . Fein, D. (2016). Restricted and repetitive behaviors as predictors of outcome in autism spectrum disorders. *J. Autism Dev. Disord.*, *46*(4), 1282–1296. doi:10.1007/s10803-015-2668-2

Vilensky, J. A., Damasio, A. R., & Maurer, R. G. (1981). Gait disturbances in patients with autistic behavior: A preliminary study. *Arch. Neurol.*, *38*(10), 646–649. doi:10.1001/archneur.1981.00510100074013

White, K. G., & Wixted, J. T. (2010). Psychophysics of remembering: To bias or not to bias. *J. Exp. Anal. Behav.*, *94*(1), 83–94. doi:10.1901/jeab.2010.94-83

Whyatt, C., & Craig, C. (2013). Sensory-motor problems in autism. *Front. Integr. Neurosci.*, *7*, 51. doi:10.3389/fnint.2013.00051

Whyatt, C., & Torres, E. B. (2017). *The social-dance: Decomposing naturalistic dyadic interaction dynamics to the micro-level.* Paper presented at the MOCO 2017, London, UK.

Whyatt, C. P., & Torres, E. B. (2018). Autism research: An objective quantitative review of progress and focus between 1994 and 2015. *Front. Psychol.*, *9*, 1526. doi:10.3389/fpsyg.2018.01526

Witt, J. K., Taylor, J. E., Sugovic, M., & Wixted, J. T. (2015). Signal detection measures cannot distinguish perceptual biases from response biases. *Perception*, *44*(3), 289–300. doi:10.1068/p7908

Wu, D., Jose, J. V., Nurnberger, J. I., & Torres, E. B. (2018). A biomarker characterizing neurodevelopment with applications in autism. *Sci. Rep.*, *8*(1), 614. doi:10.1038/s41598-017-18902-w

Received July 30, 2019; accepted November 10, 2019.