2023

Development of webcam-collected and artificial- intelligence-derived social and cognitive performance measures for neurodevelopmental genetic syndromes

Thomas Frazier

Robyn M. Busch

Patricia Klaas

Katherine Lachlan

Shafali Jeste

See next page for additional authors

Authors

Thomas Frazier, Robyn M. Busch, Patricia Klaas, Katherine Lachlan, Shafali Jeste, Alexander Kolevzon, Eva Loth, Jacqueline Harris, Leslie Speer, Tom Pepper, Kristin Anthony, J Michael Graglia, Christal G. Delagrammatikas, Sandra Bedrosian-Sermone, Constance Smith-Hicks, Katie Huba, Robert Longyear, LeeAnn Green-Snyder, Frederick Shic, Mustafa Sahin, Charis Eng, Antonio Y. Hardan, and Mirko Uljarevic

Received: 22 June 2023 | Accepted: 19 July 2023

DOI: 10.1002/ajmg.c.32058

RESEARCH ARTICLE

AMERICAN JOURNAL OF medical genetics PART C Seminars in Medical Genetics

WILEY

Development of webcam-collected and artificial-intelligence-derived social and cognitive performance measures for neurodevelopmental genetic syndromes

Thomas W. Frazier[1,2] | Robyn M. Busch[3,4] | Patricia Klaas[3] | Katherine Lachlan[5] | Shafali Jeste[6] | Alexander Kolevzon[7] | Eva Loth[8] | Jacqueline Harris[9] | Leslie Speer[10] | Tom Pepper[11] | Kristin Anthony[12] | J. Michael Graglia[13] | Christal G. Delagrammatikas[14] | Sandra Bedrosian-Sermone[15] | Constance Smith-Hicks[9] | Katie Huba[1] | Robert Longyear[16] | LeeAnne Green-Snyder[17] | Frederick Shic[18] | Mustafa Sahin[19] | Charis Eng[4] | Antonio Y. Hardan[20] | Mirko Uljarević[20,21]

[1]Department of Psychology, John Carroll University, University Heights, Ohio, USA

[2]Departments of Pediatrics and Psychiatry, SUNY Upstate Medical University, Syracuse, New York, USA

[3]Department of Neurology, Neurological Institute, Cleveland Clinic, Cleveland, Ohio, USA

[4]Genomic Medicine Institute, Lerner Research Institute, Cleveland Clinic, Cleveland, Ohio, USA

[5]Human Genetics and Genomic Medicine, Faculty of Medicine, University of Southampton and Wessex Clinical Genetics Service, University Hospital Southampton NHS Foundation Trust, Southampton, UK

[6]Division of Neurology, Children's Hospital of Los Angeles, Los Angeles, California, USA

[7]Departments of Psychiatry and Pediatrics, Seaver Autism Center for Research and Treatment, Icahn School of Medicine at Mount Sinai, New York, New York, USA

[8]Department of Forensic and Neurodevelopmental Science, Institute of Psychiatry, Psychology and Neuroscience, Kings College London, London, UK

[9]Department of Neurology, Kennedy Krieger Institute and Johns Hopkins University School of Medicine, Baltimore, Maryland, USA

[10]Frazier Behavioral Health, Cleveland, Ohio, USA

[11]PTEN Research Foundation, Cheltenham, UK

[12]PTEN Hamartoma Tumor Syndrome Foundation, Huntsville, Alabama, USA

[13]SYNGAP Research Fund, Palo Alto, California, USA

[14]Malan Syndrome Foundation, Old Bridge, New Jersey, USA

[15]ADNP Kids Foundation, Brush Prairie, Washington, USA

[16]Autism Analytica, Syracuse, New York, USA

[17]Simons Foundation, New York, New York, USA

[18]Department of Pediatrics, University of Washington and Seattle Children's Research Institute, Seattle, Washington, USA

[19]Rosamund Stone Zander Translational Neuroscience Center, Department of Neurology, Boston Children's Hospital and Harvard Medical School, Boston, Massachusetts, USA

[20]Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, California, USA

[21]Melbourne School of Psychological Sciences, Faculty of Medicine, Dentistry, and Health Sciences, The University of Melbourne, Melbourne, Victoria, Australia

[Correction added after first online publication on 18 August 2023. The author name "Graglia J. Michael" has been changed to "J. Michael Graglia".]

Correspondence
Thomas W. Frazier, Department of Psychology, John Carroll University, University Heights, OH, USA.
Email: tfrazier@jcu.edu

Funding information
Autism Speaks, Grant/Award Number: 12776; Malan Syndrome Foundation; PTEN Research Foundation, Grant/Award Number: JCU-20-001; Simons Foundation Autism Research Initiative, Grant/Award Number: 831500; PTEN Hamartoma Tumor Syndrome Foundation; SYNGAP Research Fund; ADNP Kids Foundation

Abstract

This study focused on the development and initial psychometric evaluation of a set of online, webcam-collected, and artificial intelligence-derived patient performance measures for neurodevelopmental genetic syndromes (NDGS). Initial testing and qualitative input was used to develop four stimulus paradigms capturing social and cognitive processes, including social attention, receptive vocabulary, processing speed, and single-word reading. The paradigms were administered to a sample of 375 participants, including 163 with NDGS, 56 with idiopathic neurodevelopmental disability (NDD), and 156 neurotypical controls. Twelve measures were created from the four stimulus paradigms. Valid completion rates varied from 87 to 100% across measures, with lower but adequate completion rates in participants with intellectual disability. Adequate to excellent internal consistency reliability ($\alpha = 0.67$ to 0.95) was observed across measures. Test–retest reproducibility at 1-month follow-up and stability at 4-month follow-up was fair to good ($r = 0.40$–0.73) for 8 of the 12 measures. All gaze-based measures showed evidence of convergent and discriminant validity with parent-report measures of other cognitive and behavioral constructs. Comparisons across NDGS groups revealed distinct patterns of social and cognitive functioning, including people with *PTEN* mutations showing a less impaired overall pattern and people with *SYNGAP1* mutations showing more attentional, processing speed, and social processing difficulties relative to people with *NFIX* mutations. Webcam-collected performance measures appear to be a reliable and potentially useful method for objective characterization and monitoring of social and cognitive processes in NDGS and idiopathic NDD. Additional validation work, including more detailed convergent and discriminant validity analyses and examination of sensitivity to change, is needed to replicate and extend these observations.

KEYWORDS
eye tracking, facial expressions, genetic syndromes, neurodevelopment, webcam

1 | INTRODUCTION

Advances in identifying pathogenic variation linked to neurodevelopmental disability (NDD) has accelerated the discovery of a growing number of specific neurodevelopmental genetic syndromes (NDGS). As NDGS are identified, natural history investigations have begun to characterize a wide spectrum of medical conditions and neurobehavioral strengths and weaknesses associated with each condition (Busch et al., 2023; Mulder et al., 2020; Vlaskamp et al., 2019). This work is crucial to developing patient support guidelines and ensuring that patients with NDGS receive appropriate supports that maximize their development. For example, in individuals with PTEN hamartoma tumor syndrome (PHTS) resulting from germline heterozygous mutations in *PTEN*, a spectrum of frontal-systems deficits has been identified from no impairment to very severe impairment associated with intellectual disability (ID) and autism spectrum disorder (ASD) (Busch et al., 2019; Ciaccio et al., 2018; Frazier et al., 2015; Steele et al., 2021). This pattern has been found to be stable over a period of 2 years (Busch et al., 2023), even in young children, and the specific profile of frontal systems impairment can be used to inform clinical and educational care (Frazier, 2019).

While there have been some initial attempts to provide more detailed characterization of neurobehavioral profiles across different NGDS, yield from the natural history and neurobehavioral studies have been limited by the lack of comprehensive and sensitive instruments appropriate for evaluations with geographically dispersed populations. For example, within the Rare Disease Clinical Research Network—Developmental Synaptopathies Consortium natural history study of individuals with PHTS and ASD (Busch et al., 2019), in-person cognitive assessments were limited to annual visits and often required several hours of testing to collect data from relevant neurocognitive domains. Because of the extensive effort required, the related pilot clinical trial initiated within this network was limited to three in-person assessments over a 6-month study period (Hardan et al., 2021; Srivastava et al., 2022). The infrequency, difficulty, and burden of these traditional approaches highlight the need for new phenotyping methods.

Identification of NDGS has also accelerated the development of syndrome-specific patient advocacy groups and foundations, as well

as programs of research designed to better understand and translate molecular, cellular, and circuitry findings into intervention strategies. A primary goal of these patient advocacy groups—and the research programs they support—is to develop and evaluate the efficacy of personalized interventions. Recent reviews of NDGS have emphasized the need to understand pathophysiology and neurobehavioral profiles to generate personalized therapeutic strategies (Frazier, 2019; Sahin & Sur, 2015). Yet, given the small number of specialty clinics focused on each NDGS, and practical geographic constraints, many patients remain under-served and many clinics lack resources to collect extensive neurobehavioral assessments during clinic visits. Relatedly, due to the rare nature of many NDGS, natural history studies often rely on small sample sizes, which limits their value in identifying clinical endpoints for trials. In these small-sample longitudinal contexts, it is important to have reliable, stable indicators of individual performance, as compared to larger group studies where statistical certainty can be bolstered by adding participants. Having repeatable, online measures of neurobehavioral function could substantially improve both the statistical power of translational and clinical studies and increase the ability to more rapidly and sensitively identify individual differences in the pattern of intervention response. Administration of these measures in the individual's home rather than within a clinic setting would not only broaden access to research participation but might also reduce biases resulting from collection of neurobehavioral information in an unfamiliar setting.

Research in NDGS and idiopathic NDD is also limited by reliance on subjective measurements acquired from parents/caregivers and/or observations by clinician scientists, which has precipitated a call for the development of objective measures (Sahin et al., 2018). As a result, a number of tools have begun to be developed and have shown promise for objectively evaluating and tracking key functions relevant to neurodevelopment (Amit et al., 2020; Dawson et al., 2018; Egger et al., 2018; Goodwin et al., 2019; Manfredonia et al., 2019; McPartland et al., 2020; Ness et al., 2019; Tuncgenc et al., 2021). However, with a few notable exceptions, these measures have been developed solely for in-person evaluation, limiting their application and temporal sensitivity. In addition, noted measures have predominantly focused on the evaluation of only single domains rather than providing a more detailed characterization of multiple social, developmental, and cognitive domains. Furthermore, a high percentage of individuals with NDGS have significant cognitive and functional impairments. A relatively brief and repeatable battery of objective measures that can reliably capture a wide range of cognitive and behavioral capacities could supplement existing tools while simultaneously increasing sensitivity to intervention effects.

One possibility that can increase the objectivity of NDGS evaluations and simultaneously overcome accessibility barriers is to augment traditional characterization methods with appropriately designed remotely administered measures of neurobehavioral function. Designing remote measures for maximal accessibility has the potential to lower burden for providers as well as patients. Webcam-based eye tracking is a remote data collection method that uses cameras on everyday computing devices coupled with artificial-intelligence/machine learning algorithms to capture individual looking patterns toward probes such as the presentation of videos and images. Webcam data collection also permits the frame-by-frame automated facial expression analysis using machine learning algorithms that enable prototype matching using large training datasets. The potential for these methods to inform neurodevelopment is strong and, increasingly, both webcam-collected data (Simmatis et al., 2023) and artificial intelligence/machine learning algorithms (Nerusil et al., 2021) are being applied to create novel biometric measures for assessing child development and neurological conditions. A key advantage of webcam-based data collection is that the paradigms can be administered without direct real-time clinical supervision. Thus, an online, webcam-collected patient performance battery, capturing relevant social and cognitive measurements in an objective way, could supplement in-person assessment of NDGS patients and provide a more temporally sensitive picture of neurobehavioral development in these populations. This is particularly true for individuals with medical and mental health comorbidities and cognitive impairments that merit closer surveillance but are currently underserved (Vlaskamp et al., 2019).

Unfortunately, at present, there are no accessible, scalable objective measures specifically designed for rapid and repeated evaluation of multiple social and cognitive domains important to NDGS and idiopathic NDD. The primary aim of this study was to address this limitation and develop social and cognitive stimulus paradigms that could be paired with webcam collection and artificial intelligence algorithms to measure key neurocognitive processes relevant to NDGS. Webcam-collected measures were developed in conjunction with clinician–scientist experts, patients, and parents/caregivers, following gold-standard principles of measure development (Boateng et al., 2018) and inclusive practices (FDA, 2009), to complement our recently developed and validated informant-report survey scales (Frazier et al., 2023). Individual paradigms were created to be brief (3–4 min) and to require only spontaneous or directed gaze, without motor or speech responses, thus making it appropriate for a wide range of developmental and cognitive levels. Stimuli followed best practices in gaze collection (Sasson & Elison, 2012) and test development (Boateng et al., 2018), including teaching parents to facilitate data collection (when needed) without interfering in the evaluation, presenting large elements within the visual field to limit accuracy issues in webcam gaze collection (Semmelmann & Weigelt, 2018), and, where relevant, focusing on very easy initial items with a graded increase in task difficulty. Based on careful attention to applicability to a wide range of individuals with NDGS, valid measure collection was expected to be achieved in the majority of participants, including those with ID.

A secondary aim of this study was to conduct initial psychometric evaluation of these measures in several distinct NDGS groups, people with idiopathic NDD, and neurotypical controls. Initial evaluation included estimation of scale reliability, test–retest reproducibility (1-month follow-up), and stability (4-month follow-up). Initial convergent and discriminant validity was assessed using data from other informant(parent)-reported clinical information (Frazier et al., 2023). In

addition, given the importance of detecting autism within NDGS to ensure access to appropriate services, concurrent validity with ASD diagnoses and autism symptom levels was evaluated. Finally, using baseline data, exploratory analyses examined the pattern of cognitive and behavioral functioning across NDGS and idiopathic NDD.

2 | METHODS

2.1 | Initial stimulus development

The stimulus paradigm development process is outlined in Online Appendix 1. Briefly, this included identifying or creating appropriate target items and stimuli across a wide range of ages (3–45) and ability levels (moderate to severe cognitive impairment to average ability levels); collecting feasibility data; updating items and stimuli based on initial feedback; conducting a pilot administration of performance measures with 10 clinician–scientist experts and 9 parents and patients with NDGS and/or idiopathic NDD; and administering a post-evaluation survey to collect additional feedback and create the final performance paradigms.

The social paradigm and associated stimuli were chosen based on the combination of empirical work (Frazier et al., 2018) and comprehensive review of the literature (Chita-Tegmark, 2016; Frazier et al., 2017). Specifically, a variety of social stimuli were selected, in part, due to the high rates of ASD occurrence in NDGS and the broader relevance of social attention to neurodevelopment as a transdiagnostic construct (Frazier, Uljarevic, et al., 2021; Salley & Colombo, 2016). The processing speed paradigm was selected because of the potential to use this cognitive paradigm to capture attentional scanning across the stimulus field, measure speed of object detection via gaze, the ease-of-administration in individuals with NGDS, particularly those with limited speech or motor difficulties, and the ability to create easier stimuli relevant to individuals with more significant intellectual impairments. Importantly, processing speed has been shown to be a very sensitive index of brain development and neuropathophysiological processes (Bove et al., 2021; Kail, 1991). The receptive vocabulary paradigm was selected because receptive language is a strong indicator of developmental trajectory and functional outcome (Frazier, Klingemier, et al., 2021) and can validly estimate results from standardized in-person testing using gaze to visual targets (Frazier et al., 2020). The single-word reading paradigm was developed based on a recommendation by clinician–scientist experts for identifying early reading, including in people with limited or no speech where reading is more difficulty to assess. This paradigm was also included based on its potential to monitor development of reading throughout childhood and early adulthood in NDGS. Additional information for receptive vocabulary and single-word reading target selection and stimulus creation are provided in Online Appendices 2 and 3. Example screenshots for each of the performance paradigms are included in Online Appendices 4–7, and stimulus/target order and composition information are provided in Online Appendices 8–11.

2.2 | Clinician–scientist experts and parent pilot evaluation feedback

Ten clinician–scientist experts were recruited based on their clinical and/or research expertise with a specific NDGS group or idiopathic NDD. Nine parent-patient pairs were recruited from the respective groups (6 PHTS, 1 *NFIX*, 1 *SYNGAP1*, 1 *ADNP*, and 1 idiopathic ASD). Patients were intentionally selected to represent a range of ages and cognitive levels. After completing a pilot administration of performance paradigms, clinician–scientist experts and parents—who facilitated the webcam administration for the patient participant—completed a post-evaluation survey. Questions are provided in Online Appendices 12 and 13. This information was used to generate final stimulus videos and to improve the training of parents in facilitating administration to the child.

2.3 | Parent/caregiver administration support training

Based on initial feedback, a parent/caregiver training process was developed (Online Appendix 14). This process included the following elements: (1) introduction to webcam technology, (2) training video, (3) parent completion of a "practice" stimulus set, (4) online training in valid task completion, and (5) virtual support meetings during initial and follow-up administrations. All of the elements were optional, but most participants used at least one option, and nearly all participants completed the parent "practice" stimuli.

2.4 | Webcam collection of gaze

Participants were instructed to use a device with at least a 10 in. screen size based on results of initial pilot testing, which indicated that smaller screen sizes could reduce accuracy of point-of-regard relative to specific areas-of-interest. Webcam data were collected and processed using proprietary CoolTool software. The software was originally intended as a neuromarketing tool, but initial feasibility testing, including with several young children with neurodevelopmental disabilities, indicated good potential for use as a data collection platform. The minimum required camera resolution was 720p at 30 fps. The gaze collection algorithm included a five-point calibration routine prior to each paradigm administration. This routine is coupled with a machine learning algorithm and was designed to detect webcam position within the three-dimensional (3D) space and intended to maximize gaze accuracy. On a frame-by-frame basis, gaze position relative to the two-dimensional screen was estimated. While accurate calibration is desirable, the gaze estimation model often functions adequately when less than ideal calibration data are acquired, making the system ideal for young and more impaired participants. Similar systems have been shown to have achieve ~3–5° of calibration uncertainty, translating to accurate detection of areas >10% of screen size (Semmelmann & Weigelt, 2018; Shehu et al., 2021). The present

stimulus paradigms were built with large areas-of-interest to be tolerant of higher levels of gaze uncertainty. Importantly, any reductions in gaze accuracy should reduce the reliability and validity of gaze-based measurements. Thus, observations of high reliability and evidence of convergent validity would suggest minimal impact of suboptimal gaze calibration. To offset concerns regarding possible reductions in gaze calibration and accuracy negatively impacting neurobehavioral measurements, no indices were scored if total time with eyes on screen was estimated to be less than 30 s overall (out of 15 min of possible gaze time to the screen).

Areas-of-interest were generated for each stimulus. For social attention stimuli, these include both socially relevant (e.g., faces, target objects, etc.) and socially irrelevant stimuli (e.g., foreground and background distractors, nontarget objects), based on our prior research (Frazier et al., 2018). For processing speed, receptive vocabulary, and single-word reading stimuli, areas-of-interest included target items/objects. For all stimuli, areas-of-interest are temporally defined based on expected gaze patterns from prior research (social attention) (Frazier et al., 2018) or after the verbal directive has been given (cognitive paradigms) (Frazier et al., 2020).

2.5 | Automated scoring of facial expressions

The webcam software also includes a proprietary algorithm for automatically scoring facial expressions. Facial landmarks are identified in the 3D space and the artificial intelligence algorithm is applied to these landmarks on a frame-by-frame basis to generate probability scores based on accuracy of classification from training data (Kuntzler et al., 2021). Probability scores represent a match between the facial landmark configuration and known sets of facial expressions (fear, anger, disgust, sadness, surprise, joy, and neutral), with closer matches being interpreted as higher intensities of expression (range 0–100%). For the present study, and because specific affect recognition intensities can be prone to error for more subtle expressions (Kuntzler et al., 2021), specific expressions were aggregated into positive and negative categories to maximize reliability. Facial expression measures were only collected to the social attention stimuli, as these showed the greatest range of non-neutral expressions in preliminary data.

2.6 | Development of a priori validity criteria and scoring

For each social and cognitive paradigm, the investigative team a priori identified possible gaze and facial expression measures that would be relevant to evaluating social and cognitive processes in NDGS and idiopathic NDD. The only exception to this is the social attention measure which was empirically developed following our prior published methodology (Frazier et al., 2018) (see Online Appendix 15 for additional information). Online Appendix 16 presents operational definitions for each performance measure. Each gaze-based measure was only scored if stringent validity criteria were met. Online Appendix 17 includes validity criteria for all 12 webcam-collected measures. For each measure, validity criteria ensure that the participant attended to the stimuli for at least 30 s, and at least eight valid targets or four valid stimuli were collected. Fixations were scored by identifying at least 66 ms of gaze point samples within a 100-pixel dispersion. Four gaze metrics are calculated for each area-of-interest – fixation duration, fixation count, glance count, and time-to-first fixation (Online Appendix 18). These metrics were used to score the 12 performance measures evaluated in this study.

2.7 | Participants for initial measure evaluation

NDGS groups included participants with PHTS, *ADNP*, *SYNGAP1*, or *NFIX* recruited via contacts through the PHTS Foundation with the support of the PTEN Research Foundation, the ADNP Kids Foundation, the SYNGAP Research Fund, and the Malan Syndrome Foundation. Other individuals with NDGS were recruited via the Simons Foundation Searchlight registry and included people with mutations in *GRIN2B*, *CSNK2A1*, *HIVEP2*, *SCN2A*, *MED13L*, and *STXBP1*. Given the relatively small sample sizes for ADNP ($n = 11$) and these NDGS groups, they were combined into a single "other NDGS" group ($n = 63$). Individuals were included if they were between the ages of 3 and 45 at enrollment and had an available parent or other close relative/caregiver to complete informant-report measures. Siblings of individuals with NDGS were also eligible to participate, and unrelated neurotypical controls were recruited using StudyKik, a national recruitment service. Siblings and unrelated controls who were reported to have an idiopathic NDD were included in a separate group.

2.8 | Procedure

Parent/caregiver informants first completed a demographic and clinical information questionnaire followed by 11 neurobehavioral evaluation tool (NET) survey scales (Frazier et al., 2023). These survey scales included six measures of symptoms/problems (anxiety, attention-deficit/hyperactivity disorder, restricted/repetitive behavior, challenging behavior, mood, and sleep problems) and five measures of skills/functioning (motor skills, daily living skills, social communication/interaction skills, executive functioning, and quality of life). After NET survey completion, informants and participants were instructed to complete webcam-collected performance measures and were sent links via email or text to facilitate completion. For young and/or impaired children, performance measure administration began by having the parent complete a practice version, so that they understood how the webcam collection works and how best to help their child. Parents and older patients also were offered a video call with the research coordinator to review best practices in performance measure administration and were provided a set of recommendations to improve evaluation validity.

Performance measure administration began with the five-point calibration that included dots presented in the four corners and center

of the screen. Next, videos were presented for each paradigm in succession—social attention, receptive language, processing speed, and single-word reading. Recalibration automatically occurred prior to each paradigm.

Survey and webcam measures were collected at baseline, 1-month, and 4-month follow-up timepoints. The maximum total administration time across all paradigms was 15 min (social attention—4 min, receptive vocabulary—4 min, processing speed—3 min, single-word reading—4 min) with videos separated into 1-min segments to permit breaks. A button press was required to advance to the next video. Participants were instructed to complete all of the social attention and processing speed videos, but were permitted to complete only the first 2 min of the receptive vocabulary paradigm and complete only 1 min of the single-word reading paradigm dependent on the parent's appraisal of the patient's capacity to engage with the paradigm. Participants could proceed through all paradigms or take breaks between paradigms but were encouraged to finish all videos in one sitting if possible.

IRB approval was obtained for all of the qualitative and quantitative procedures of the study, including administration of the performance measures, and parents/legally authorized representatives and adult patients provided informed consent prior to completing any study procedures. Assent for minors was also obtained, where appropriate.

2.9 | Statistical analyses

2.9.1 | Sample characterization

Descriptive statistics for demographic and clinical factors were computed to characterize the sample, and chi-square or univariate ANOVA were used to compare across the seven study groups (PHTS, *SYNGAP1*, *NFIX*, other NDGS, idiopathic NDD, sibling controls, and unrelated neurotypical controls).

2.9.2 | Evaluation and measure validity

Using validity criteria for each of the 12 performance measures, the sum of valid measures was computed and compared across study groups using univariate ANOVA. Proportions of validity by measure were also computed overall and by parent-reported ID status.

2.9.3 | Reliability

Scale reliability (internal consistency) was calculated using Cronbach's alpha (α) (Streiner & Norman, 1995). Scale reliability estimates falling in the ranges 0.70 to 0.79, 0.80 to 0.89, and >0.90 were considered fair, good, and excellent (Nunnally & Bernstein, 1994), respectively. Test–retest reproducibility (1-month follow-up) and stability (4-month follow-up) were estimated using Pearson's bivariate correlations. Test–retest estimates <0.40 were considered poor, 0.40 to 0.59 fair, 0.60 to 0.74 good, and 0.75+ excellent (Cicchetti et al., 2006).

2.9.4 | Convergent and discriminant validity

To evaluate convergent and discriminant validity, other clinical information based on informant-report was a priori selected as either measuring similar constructs (convergent validity) or measuring dissimilar constructs (discriminant validity) for each performance measure. Informant-report information included: estimated IQ; speech level (5-point scale from non or minimally speaking to fluent speech); reading level (5-point scale from no reading to paragraph level or higher). ADHD, anxiety, mood, challenging behavior, social communication/interaction, and restricted repetitive behavior symptoms; sleep problems; daily living skills; executive functioning; and motor skills. Bivariate correlations were computed between each performance measure and the convergent and discriminant validity measures selected. To compute aggregate correlations over multiple measures, correlations were converted to Fisher's *z*, averaged, and transformed back to a correlation metric. The test of the significance of the difference in dependent correlations was used to examine whether convergent validity correlations were higher than discriminant validity correlations (Cohen & Cohen, 1983).

2.9.5 | Concurrent validity with ID, ASD diagnoses, and autism symptom levels

To examine concurrent validity of performance measures with parent-report clinical ID diagnosis, independent samples *t* tests were computed with each measure as the dependent variable and ID status (yes, no) as the grouping variable. Cohen's *d* was computed to estimate the magnitude of group differences. To evaluate potential diagnostic validity, receiver operating characteristic (ROC) curve analyses were calculated in the training, testing, validation, and testing plus validation subsamples, separately for baseline, 1-month, and 4-month follow-up data. Areas under the curve (AUCs) evaluated diagnostic validity. A rough guideline for evaluating AUC values is: < 0.60 = poor, 0.60–0.69 = fair, 0.70 to 0.79 = good, 0.80–0.89 = excellent if the comparison group is clinically meaningful; and 0.90–1.00 = exceptional only if the design and comparison are appropriate (Youngstrom et al., 2019). To evaluate concurrent validity with autism symptom levels, autism symptom levels derived from neurobehavioral evaluation survey scales were calculated and correlations were computed in the same subsamples as ROC analyses.

2.9.6 | Neurobehavioral patterns across NDGS and idiopathic NDD groups

To explore unique patterns of social and cognitive function, webcam measures were first normed using regression-based norming in

unrelated healthy controls, with age, the square of age (to capture nonlinear developmental trends), and sex included as predictors in each equation. This approach puts each measure on a *z*-score metric relative to healthy controls. Using these standardized residual scores, univariate analysis of variance models were computed, with each of the seven groups as the independent variable and the performance measure scores as dependent variables in separate analyses.

2.9.7 | Statistical power

Assuming total sample sizes of 200+ for reliability and validity analyses, statistical power to detect a bivariate correlation of r ≥ 0.40 was excellent (>0.99; one-tailed *p*-value of 0.05). Assuming minimal subsample sizes of at least 18 ASD and 40 non-ASD diagnosed individuals, power to detect AUCs ≥0.72 was at least good (≥0.80). Statistical power to detect group differences across webcam performance measures, assuming a minimum sample size of 24, was at least adequate (>0.82) if large group differences were observed ($d \geq 0.80$; $\alpha = 0.05$, two-tailed). For larger group sizes ($n > 40$), power was adequate, even for medium effects ($d \geq 0.50$).

2.9.8 | Statistical analysis implementation

Statistical significance was set at $\alpha = 0.05$, two-tailed, and effect size magnitude was emphasized. Data preparation, descriptive analyses, internal consistency reliability using Cronbach's alpha (α), and bivariate correlations were computed in SPSS v28 (IBM Corp, 2021). ROC analyses were computed using the R package *pROC* and implemented in version 4.1.2 (R Core Team, 2021) using *R Studio* version 2021.09.1.

3 | RESULTS

3.1 | Pilot evaluation results

Clinicians used a wide range of hardware setups and reported high relevance of the paradigms to their respective NDGS or idiopathic NDD group (Online Appendix 19). Clarity of instructions and quality of audio and visual stimuli was rated as high. Timing was rated as generally moderate (neither fast nor slow). Several potential concerns about target difficulty levels were raised and used to adjust the final stimuli.

Parents rated the overall experience as positive and of relatively moderate difficulty across paradigms (Online Appendix 20). Patient participants did not require breaks, looked away from the screen with variable frequency (every 5–10 s to only a few losses of attention to screen), covered or touched their face only infrequently, and required variable levels of physical, gestural, or verbal assistance to maintain motivation and attention. Unexpected intrusions and adjustments to lighting were infrequent. Overall attention was rated as average to good. Paradigm relevance to the patient's condition was rated as "relevant" to "highly relevant" across paradigms. Quality of audio and visual stimuli was rated as high, and timing was judged to be generally moderate to fast. These data were used to adjust parent training processes and to include reminders to limit assistance to motivation and general attention (not specific to a stimulus or desired response).

3.2 | Sample characteristics

A total of 395 individuals enrolled to participate before May 4, 2023 (recruitment is ongoing). Of these, 20 did not attempt baseline webcam paradigms, but of the 375 who did attempt the paradigms, all achieved at least 1 valid measure (Online Appendix 21). Longitudinal attrition was modest at 1-month follow-up ($n = 54$ did not attempt; $n = 341$ attempted) but higher at 4-month follow-up ($n = 100$ did not attempt; $n = 295$ attempted).

Table 1 presents sample characteristics. Findings were highly consistent with findings in our recent survey validation study (Frazier et al., 2023). Specifically, participants were younger in the *NFIX* and *SYNGAP1* groups and older in the PHTS and idiopathic NDD groups, with high rates of spousal informants in the latter groups. All groups had very high proportions of White/Caucasian participants, although Hispanic ethnicity approximated US population proportions in most groups, and the sample had a wide range of household incomes. Estimated cognitive levels were lowest in the *NFIX*, *SYNGAP1*, and other NDGS groups and to a lesser extent in the PHTS group relative to control groups. Informant-reported developmental diagnoses were highly variable across NDGS groups, but with elevated rates of ASD, ID, anxiety, and motor disorder in *NFIX*, *SYNGAP1*, and other NDGS groups compared to controls. Participants were predominantly from the United States ($n = 325$, 87%), but a small minority of participants with informants fluent in English were also included from other countries (United Kingdom $n = 17$, Canada $n = 24$, Australia $n = 4$, New Zealand $n = 1$, Ireland $n = 2$, Netherlands $n = 1$, Israel n = 1).

3.3 | Evaluation validity

Evaluation validity was high across all groups, but *NFIX*, *SYNGAP1*, and other NDGS groups had higher proportions of individuals with at least one invalid measure (Table 1). On average, all groups had at least 10 valid performance measures. Participants with reported ID had lower measure validity proportions than participants without ID, but measure validity never dropped below 84% (Table 2).

Score distributions were variable across measures, with many showing near normal distributions, and all but negative emotion suggesting a good quantitative range (Online Appendix 22). The latter was highly skewed and kurtosis with scores clustered close to 0%.

3.4 | Reliability

Internal consistency reliability was good to excellent for all performance measures ($\alpha = 0.89$–0.95; Table 2), with the exception of

TABLE 1 Demographic and clinical characteristics by study group.

	Sibling controls	Unrelated controls	PHTS	*NFIX*	*SYNGAP1*	Other NDGS	NDD	X^2/F (*p*)
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	
N	40	116	33	24	43	63	56	
***Informant age* (*M*, *SD*)**	42 (6)	42 (9)	43 (8)	41 (10)	42 (8)	44 (8)	42 (8)	0.6 (0.718)
***Informant sex* (% female)**	37 (93%)	95 (82%)	28 (85%)	21 (88%)	39 (91%)	61 (97%)	51 (91%)	12.3 (0.424)
Informant relationship to participant								39.3 (0.003)
Biological parent	39 (98%)	99 (85%)	25 (76%)	23 (96%)	40 (93%)	59 (93%)	44 (79%)	
Adoptive or custodial parent	0 (0%)	3 (3%)	1 (3%)	1 (4%)	1 (2%)	4 (6%)	2 (4%)	
Other biological relative/sibling	1 (2%)	7 (6%)	0 (0%)	0 (0%)	1 (2%)	0 (0%)	3 (5%)	
Spouse/other non-biological relative	0 (0%)	7 (6%)	7 (21%)	0 (0%)	1 (2%)	0 (0%)	7 (12%)	
***Household income* (*US $*)**								79.7 (0.013)
< $25,000	1 (3%)	5 (4%)	2 (6%)	0 (0%)	0 (0%)	2 (3%)	8 (14%)	
$25,000–$34,999	2 (5%)	8 (7%)	0 (0%)	2 (8%)	1 (2%)	1 (2%)	2 (4%)	
$35,000–$49,999	1 (3%)	5 (4%)	1 (3%)	3 (13%)	3 (7%)	3 (5%)	6 (11%)	
$50,000–$74,999	6 (15%)	18 (16%)	9 (27%)	4 (17%)	3 (7%)	4 (6%)	11 (20%)	
$75,000–$99,999	2 (5%)	21 (18%)	3 (9%)	3 (13%)	4 (9%)	6 (10%)	4 (7%)	
$100,000–$149,999	7 (18%)	28 (24%)	7 (21%)	4 (17%)	10 (23%)	16 (25%)	11 (20%)	
$150,000–$199,999	7 (18%)	14 (12%)	4 (12%)	5 (21%)	10 (23%)	6 (10%)	6 (11%)	
$200,000+	6 (15%)	13 (11%)	2 (6%)	2 (8%)	7 (16%)	12 (19%)	5 (9%)	
Did not report	8 (20%)	4 (3%)	5 (15%)	1 (4%)	5 (11%)	13 (21%)	3 (5%)	
***Participant age* (*M*, *SD*)**	11 (5)	12 (8)	17 (13)	10 (7)	10 (7)	11 (6)	16 (9)	4.8 (<0.001)
***Participant sex* (% female)**	23 (58%)	63 (54%)	13 (39%)	12 (50%)	19 (44%)	36 (57%)	21 (38%)	8.6 (0.197)
Participant race/ethnicity								
White/Caucasian	36 (90%)	95 (82%)	30 (91%)	24 (100%)	37 (86%)	58 (92%)	46 (82%)	9.6 (0.142)
Black/African American	3 (8%)	9 (8%)	2 (6%)	0 (0%)	5 (12%)	5 (8%)	8 (14%)	5.6 (0.473)
Middle Eastern or North African	2 (5%)	1 (1%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	9.1 (0.167)
East Asian	2 (5%)	9 (8%)	3 (9%)	0 (0%)	2 (5%)	5 (8%)	2 (4%)	3.8 (0.697)
South Asian	2 (5%)	8 (7%)	0 (0%)	0 (0%)	1 (2%)	3 (5%)	0 (0%)	8.2 (0.223)
Native American/Alaskan Native	0 (0%)	3 (3%)	1 (3%)	1 (4%)	0 (0%)	0 (0%)	1 (2%)	4.5 (0.605)
Native Hawaiian/Pacific Islander	0 (0%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	2.2 (0.896)
Hispanic	7 (18%)	21 (18%)	1 (3%)	5 (21%)	7 (17%)	2 (3%)	11 (20%)	18.7 (0.096)
Unknown	0 (0%)	2 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	4.5 (0.611)
Did not report	0 (0%)	2 (2%)	0 (0%)	0 (0%)	0 (0%)	0 (0%)	1 (2%)	3.6 (0.734)
***Cognitive level* (informant-estimated)**								337.9 (<0.001)
Very high or above (120+)	6 (15%)	12 (10%)	3 (9%)	0 (0%)	0 (0%)	1 (2%)	10 (18%)	
High average (110–119)	18 (45%)	58 (50%)	6 (18%)	0 (0%)	0 (0%)	0 (0%)	19 (34%)	
Average (90–109)	13 (33%)	42 (36%)	15 (46%)	0 (0%)	1 (2%)	2 (3%)	22 (39%)	
Below average (80–89)	0 (0%)	0 (0%)	1 (3%)	2 (8%)	4 (9%)	6 (10%)	2 (4%)	
Borderline impairment (70–79)	0 (0%)	0 (0%)	2 (6%)	2 (8%)	1 (2%)	2 (3%)	0 (0%)	
Mild impairment (55–69)	0 (0%)	0 (0%)	1 (3%)	5 (21%)	6 (14%)	12 (19%)	3 (5%)	
Moderate impairment (40–54)	0 (0%)	0 (0%)	2 (6%)	9 (38%)	11 (26%)	17 (27%)	0 (0%)	
Severe impairment (21–39)	0 (0%)	0 (0%)	0 (0%)	2 (8%)	10 (23%)	12 (19%)	0 (0%)	
Profound impairment (<20)	0 (0%)	0 (0%)	0 (0%)	2 (8%)	5 (12%)	3 (5%)	0 (0%)	
Did not report	3 (8%)	4 (3%)	3 (9%)	2 (8%)	5 (12%)	8 (13%)	0 (0%)	

TABLE 1 (Continued)

	Sibling controls	Unrelated controls	PHTS	*NFIX*	*SYNGAP1*	Other NDGS	NDD	X^2/F (*p*)
	n (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	*n* (%)	
Cognitive estimate from prior testing	6 (15%)	19 (16%)	16 (49%)	13 (54%)	21 (49%)	30 (48%)	26 (46%)	57.1 (<0.001)
Developmental diagnoses (n, %)								
ASD	-	-	9 (27%)	5 (21%)	35 (81%)	32 (51%)	8 (14%)	54.8 (<0.001)
ID/GDD	-	-	10 (30%)	21 (88%)	39 (91%)	58 (92%)	1 (2%)	141.3 (<0.001)
Speech/language disorder	-	-	9 (27%)	11 (46%)	32 (74%)	40 (64%)	10 (18%)	44.2 (<0.001)
ADHD	-	-	5 (15%)	1 (4%)	6 (14%)	16 (25%)	26 (46%)	24.0 (<0.001)
ODD/CD	-	-	0 (0%)	1 (4%)	4 (9%)	2 (3%)	4 (7%)	4.4 (0.353)
Anxiety disorder	-	-	7 (21%)	8 (33%)	8 (19%)	10 (16%)	18 (32%)	6.4 (0.174)
Specific learning disorder	-	-	2 (6%)	0 (0%)	1 (2%)	4 (6%)	5 (9%)	3.6 (0.460)
Motor/coordination disorder	-	-	4 (12%)	6 (25%)	24 (56%)	21 (33%)	0 (0%)	45.5 (<0.001)
Depressive disorder	-	-	5 (15%)	0 (0%)	0 (0%)	0 (0%)	10 (18%)	23.8 (<0.001)
Bipolar disorder/mania	-	-	0 (0%)	0 (0%)	0 (0%)	1 (2%)	1 (2%)	1.7 (0.789)
Obsessive compulsive disorder	-	-	0 (0%)	0 (0%)	4 (9%)	2 (3%)	2 (4%)	6.1 (0.192)
Tic disorder	-	-	0 (0%)	0 (0%)	1 (2%)	1 (2%)	1 (2%)	1.2 (0.882)
Feeding/eating disorder	-	-	0 (0%)	0 (0%)	11 (26%)	10 (16%)	0 (0%)	27.5 (<0.001)
Baseline webcam evaluation validity								57.4 (<0.001)
1–3 measures valid (n, %)	0 (0%)	0 (0%)	0 (0%)	1 (4%)	0 (0%)	3 (6%)	0 (0%)	
4–11 measures valid (n, %)	9 (22%)	29 (25%)	7 (21%)	10 (42%)	30 (69.8%)	30 (47%)	13 (23%)	
All measures valid (n, %)	31 (78%)	87 (75%)	26 (79%)	13 (54%)	13 (30.2%)	30 (47%)	43 (77%)	
Number of valid measures (M, SD)	11.3 (1)	11.2 (2)	11.1 (2)	9.9 (3)	10.0 (2)	9.8 (3)	11.3 (2)	6.3 (<0.001)

Note: Diagnoses do not sum to 100% because children could be diagnosed with more than one condition. Note that race/ethnicity categories are not mutually exclusive and participants were encouraged to select all options that apply. For statistical tests with low cell sizes, Fisher's exact test was also computed, but results were highly consistent with the chi-square analysis. For this reason, chi-square is reported with the associated *p*-value.
Abbreviations: ADHD, attention-deficit/hyperactivity disorder; ASD, autism spectrum disorder; ID/GDD, intellectual disability/global developmental delay; ODD/CD, oppositional defiant disorder/conduct disorder.

nonsocial attention, where reliability was lower but still adequate for a low frequency behavior ($\alpha = 0.67$). Test–retest reproducibility estimates were fair or above across 9 of the 12 scales ($r = 0.44$–0.73), with the two measures based on face processing and the nonsocial preference measure showing less stability. Test–retest stability was fair or above for 8 of the 12 measures ($r = 0.40$–0.72), and the highest stability estimates were for receptive vocabulary and single-word reading. Face processing, nonsocial preference, and negative emotional expression scales showed lower stability, the latter of which just missed the cutoff for fair test–retest stability. Similar levels were observed when only NDGS patients were examined.

3.5 | Convergent and discriminant validity

All performance measures, except positive and negative emotional expressiveness, showed strong evidence of convergent and discriminant validity (Table 3). Given the unique nature of gaze-based measures and the difference in measurement modality (gaze vs. informant-report), convergent validity was generally quite good ($r = 0.21$–0.62). Similarly, discriminant validity estimates were generally quite low ($r = 0.07$–0.24). The lack of convergent validity for emotional expressiveness measures is likely because there were no close behavioral constructs assessed by any available informant-report measure.

Intercorrelations among the performance measures tended to be small to moderate (Online Appendix 23), with a few notable exceptions (speed to faces with face preference $r = -0.79$ and receptive vocabulary with reading accuracy $r = 0.78$). The former may suggest redundancy of these measures but the latter correlation is likely due to the close relationship between vocabulary and reading and represents a realistic estimate of the association of these two constructs.

3.6 | Concurrent validity with ID, ASD diagnosis, and autism symptom level

Participants with ID showed statistically significant differentiation across all performance measures (Table 4), including lower levels of

TABLE 2 Valid administration and reliability metrics for webcam-based performance measures.

#	Measure	Stimulus paradigm	Number of indicators	Evaluation validity overall %	% valid no ID	% valid ID	Internal consistency reliability (Cronbach's α)	1-Month test–retest reproducibility (r)	4-Month test–retest stability (r)
1	Overall attention	All	15	100%	100%	100%	0.89	0.52	0.50
2	Attentional scanning	Processing speed	12	87%	89%	84%	0.94	0.66	0.64
3	Positive emotion	Social	32	100%	100%	100%	0.93	0.63	0.62
4	Negative emotion	Social	32	100%	100%	100%	0.95	0.44	0.38
5	Social attention	Social	141	92%	95%	89%	0.89	0.62	0.64
6	Social preference	Social	69	92%	95%	89%	0.75	0.48	0.40
7	Face preference	Social	28	92%	94%	88%	0.90	0.37	0.29
8	Nonsocial preference	Social	42	92%	95%	89%	0.67	0.31	0.31
9	Receptive vocabulary	Receptive vocabulary	39	94%	96%	89%	0.93	0.73	0.72
10	Speed to faces	Social	28	92%	94%	88%	0.93	0.29	0.29
11	Speed to object	Processing speed	12	87%	89%	84%	0.95	0.53	0.51
12	Reading accuracy	Single-word reading	46	96%	99%	91%	0.91	0.68	0.72

Note: Number of indicators refers to the number of areas-of-interest (these could be whole videos or whole stimuli if areas-of-interest are combined) included in computing the measure. Validity proportions are given for baseline data and are estimated by including all individuals who attempted to complete the webcam performance paradigm. Fair test–retest reliability values for overall attention are likely due in part to restricted range as many individuals obtain near 95–100% values. Low test–retest reliability values for negative emotion is likely a function of very limited score range with many individuals falling at 0% expression intensity values.

TABLE 3 Predicted convergent and discriminant validity associations for selected webcam measures.

	Convergent validity		Discriminant validity		
Webcam measure	Measures	Average $\lvert r \rvert$	Measures	Average $\lvert r \rvert$	t (p)
Overall attention	Estimated IQ, ADHD symptoms, executive functioning	0.30	Anxiety, mood, challenging behavior	0.17	2.53 (0.012)
Attentional scanning	Estimated IQ, ADHD symptoms, executive functioning	0.43	Anxiety, mood, challenging behavior	0.23	4.10 (<0.001)
Positive emotion	Mood-hypomania, anxiety	0.10	Motor, daily living skills	0.07	0.47 (0.635)
Negative emotion	Mood-irritability, anxiety	0.09	Motor, daily living skills	0.11	−0.33 (0.742)
Social attention	Autism symptoms	0.55	Anxiety, mood	0.23	6.95 (<0.001)
Social preference	Social communication/interaction Symptoms	0.36	Anxiety, mood	0.16	3.69 (<0.001)
Face preference	Social communication/interaction symptoms	0.26	Anxiety, mood	0.12	2.50 (0.013)
Nonsocial preference	Social communication/interaction symptoms, restricted/repetitive behavior	0.21	Anxiety, mood	0.09	2.27 (0.024)
Receptive vocabulary	Estimated IQ, speech level, social communication/interaction symptoms	0.29	Anxiety, mood, sleep	0.14	2.38 (0.018)
Speed to faces	Social communication/interaction symptoms	0.25	Anxiety, mood, challenging behavior	0.12	2.43 (0.016)
Speed to object	Estimated IQ	0.47	Anxiety, mood, challenging behavior	0.24	3.70 (<0.001)
Reading accuracy	Reading fluency level	0.62	Anxiety, mood, sleep	0.14	8.05 (<0.001)

Note: Convergent and discriminant validity correlations were averaged after conversion to Fisher's z and then reconverted to correlations. Average convergent and discriminant validity correlations were compared using the test of dependent correlations with the nuisance correlation being the average of the intercorrelations between the convergent and discriminant validity measures.

TABLE 4 Descriptive statistics for webcam-collected performance measures across cases with and without ID.

	No ID $n = 224$	ID $n = 151$			
	M (*SD*)	*M* (*SD*)	Raw Δ	*t* (*p*)	Cohen's *d*
Overall attention (%)	82.1 (14)	70.5 (17)	+11.6% (1.8 min total)	7.1 (<0.001)	0.75
Attentional scanning (count)	11.6 (3.4)	7.9 (2.5)	+3.7 glances to each target	9.9 (<0.001)	1.20
Positive emotion (%)	6.4 (8.7)	10.3 (9.1)	−3.9% intensity	−4.2 (<0.001)	−0.44
Negative emotion (%)	2.2 (3.0)	3.4 (4.1)	−1.2% intensity	−3.3 (0.001)	−0.35
Social attention (*z*)	−0.02 (1.0)	−1.52 (1.3)	+1.5 control *SD*s	11.9 (<0.001)	1.14
Social preference (FD)	1.4 (0.3)	1.2 (0.3)	+0.2 seconds per AOI	6.0 (<0.001)	0.68
Face preference (FD)	1.3 (0.8)	0.8 (0.5)	+0.5 seconds per AOI	6.1 (<0.001)	0.70
Nonsocial preference (FD)	1.1 (0.4)	1.2 (0.4)	−0.1 seconds per AOI	−2.1 (0.038)	−0.24
Receptive vocabulary (FD)	41.9 (25.7)	17.1 (13.6)	+24.8 seconds to all targets	10.1 (<0.001)	1.13
Speed to faces (TFF)	7.2 (2.1)	8.0 (1.8)	−0.8 seconds per AOI	−3.3 (<0.001)	−0.37
Speed to object (TFF)	4.9 (1.3)	6.1 (1.2)	−1.2 seconds per AOI	−7.2 (<0.001)	−0.87
Reading accuracy (FD)	37.9 (22.8)	16.6 (14.1)	+21.3 seconds to all targets	9.2 (<0.001)	1.06

Note: ID (defined as parent-report of ID/GDD or estimated IQ < 70). Overall attention (%) is the percentage of time on screen throughout all stimulus paradigms. Count = sum of glances to all targets averaged across stimuli. TFF—values represent averages across all stimuli, including those that were not fixated where the length of the stimulus was imputed. Values for positive and negative emotion represent estimated intensities with a range of 0–100%. Higher values are preferable for all measures except speed to faces and speed to objects where higher values indicate slower time to the AOIs, nonsocial preference where higher values indicate a preference for nonsocial information, and positive and negative emotion measures where higher scores simply indicate more expressiveness. Social attention is presented as a z-score (based on the neurotypical control mean) because this measure is created by averaging multiple different metrics (fixation duration, fixation count, and time-to-first fixation) after standardization.
Abbreviations: AOI, area-of-interest; GDD, global developmental delay; ID, intellectual disability; TFF, time to first fixation.

general attention, attentional scanning, social attention, social preference, face preference, receptive vocabulary, single-word reading, and slower speed to faces and objects. Interestingly, individuals with ID showed high positive and negative emotional expressiveness.

Across subsamples, timepoints, and ages, the social attention measure showed moderate to high correlations ($r = 0.32$–0.62) with autism symptom level (Online Appendix 24). Similarly, concurrent validity with ASD diagnosis consistently fell in the good to excellent range (AUC = 0.69–0.88; Online Appendix 25), with evidence that diagnostic validity is maintained across evaluation timepoints. Dividing the social attention measure into clinically useful score ranges, multilevel likelihood ratios suggest meaningful reductions in ASD probability for low scores ($z \leq 0.1$) and increases in ASD probability for high scores ($z \geq 1.81$). The optimal cut score was 1.49 resulting in 70% sensitivity and 87% specificity (Online Appendix 26).

3.7 | Group profiles across performance measures

Group differences were statistically significant across all performance measures (largest $p = 0.041$; eta-squared = 0.04–0.36). In general, *NFIX*, *SYNGAP1*, and other NDGS showed a more impaired neurobehavioral phenotype, including lower attention, higher nonsocial preference, worse receptive vocabulary and single-word reading, and slower speed to faces and objects (Figure 1). PHTS patients showed lower social attention and social preference and higher nonsocial preference, consistent with high rates of ASD in this group, but only mild reductions in receptive vocabulary and single-word reading and no deficits in overall attention or attentional scanning. Interestingly, *SYNGAP1* and other NDGS patients had higher negative emotional expressiveness scores, while *NFIX* patients and other NDGS showed higher positive emotional expressiveness scores, implying syndrome-specific patterns even among more significantly impaired groups (Online Appendix 27). Taken together, these findings provide preliminary evidence of concurrent (known-groups) validity of performance measures.

4 | DISCUSSION

This research aimed to describe a comprehensive process of creating a set of objective webcam-collected measures, derived using artificial intelligence algorithms for capturing gaze and facial expression information, and based on the gold-standard measurement development guidelines (Boateng et al., 2018) as well as principles of inclusive research practices (FDA, 2009). The process involved both clinicians-scientists and families and was undertaken to provide a preliminary validation of these patient performance measures by examining a range of key psychometric characteristics. Results suggest that these measures a promising new objective evaluation tools that can be useful complements to our recently validated informant-report survey scales (Frazier et al., 2023), permitting multi-method characterization of key social and cognitive characteristics among individuals with NGDS. To our knowledge, the webcam measures and associated survey instruments are the first dedicated set specifically developed to assess a wide range of neurobehavioral and neurodevelopmental

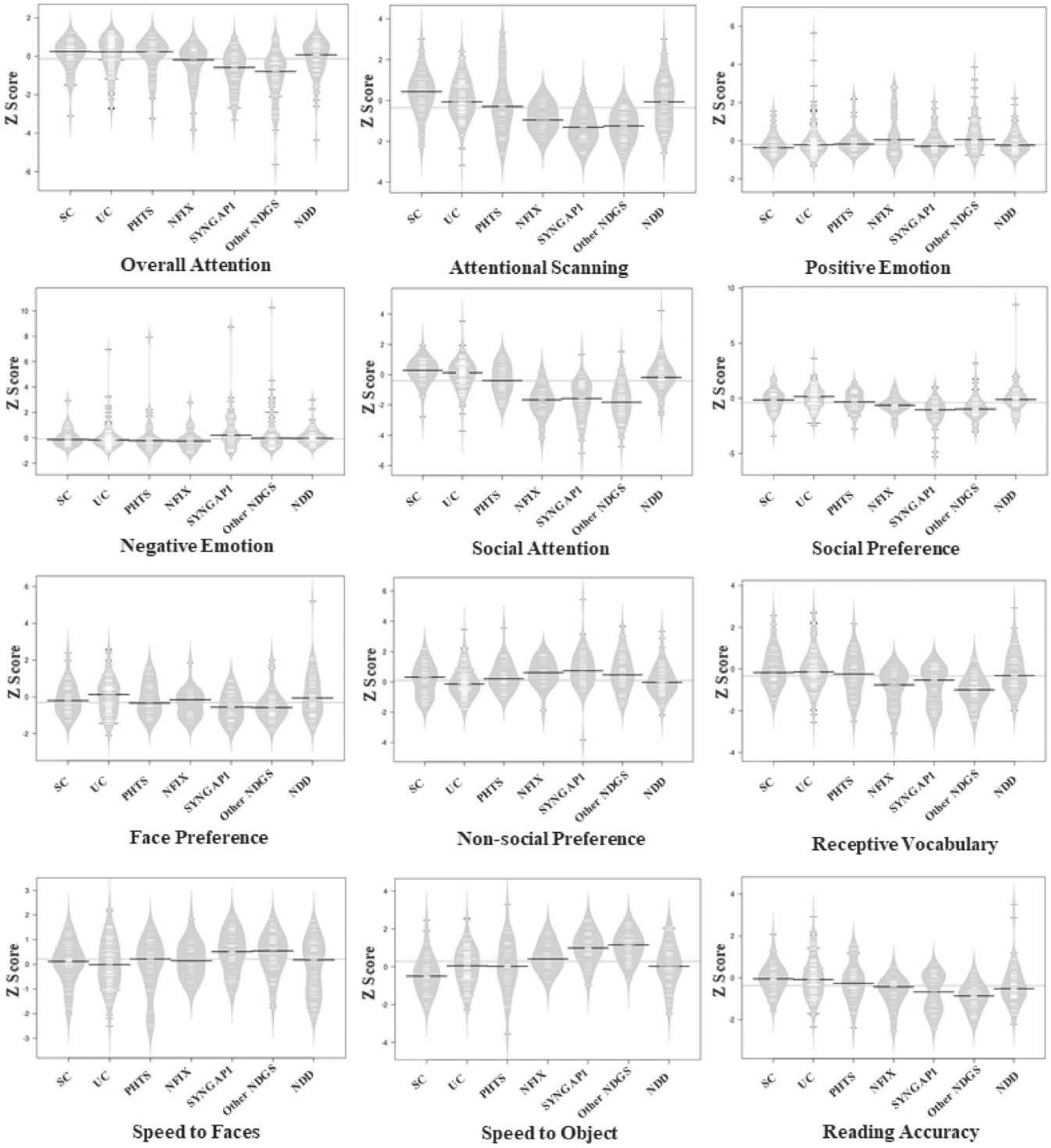

FIGURE 1 NDGS group differences across webcam measures. NDD, idiopathic neurodevelopmental disability; Other NDGS, other neurodevelopmental genetic syndromes; PHTS, PTEN Hamartoma Tumor Syndrome; SC, sibling controls; UC, unrelated controls

presentations seen in NGDS, including individuals with significant cognitive challenges. This initial validation demonstrated that the performance measures are psychometrically sound instruments with potential utility in characterizing the varied clinical and functional spectra seen in many people with NDGS and idiopathic NDD. The validation further highlights the potential value of artificial intelligence/machine learning algorithms for collecting key biometric information that can be used to better understanding individuals with NDGS.

All of the measures showed strong evaluation validity and can be collected in many individuals with mild to moderate cognitive dysfunction. There was a clear gradient of invalid collection in people with more severe cognitive dysfunction, but some individuals

reported to be at the more severe levels could validly complete one or more performance measures. Scale reliability was fair to excellent across all webcam measures, indicating good ability to measure individual differences cross-sectionally across each of the neurobehavioral processes assessed. Test–retest reproducibility and stability were at least acceptable across the majority of measures. Specifically, test–retest reliability was good for attentional scanning, positive emotional expressiveness, social attention, receptive vocabulary, and single-word reading and was fair for sustained attention, social preference, and speed to objects. This indicates that changes in these measures are relatively stable over time, increasing the likelihood that changes reflect real differences in neurobehavioral functioning. Test–retest reliability estimates were lower for negative emotional expressiveness, nonsocial attentional preference, face preference, and speed to faces. When considered in light of adequate or better scale reliability for these measures, the present results suggest these measures may be more state-like in nature. Observations of the score distributions for negative emotional expression and nonsocial preference suggest that lower test–retest reliability for these measures may be influenced by floor effects and, therefore, may be underestimated. Future work is needed to examine score stability over a longer time interval to ensure an adequate balance of stability and sensitivity to change. If sensitivity to change is demonstrated, the quantitative nature, relative brevity, and high evaluation validity of webcam measures might allow for more frequent assessments in the context of intervention studies, thereby increasing statistical power and reducing the sample size needed for clinical trials. This is particularly important for studies of rare NDGS.

Lower test–retest reliability for measures of face processing is intriguing and may be due to factors influencing attention to faces, including the fact that many stimuli included multiple faces as well as other target or background stimuli. It is possible that follow-up evaluations may bias attention toward novel faces (faces not processed as comprehensively in the baseline assessment) or other novel environmental stimuli. It is also possible that face processing is simply more state-like in nature, with reliable collection at each assessment, but rapid changes in quantitative level across hours or days. Future work is needed to tease out these possibilities and examine whether stimulus complexity moderates stability for these measures. Beyond floor effects, lower stability for nonsocial preference is likely, in part, a function of the less frequent nature of attention to socially irrelevant information. It may also be useful for future iterations of the social stimuli to include a larger number of nonsocial or background objects to increase the reliability of this measure. Lower stability for negative emotional expressiveness may be, at least partly, due to the low number of negative facial expressions observed across all participants and is likely influenced by the state-like nature of emotional expressiveness. Adding stimuli that specifically pull for negative emotionality could enhance the test–retest reliability of this measure. Even with these exceptions, all performance measures showed group differences in the baseline data collection, suggesting good known-groups validity and potential value for cross-sectional characterization.

Given their scalability, webcam-collected performance measures also may have utility in clinical contexts for supplementing collection of traditional neurobehavioral measures, allowing more frequent collection between clinical visits, great inclusion in research, and higher quality data via home-based collection. If offered at minimal cost with automated administration, scoring, and reporting functions to reduce clinician burden, these measures could become a key part of ongoing developmental monitoring strategies. This is further supported by the brevity (max 15 min) of administering all four paradigms and the potential to collect only those measures that are relevant to a given patient in future clinical assessments. Future research and collection of large-scale normative data is warranted to determine whether this potential clinical value might be realized and, more importantly, to further evaluate psychometric performance.

Finally, the present results provide preliminary evidence of concurrent (known-groups) validity of webcam measures across NDGS and in comparison to neurotypical controls and idiopathic NDD. The pattern of substantial reductions in many cognitive processes in *NFIX*, *SYNGAP1*, and other NDGS is consistent with our recently published informant-report patterns for many neurobehavioral domains (Frazier et al., 2023). Interestingly, there are some unique patterns among these groups, particularly in the pattern for positive and negative emotional expressiveness, but also in the magnitude of impairments for other domains. For example, people with *SYNGAP1* mutations showed generally worse attention, slower processing speed to faces and objects, and lower social but higher nonsocial preference than people with *NFIX* mutations.

Relative to other NDGS groups, individuals with PHTS tended to show a less impacted social and cognitive profile. Specifically, this group showed no significant impairment in overall attention, attentional scanning, or processing speed measures and only slight reductions in receptive vocabulary and reading accuracy. This is consistent with a spectrum of neurobehavioral dysfunction in PHTS (Busch et al., 2023) and the observation that many individuals have either no or mild reductions in neurocognitive function relative to normative expectation (Busch et al., 2013). Additional data collection in larger NDGS samples will be required to replicate and extend the findings reported here. This work will also need to evaluate the influence of additional clinical factors (e.g., seizures, ID, etc.) on developmental trends.

Several limitations of the current study warrant mention. The genetic syndromes included in this study have a low prevalence and, thus, sample sizes remain modest, particularly given the wide age range. While our power analysis indicated at least adequate power for group comparisons and psychometric analyses were well powered in the full sample, our current data should nevertheless be treated as preliminary, and studies with larger group sample sizes should be completed to replicate our findings and ensure they generalize to the larger population of these NDGS. Given the online nature of the research, it was not feasible to conduct in-person clinical characterization. As a result, this study could not independently confirm the diagnostic status of participants and was not able to administer

dedicated in-person cognitive and behavioral assessments. However, previous studies have demonstrated that parent-report of children's IQ strongly correlates with standardized clinical IQ testing (Shu et al., 2022), and a substantial minority of estimates in this study were based on prior testing (42%). Future work should collect well-validated in-person cognitive assessments to more accurately characterize the sample and examine how webcam measures relate to traditional standardized measures of cognitive and behavioral functioning.

Longitudinal investigations with larger NDGS samples and longer follow-up will also be critical for evaluating age effects and changes in neurobehavioral processes across development, as well as sensitivity to intervention effects. Further, given the preliminary nature of this study, it was not possible to include a comprehensive set of additional instruments to establish convergent and divergent validity. Thus, additional validation work, including convergent and discriminant validity analyses, is needed to provide further support for these webcam measures.

In spite of noted limitations, the present results suggest that webcam-collected gaze and facial expression-based performance measures are promising with evidence that they may function as reliable and valid assessment tools, covering key social and cognitive domains not easily evaluated by informant-report surveys. As such, they may be useful for detailed phenotypic characterization and, ultimately, as reliable, objective, and feasible outcome measures in clinical trials. With additional validation, and sufficient norming, these measures could also facilitate surveillance and clinical assessment for NDGS and idiopathic NDD.

5 | CONCLUSIONS

The present study provides preliminary evidence that webcam-collected performance measures, derived using artificial intelligence algorithms for capturing gaze and facial expression data, can reliably capture individual and between group differences in neurobehavioral function. Future longitudinal investigations with larger NDGS and idiopathic NDD samples will be crucial to further evaluate these measures and determine their potential clinical and research utility.

AUTHOR CONTRIBUTIONS

Thomas W. Frazier, Mirko Uljarević, and Antonio Y. Hardan designed the study. Thomas W. Frazier and Mirko Uljarević collected the data. Thomas W. Frazier and Mirko Uljarević had full access to the data and conducted the analyses. Thomas W. Frazier, Mirko Uljarević, and Antonio Y. Hardan drafted the initial manuscript. All authors critically reviewed and provided the feedback on the initial version of manuscript. All authors approved the final version of the manuscript.

ACKNOWLEDGMENTS

The authors are sincerely indebted to the generosity of the families and individuals who contributed their time and effort to this study. The authors would also like to thank the PTEN Hamartoma Tumor Syndrome Foundation, the PTEN Research Foundation, the SYNGAP Research Fund, the Malan Syndrome Foundation, and the ADNP Kids Foundation for their support of this project. The authors are grateful to all of the families at the participating Simons Searchlight sites as well as the Simons Searchlight Consortium, formerly the Simons VIP Consortium. The authors also appreciate obtaining access to the phenotypic data on SFARI Base. Approved researchers can obtain the Simons Searchlight population dataset described in this study by applying at https://base.sfari.org. CE is the Sondra J. and Stephen R. Hardis Endowed Chair of Cancer Genomic Medicine at the Cleveland Clinic and an ACS Clinical Research Professor. MS is the Rosamund Stone Zander Chair at Boston Children's Hospital.

FUNDING INFORMATION

This study was funded by the PTEN Research Foundation (JCU-20-001) (to Frazier and Uljarević), with additional support from the SYNGAP Research Fund, the Malan Syndrome Foundation, the ADNP Kids Foundation, Autism Speaks (12776), and the Simons Foundation Autism Research Initiative (831500). The content is solely the responsibility of the authors and does not necessarily represent the official views of any of the funding agencies.

CONFLICT OF INTEREST STATEMENT

Dr. Frazier has received funding or research support from, acted as a consultant to, received travel support from, and/or received a speaker's honorarium from the PTEN Research Foundation, SYNGAP Research Fund, Malan Syndrome Foundation, ADNP Kids Research Foundation, Quadrant Biosciences, Autism Speaks, Impel NeuroPharma, F. Hoffmann-La Roche AG Pharmaceuticals, the Cole Family Research Fund, Simons Foundation, Ingalls Foundation, Forest Laboratories, Ecoeos, IntegraGen, Kugona LLC, Shire Development, Bristol-Myers Squibb, National Institutes of Health, and the Brain and Behavior Research Foundation, is employed by and has equity options in Quadrant Biosciences/Autism Analytica, has equity options in MaraBio and Springtide, and has an investor stake in Autism EYES LLC and iSCAN-R. Dr. Kolevzon has received funding or research support from, or acted as a consultant to ADNP Kids Research Foundation, David Lynch Foundation, Klingestein Third Generation Foundation, Ovid Therapeutics, Ritrova Therapeutics, Acadia, Alkermes, Jaguar Therapeutics, GW Pharmaceuticals, Neuren Pharmaceuticals, Scioto Biosciences, and Biogen. Dr. Sahin reports grant support from Novartis, Biogen, Astellas, Aeovian, Bridgebio, and Aucta. He has served on Scientific Advisory Boards for Novartis, Roche, Regenxbio, SpringWorks Therapeutics, Jaguar Therapeutics and Alkermes. Dr. Hardan is a consultant to Beaming Health and IAMA therapeutics. He also has equity options in Quadrant Biosciences/Autism Analytica, and has an investor stake in iSCAN-R. Dr. Shic has acted as a consultant to F. Hoffmann-La Roche AG Pharmaceuticals and Jansen Pharameuticals. The remaining authors have no competing interests to disclose.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Thomas W. Frazier https://orcid.org/0000-0002-6951-2667
Constance Smith-Hicks https://orcid.org/0000-0001-8241-9574
Frederick Shic https://orcid.org/0000-0002-9040-1259

REFERENCES

Amit, M., Chukoskie, L., Skalsky, A. J., Garudadri, H., & Ng, T. N. (2020). Flexible pressure sensors for objective assessment of motor disorders. *Advanced Functional Materials*, *30*(20), 1905241. https://doi.org/10.1002/adfm.201905241

Boateng, G. O., Neilands, T. B., Frongillo, E. A., Melgar-Quinonez, H. R., & Young, S. L. (2018). Best practices for developing and validating scales for health, social, and behavioral research: A primer. *Frontiers in Public Health*, *6*, 149. https://doi.org/10.3389/fpubh.2018.00149

Bove, R., Rowles, W., Zhao, C., Anderson, A., Friedman, S., Langdon, D., Alexander, A., Sacco, S., Henry, R., Gazzaley, A., Feinstein, A., & Anguera, J. A. (2021). A novel in-home digital treatment to improve processing speed in people with multiple sclerosis: A pilot study. *Multiple Sclerosis*, *27*(5), 778–789. https://doi.org/10.1177/1352458520930371

Busch, R. M., Chapin, J. S., Mester, J., Ferguson, L., Haut, J. S., Frazier, T. W., & Eng, C. (2013). Cognitive characteristics of PTEN hamartoma tumor syndromes. *Genetics in Medicine*, *15*(7), 548–553. https://doi.org/10.1038/gim.2013.1

Busch, R. M., Frazier Ii, T. W., Sonneborn, C., Hogue, O., Klaas, P., Srivastava, S., Hardan, A. Y., Martinez-Agosto, J. A., Sahin, M., & Eng, C. (2023). Longitudinal neurobehavioral profiles in children and young adults with PTEN hamartoma tumor syndrome and reliable methods for assessing neurobehavioral change. *Journal of Neurodevelopmental Disorders*, *15*(1), 3. https://doi.org/10.1186/s11689-022-09468-4

Busch, R. M., Srivastava, S., Hogue, O., Frazier, T. W., Klaas, P., Hardan, A., Martinez-Agosto, J. A., Sahin, M., Eng, C., & Developmental Synaptopathies Consortium. (2019). Neurobehavioral phenotype of autism spectrum disorder associated with germline heterozygous mutations in PTEN. *Translational Psychiatry*, *9*(1), 253. https://doi.org/10.1038/s41398-019-0588-1

Chita-Tegmark, M. (2016). Social attention in ASD: A review and meta-analysis of eye-tracking studies. *Research in Developmental Disabilities*, *48*, 79–93. https://doi.org/10.1016/j.ridd.2015.10.011

Ciaccio, C., Saletti, V., D'Arrigo, S., Esposito, S., Alfei, E., Moroni, I., Tonduti, D., Chiapparini, L., Pantaleoni, C., & Milani, D. (2018). Clinical spectrum of PTEN mutation in pediatric patients. A bicenter experience. *European Journal of Medical Genetics*, *62*(12), 103596. https://doi.org/10.1016/j.ejmg.2018.12.001

Cicchetti, D., Bronen, R., Spencer, S., Haut, S., Berg, A., Oliver, P., & Tyrer, P. (2006). Rating scales, scales of measurement, issues of reliability: Resolving some critical issues for clinicians and researchers. *The Journal of Nervous and Mental Disease*, *194*(8), 557–564.

Cohen, J., & Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Dawson, G., Campbell, K., Hashemi, J., Lippmann, S. J., Smith, V., Carpenter, K., Egger, H., Espinosa, S., Vermeer, S., Baker, J., & Sapiro, G. (2018). Atypical postural control can be detected via computer vision analysis in toddlers with autism spectrum disorder. *Scientific Reports*, *8*(1), 17008. https://doi.org/10.1038/s41598-018-35215-8

Egger, H. L., Dawson, G., Hashemi, J., Carpenter, K. L. H., Espinosa, S., Campbell, K., Brotkin, S., Schaich-Borg, J., Qiu, Q., Tepper, M., Baker, J. P., Bloomfield, R. A., Jr., & Sapiro, G. (2018). Automatic emotion and attention analysis of young children at home: A ResearchKit autism feasibility study. *npj Digit Medicine*, *1*, 20. https://doi.org/10.1038/s41746-018-0024-6

FDA. (2009). Patient-reported outcome measures: Use in medical product development to support labeling claims. United States Food and Drug Administration, Guidance for Industry.

Frazier, T. W. (2019). In C. Eng, J. Ngeow, & V. Stambolic (Eds.), *Autism spectrum disorder associated with germline heterozygous PTEN mutations* (Vol. 9, a037002). Cold Spring Harbor Laboratory Press. https://doi.org/10.1101/cshperspect.a037002

Frazier, T. W., Busch, R. M., Klaas, P., Lachlan, K., Jeste, S., Kolevzon, A., Loth, E., Harris, J., Speer, L., Pepper, T., Anthony, K., Graglia, J. M., Delagrammatikas, C., Bedrosian-Sermone, S., Beekhuyzen, J., Smith-Hicks, C., Sahin, M., Eng, C., Hardan, A. Y., & Uljarevic, M. (2023). Development of informant-report neurobehavioral survey scales for PTEN hamartoma tumor syndrome and related neurodevelopmental genetic syndromes. *American Journal of Medical Genetics. Part A*, *191*, 1741–1757. https://doi.org/10.1002/ajmg.a.63195

Frazier, T. W., Embacher, R., Tilot, A. K., Koenig, K., Mester, J., & Eng, C. (2015). Molecular and phenotypic abnormalities in individuals with germline heterozygous PTEN mutations and autism. *Molecular Psychiatry*, *20*(9), 1132–1138. https://doi.org/10.1038/mp.2014.125

Frazier, T. W., Hauschild, K. M., Klingemier, E., Strauss, M. S., Hardan, A. Y., & Youngstrom, E. A. (2020). Rapid eye-tracking evaluation of language in children and adolescents referred for assessment of neurodevelopmental disorders. *Journal of Intellectual & Developmental Disability*, *45*(3), 222–235. https://doi.org/10.3109/13668250.2019.1698287

Frazier, T. W., Klingemier, E. W., Anderson, C. J., Gengoux, G. W., Youngstrom, E. A., & Hardan, A. Y. (2021). A longitudinal study of language trajectories and treatment outcomes of early intensive behavioral intervention for autism. *Journal of Autism and Developmental Disorders*, *51*(12), 4534–4550. https://doi.org/10.1007/s10803-021-04900-5

Frazier, T. W., Klingemier, E. W., Parikh, S., Speer, L., Strauss, M. S., Eng, C., Hardan, A. Y., & Youngstrom, E. A. (2018). Development and validation of objective and quantitative eye tracking-based measures of autism risk and symptom levels. *Journal of the American Academy of Child and Adolescent Psychiatry*, *57*(11), 858–866. https://doi.org/10.1016/j.jaac.2018.06.023

Frazier, T. W., Strauss, M., Klingemier, E. W., Zetzer, E. E., Hardan, A. Y., Eng, C., & Youngstrom, E. A. (2017). A meta-analysis of gaze differences to social and nonsocial information between individuals with and without autism. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(7), 546–555. https://doi.org/10.1016/j.jaac.2017.05.005

Frazier, T. W., Uljarevic, M., Ghazal, I., Klingemier, E. W., Langfus, J., Youngstrom, E. A., Aldosari, M., Al-Shammari, H., El-Hag, S., Tolefat, M., Ali, M., & Al-Shaban, F. A. (2021). Social attention as a cross-cultural transdiagnostic neurodevelopmental risk marker. *Autism Research*, *14*, 1873–1885. https://doi.org/10.1002/aur.2532

Goodwin, M. S., Mazefsky, C. A., Ioannidis, S., Erdogmus, D., & Siegel, M. (2019). Predicting aggression to others in youth with autism using a wearable biosensor. *Autism Research*, *12*(8), 1286–1296. https://doi.org/10.1002/aur.2151

Hardan, A. Y., Jo, B., Frazier, T. W., Klaas, P., Busch, R. M., Dies, K. A., Filip-Dhima, R., Snow, A. V., Eng, C., Hanna, R., Zhang, B., & Sahin, M. (2021). A randomized double-blind controlled trial of everolimus in individuals with PTEN mutations: Study design and statistical considerations. *Contemporary Clinical Trials Communications*, *21*, 100733. https://doi.org/10.1016/j.conctc.2021.100733

IBM Corp. (2021). *IBM SPSS statistics for windows. In (version 28.0)*. IBM Corp.

Kail, R. (1991). Developmental change in speed of processing during childhood and adolescence. *Psychological Bulletin*, *109*(3), 490–501. https://doi.org/10.1037/0033-2909.109.3.490

Kuntzler, T., Hofling, T. T. A., & Alpers, G. W. (2021). Automatic facial expression recognition in standardized and non-standardized emotional expressions. *Frontiers in Psychology*, *12*, 627561. https://doi.org/10.3389/fpsyg.2021.627561

Manfredonia, J., Bangerter, A., Manyakov, N. V., Ness, S., Lewin, D., Skalkin, A., Boice, M., Goodwin, M. S., Dawson, G., Hendren, R., Leventhal, B., Shic, F., & Pandina, G. (2019). Automatic recognition of posed facial expression of emotion in individuals with autism Spectrum disorder. *Journal of Autism and Developmental Disorders*, *49*(1), 279–293. https://doi.org/10.1007/s10803-018-3757-9

McPartland, J. C., Bernier, R. A., Jeste, S. S., Dawson, G., Nelson, C. A., Chawarska, K., Earl, R., Faja, S., Johnson, S. P., Sikich, L., Brandt, C. A., Dziura, J. D., Rozenblit, L., Hellemann, G., Levin, A. R., Murias, M., Naples, A. J., Platt, M. L., Sabatos-DeVito, M., … Autism Biomarkers Consortium for Clinical Trials. (2020). The Autism Biomarkers Consortium for Clinical Trials (ABC-CT): Scientific context, study design, and progress toward biomarker qualification. *Frontiers in Integrative Neuroscience*, *14*, 16. https://doi.org/10.3389/fnint.2020.00016

Mulder, P. A., van Balkom, I. D. C., Landlust, A. M., Priolo, M., Menke, L. A., Acero, I. H., Alkuraya, F. S., Arias, P., Bernardini, L., Bijlsma, E. K., Cole, T., Coubes, C., Dapia, I., Davies, S., Di Donato, N., Elcioglu, N. H., Fahrner, J. A., Foster, A., Gonzalez, N. G., … Hennekam, R. C. (2020). Development, behaviour and sensory processing in Marshall-Smith syndrome and Malan syndrome: Phenotype comparison in two related syndromes. *Journal of Intellectual Disability Research*, *64*(12), 956–969. https://doi.org/10.1111/jir.12787

Nerusil, B., Polec, J., Skunda, J., & Kacur, J. (2021). Eye tracking based dyslexia detection using a holistic approach. *Scientific Reports*, *11*(1), 15687. https://doi.org/10.1038/s41598-021-95275-1

Ness, S. L., Bangerter, A., Manyakov, N. V., Lewin, D., Boice, M., Skalkin, A., Jagannatha, S., Chatterjee, M., Dawson, G., Goodwin, M. S., Hendren, R., Leventhal, B., Shic, F., Frazier, J. A., Janvier, Y., King, B. H., Miller, J. S., Smith, C. J., Tobe, R. H., & Pandina, G. (2019). An observational study with the Janssen Autism Knowledge Engine (JAKE((R))) in individuals with autism spectrum disorder. *Frontiers in Neuroscience*, *13*, 111. https://doi.org/10.3389/fnins.2019.00111

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). McGraw-Hill.

R Core Team. (2021). *R: A language and environment for statistical computing*. R Core Team Retrieved from https://www.R-project.org/

Sahin, M., Jones, S. R., Sweeney, J. A., Berry-Kravis, E., Connors, B. W., Ewen, J. B., Hartman, A. L., Levin, A. R., Potter, W. Z., & Mamounas, L. A. (2018). Discovering translational biomarkers in neurodevelopmental disorders. *Nature Reviews. Drug Discovery*. https://doi.org/10.1038/d41573-018-00010-7

Sahin, M., & Sur, M. (2015). Genes, circuits, and precision therapies for autism and related neurodevelopmental disorders. *Science*, *350*(6263). https://doi.org/10.1126/science.aab3897

Salley, B., & Colombo, J. (2016). Conceptualizing social attention in developmental research. *Social Development*, *25*(4), 687–703. https://doi.org/10.1111/sode.12174

Sasson, N. J., & Elison, J. T. (2012). Eye tracking young children with autism. *Journal of Visual Experiments*, *61*, 3675. https://doi.org/10.3791/3675

Semmelmann, K., & Weigelt, S. (2018). Online webcam-based eye tracking in cognitive science: A first look. *Behavior Research Methods*, *50*(2), 451–465. https://doi.org/10.3758/s13428-017-0913-7

Shehu, I. S., Wang, Y. F., Athuman, A. M., & Fu, X. P. (2021). Remote eye gaze tracking research: A comparative evaluation on past and recent progress. *Electronics*, *10*(24), 3165. https://doi.org/10.3390/electronics10243165

Shu, C., Green Snyder, L., Shen, Y., Chung, W. K., & SPARK Consortium. (2022). Imputing cognitive impairment in SPARK, a large autism cohort. *Autism Research*, *15*(1), 156–170. https://doi.org/10.1002/aur.2622

Simmatis, L., Alavi Naeini, S., Jafari, D., Xie, M. K. Y., Tanchip, C., Taati, N., McKinlay, S., Sran, R., Truong, J., Guarin, D. L., Taati, B., & Yunusova, Y. (2023). Analytical validation of a webcam-based assessment of speech kinematics: Digital biomarker evaluation following the V3 framework. *Digital Biomarkers*, *7*(1), 7–17. https://doi.org/10.1159/000529685

Srivastava, S., Jo, B., Zhang, B., Frazier, T., Gallagher, A. S., Peck, F., Levin, A. R., Mondal, S., Li, Z., Filip-Dhima, R., Geisel, G., Dies, K. A., Diplock, A., Eng, C., Hanna, R., Sahin, M., Hardan, A., & Developmental Synaptopathies Consortium. (2022). A randomized controlled trial of everolimus for neurocognitive symptoms in PTEN hamartoma tumor syndrome. *Human Molecular Genetics*, *31*(20), 3393–3404. https://doi.org/10.1093/hmg/ddac111

Steele, M., Uljarevic, M., Rached, G., Frazier, T. W., Phillips, J. M., Libove, R. A., Busch, R. M., Klaas, P., Martinez-Agosto, J. A., Srivastava, S., Eng, C., Sahin, M., & Hardan, A. Y. (2021). Psychiatric characteristics across individuals with PTEN mutations. *Frontiers in Psychiatry*, *12*, 672070. https://doi.org/10.3389/fpsyt.2021.672070

Streiner, D. L., & Norman, G. R. (1995). *Health measurement scales: A practical guide to their development and use* (2nd ed.). Oxford University Press.

Tuncgenc, B., Pacheco, C., Rochowiak, R., Nicholas, R., Rengarajan, S., Zou, E., Messenger, B., Vidal, R., & Mostofsky, S. H. (2021). Computerized assessment of motor imitation as a scalable method for distinguishing children with autism. *Biological Psychiatry: Cognitive Neuroscience and Neuroimaging*, *6*(3), 321–328. https://doi.org/10.1016/j.bpsc.2020.09.001

Vlaskamp, D. R. M., Shaw, B. J., Burgess, R., Mei, D., Montomoli, M., Xie, H., Myers, C. T., Bennett, M. F., XiangWei, W., Williams, D., Maas, S. M., Brooks, A. S., Mancini, G. M. S., van de Laar, I., van Hagen, J. M., Ware, T. L., Webster, R. I., Malone, S., Berkovic, S. F., … Scheffer, I. E. (2019). SYNGAP1 encephalopathy: A distinctive generalized developmental and epileptic encephalopathy. *Neurology*, *92*(2), e96–e107. https://doi.org/10.1212/WNL.0000000000006729

Youngstrom, E. A., Salcedo, S., Frazier, T. W., & Perez Algorta, G. (2019). Is the finding too good to be true? Moving from "more is better" to thinking in terms of simple predictions and credibility. *Journal of Clinical Child and Adolescent Psychology*, *48*(6), 811–824. https://doi.org/10.1080/15374416.2019.1669158

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Frazier, T. W., Busch, R. M., Klaas, P., Lachlan, K., Jeste, S., Kolevzon, A., Loth, E., Harris, J., Speer, L., Pepper, T., Anthony, K., Graglia, J. M., Delagrammatikas, C. G., Bedrosian-Sermone, S., Smith-Hicks, C., Huba, K., Longyear, R., Green-Snyder, L., Shic, F., … Uljarević, M. (2023). Development of webcam-collected and artificial-intelligence-derived social and cognitive performance measures for neurodevelopmental genetic syndromes. *American Journal of Medical Genetics Part C: Seminars in Medical Genetics*, *193*C:e32058. https://doi.org/10.1002/ajmg.c.32058