Contents lists available at ScienceDirect

Research in Autism

journal homepage: www.elsevier.com/locate/rasd

Maximizing accurate detection of divergence from normative expectation in behavioral intervention outcome assessment

Thomas W. Frazier [a,b,c,*], Katie Huba [a], Allison R. Frazier [d], Rebecca A. Womack [e], Eric A. Youngstrom [f], Lacey Chetcuti [g], Antonio Y. Hardan [g], Mirko Uljarevic [g]

[a] *Department of Psychology, John Carroll University, University Heights, OH 44118, United States*
[b] *Departments of Pediatrics and Psychiatry, SUNY Upstate Medical University, Syracuse, NY 13210, United States*
[c] *Autism Speaks, New York, NY 10016, United States*
[d] *OHElevate, LLC and AI-Measures, LLC, Cleveland, OH 44124, United States*
[e] *RAW Consulting Solutions, Stilwell, OK 74960, United States*
[f] *Institute for Mental and Behavioral Health Research and the Department of Psychiatry, Nationwide Children's Hospital and The Ohio State University, Columbus, OH 43205, United States*
[g] *Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA 94305, United States*

ARTICLE INFO

Keywords:
Norms
Behavioral intervention
Outcomes
Progress monitoring
Generalized additive models

ABSTRACT

Behavioral interventions have shown substantial positive effects at the group level in improving the developmental trajectory of individuals with autism spectrum disorder (ASD), including a wide range of benefits from symptom reductions to skill development. However, there remain pronounced individual differences in the response to interventions and substantial practice variability in the choice and implementation of outcome assessments to evaluate progress for individual cases. Unfortunately, legacy outcome assessments were not specifically designed for the behavioral intervention context or for use with individuals with ASD. Furthermore, legacy instruments have been normed using traditional approaches that are often very inefficient and have limited sensitivity to divergence from neurotypical expectation. Recently, new measures, specifically designed for ASD and related neurodevelopmental conditions, have been developed and revised for use as behavioral intervention outcome assessments. To maximize the value of these measures, the present study aimed to identify optimal norming methods by comparing five distinct continuous norming models. Results indicated that more complex models that include estimation of non-linear age trends fit better and appear to provide more accurate identification of deviation from normative expectation, especially at younger ages where normative data is dense. For some symptom and skill domains, inclusion of sex-specific age-trends was necessary for best fit and most accurate performance. These findings support the use of continuous norming methods using non-linear modeling of developmental trends in the norming of outcome measures for behavioral intervention. Behavior intervention outcome assessments would benefit from implementing these norming approaches to improve the ability to detect deviation from neurotypical symptom and skill levels.

[*] Correspondence to: Department of Psychology, John Carroll University, 1 John Carroll Boulevard, University Heights, OH 44118, United States.
E-mail addresses: tfrazier@jcu.edu (T.W. Frazier), khuba@jcu.edu (K. Huba), allison@ohelevate.com (A.R. Frazier), rebecca@rawconsultingsolutions.com (R.A. Womack), Eric.Youngstrom@nationwidechildrens.org (E.A. Youngstrom), laceyc1@stanford.edu (L. Chetcuti), hardanay@stanford.edu (A.Y. Hardan).

https://doi.org/10.1016/j.reia.2025.202646
Received 3 December 2024; Received in revised form 27 May 2025; Accepted 9 June 2025
Available online 19 June 2025

Introduction

Autism Spectrum Disorder (ASD) is a highly heterogeneous neurodevelopmental condition with two core symptom domains: impairments in social communication and interaction behaviors (SCI) and the presence of restricted / repetitive behaviors (RRB) (American Psychiatric Association, 2013). ASD is also associated with high rates of co-occurring medical and mental health conditions (Al-Beltagi, 2021; Lai et al., 2019; Micai et al., 2023) as well as reductions in individual and family quality of life (Markowitz et al., 2016). High ASD prevalence (Maenner et al., 2023) and appreciation for the substantial functional and quality of life impacts has driven efforts to identify effective interventions that can enhance developmental trajectories for autistic people.

Existing data support the use of both focused and comprehensive behavioral intervention approaches, individually-tailed to the person with ASD (Eckes et al., 2023; Kasari et al., 2014; Kasari et al., 2008; Sandbank et al., 2023). These data have supported increased insurance coverage for behavioral interventions, permitting greater access to early intensive behavioral interventions (Trump & Ayres, 2020). As a result of greater coverage and the impetus to enhance outcomes for individual patients, existing guidelines and many insurance payor policies require baseline and ongoing assessments (typically every six months), including the administration and interpretation of standardized, norm-referenced assessments (Council of Autism Service Providers, 2024). However, there are significant concerns regarding appropriate batteries of outcome assessments for applied behavior analysis practice (National Academies of Sciences, 2024) and significant variability in recommendations, requirements, and implementation exists across payors and providers. This concern and the growing practice requirement highlight the need for measures developed and validated specifically for behavioral intervention outcome assessment of core features and associated behaviors in people with ASD.

At present, behavioral intervention outcome assessments often use a combination of skills-based assessments (Padilla et al., 2023) and norm-referenced instruments (Padilla et al., 2024). The norm-referenced measures included in these assessments (Joseph et al., 2024) are often legacy measures that have a longstanding history of use and proven value in cross-sectional diagnostic assessments, including the Vineland Adaptive Behavior Scales – Third Edition (Vineland-3) (Sparrow et al., 2016), Social Responsiveness Scale – Second Edition (SRS-2) (Constantino & Gruber, 2012), Pervasive Developmental Disorders Behavior Inventory (PDDBI) (Cohen & Sudhalter, 2005), and Behavior Rating Inventor of Executive Function – Second Edition (BRIEF-2) (Gioia et al., 2015). Many of these tools were not specifically developed for ASD or for intervention outcome evaluation and some of these tools use outdated norming methods for comparing client scores to a neurotypical reference population. For example, the SRS-2 and PDDBI use traditional norming methods with limited consideration of age-related changes, where appropriate. Several legacy measures also have limited or incomplete psychometric information or uncertain psychometric properties (Faja et al., 2023; Farmer et al., 2021; Frazier et al., 2014; Greiner de Magalhaes et al., 2022; Lace et al., 2022; McClain et al., 2023; Uljarevic et al., 2020) relevant to use in behavioral intervention outcome assessments. For example, even in measures with a longstanding history of clinical and research application, such as the Vineland-3 (and prior versions), only a handful of prior studies have examined the instrument factor structure to support domain and subdomain scoring (de Bildt et al., 2005; Farmer et al., 2021; Wilkinson et al., 2023), including that no published studies that examined item level structure could be identified by the authors and there is significant disagreement across studies using subdomain scores (Wilkinson et al., 2023). Additionally, few studies have evaluated measurement invariance or differential item or scale functioning across relevant demographic and clinical characteristics (McClain et al., 2023; Sparrow et al., 2016). The result is measurement of broad social and communication domains but no explicit measurement of key social communication subdomains identified in the literature (Frazier & Hardan, 2017; Happe et al., 2017; Simmons et al., 2024).

Furthermore, even when legacy assessments are administered online with automated scoring and reporting, they typically have limited clinician decision support and lack the ability to comprehensively monitor patient functioning over longer intervention periods. Compounding the limitations of existing commonly-used instruments, many providers have been recently credentialed (Behavior Analyst Certification Board, 2024) and have limited training in the application of standardized, norm-referenced assessments to real-world clinical cases. Unfortunately, this leads to administration, scoring, reporting, and interpretation challenges that could impair the ability to implement assessment-driven behavioral intervention. Thus, there is a strong need for a comprehensive assessment platform with measures developed specifically for intervention outcome assessment in ASD.

To address this need, the neurobehavioral evaluation tool (NET) (Frazier, et al., 2023a, 2023b, 2023c, 2023d) was developed with patient, caregiver, and clinician-scientist stakeholders involved at each stage of the measure development process and followed modern measure development recommendations (Boateng et al., 2018; FDA, 2009; PROMIS® Validity Standards Committee on behalf of the PROMIS® Network, 2013). Original NET (NET V1.0) development included both qualitative (concept elicitation and cognitive interviewing) and quantitative (stakeholder ratings of relevance, applicability, and readability) processes. Initial psychometric validation of NET V1.0 showed adequate or better scale and subscale reliability, good test-retest reproducibility and stability, strong conditional reliability of scales across a wide range of the latent trait and good construct validity, including convergent and discriminant validity and concurrent validity with ASD and other mental health diagnoses (Frazier, et al., 2023a, 2023b, 2023c, 2023d). Over the last four years, two successive revisions of the NET V1.0 measure set were undertaken (V2.0 in 2021/2022 and V3.0 in 2024) to further enhance construct coverage relevant to behavioral intervention. Ongoing validation of the revised measures has supported improved reliability and construct coverage of the NET V2.0 and V3.0 measures (Frazier, et al., 2023a, 2023b, 2023c, 2023d; Frazier et al., 2020; Frazier et al., 2022b; Uljarevic et al., 2022; Uljarevic, Spackman et al., 2022).

Having a measure set that is specifically developed for ASD and with an eye toward informing development of, monitor progress in, and supporting clinical management decisions within behavioral intervention is a crucial step forward. But it is also important to ensure that the scores generated from these measures can accurately detect divergence from neurotypical expectations in symptoms and skills/functioning across domains relevant to the child and intervention context. Fig. 1 (panel a) depicts the importance of precise norms for detecting elevated symptom and low skill levels at baseline to inform intervention targeting as well as monitoring change in those scores as they approach and (hopefully) return to the neurotypical range with intervention. Deviations from neurotypical expectation at baseline are particularly useful for identifying domains from which to generate intervention targets for building behavioral intervention plans. By closely monitoring these deviations throughout the intervention process, clinicians can make informed, assessment-driven adjustments, including selecting additional intervention targets as the client shows progress. They can also serve as an indication of when the client may be ready to transition to a lower level of services and/or receive treatment across different service settings. Conversely, when the measurement of these deviations is imprecise, it can hinder the ability to tailor interventions effectively, potentially leading to missed opportunities for improvement and suboptimal outcomes.

Traditionally, normative data for standardized assessments have been collected in very large samples using discrete binning approaches by age, sex, and/or other demographics (e.g., 2–3 year old females) to identify a reference comparison group for each patient (A. Lenhard et al., 2018). Age can be a particularly important factor to include in norming approaches as many cognitive and behavioral processes show significant development, particularly early in life, and even small age effects can dramatically influence an individual's score interpretation relative to normative expectation (Lamsal et al., 2018). However, traditional norming methods can be inefficient, potentially requiring substantial resources and time to collect and deploy and necessitating very small age bin intervals (e.g., intervals of 3 or 6 months) to capture secular or non-linear trends in development of symptoms or skills / functioning. This can be particularly problematic for domains where rapid developmental progress is observed, such as social communication, practical everyday living skills, and executive functioning. Legacy assessments, including those with large overall normative sample sizes, such as the BRIEF-2, are limited in the ability to capture fine-grained developmental trends due to the use of broad age bands (e.g., ages 5–7 years old) or increasingly smaller bin sizes within narrow age bands. Thus, using traditional norming methods that require very large bin sizes within narrow age bands to capture developmental trajectories limits the ability to rapidly create improved test versions that respond to advances in our understanding of ASD and the needs of patients and families. This situation also results in the deployment of sub-optimal assessments with less than ideal ability to detect and monitor symptom and skill deficits.

In response to these limitations, researchers have identified continuous norming methods that have greater statistical power while improving the accuracy of representation of normative distributions (Lenhard & Lenhard, 2021; Zhu & Chen, 2011). Continuous norming methods also have the potential to simultaneously account for more than one demographic factor (e.g. age and sex) without having to substantially increase the size of the normative sample. Further, when continuous norming methods include terms for

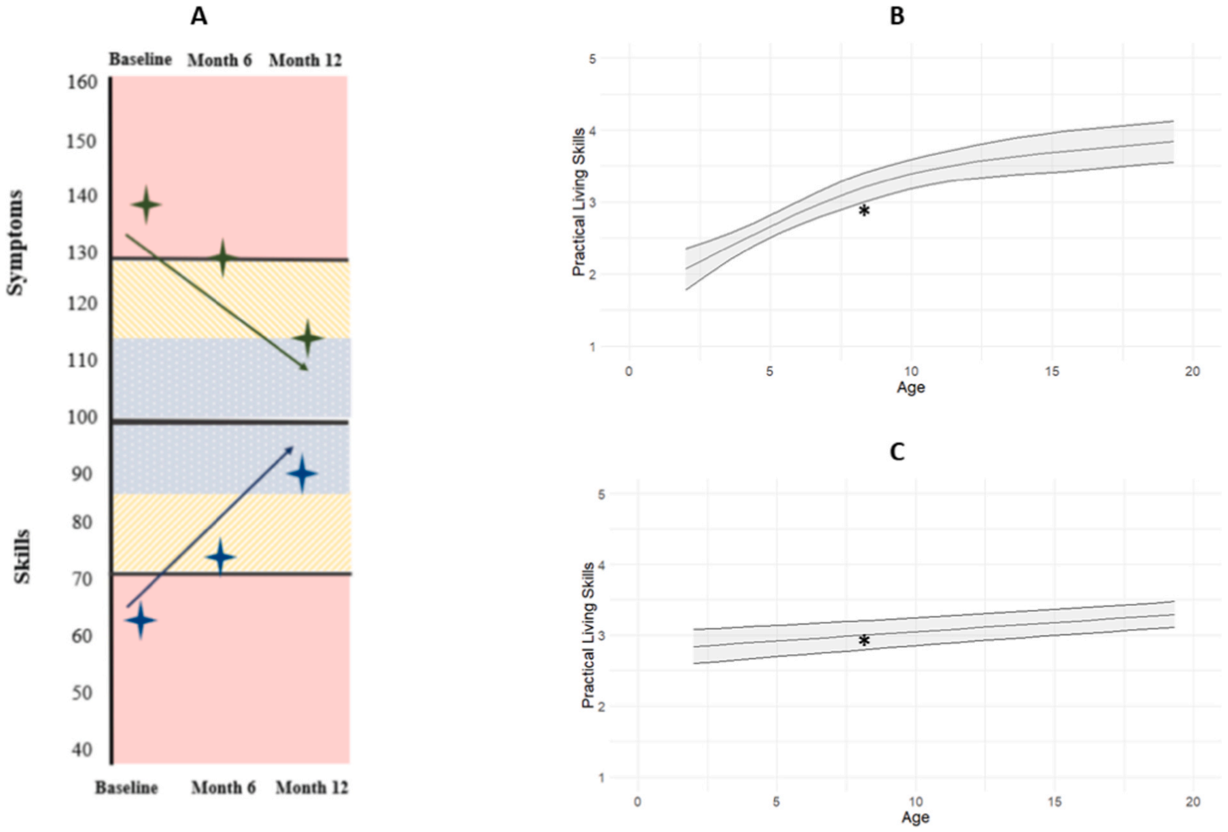

Fig. 1. Using norm-referenced assessments to detect divergence from neurotypical expectation and monitor this divergence during the course of therapy (panel a), with a non-linear model (panel b) providing greater detection accuracy of deviation than a linear model (panel c) for a rapidly developing skill in an 8-year-old child.

assessing non-linear effects of predictors, these methods may be useful for maintaining sensitivity to age or developmental trends (and stratification across other demographic factors in these trends), a particularly important consideration in samples with a wide age span. Fig. 1 (panel b) depicts an example of how, for a skill that develops rapidly in early childhood, non-linear modeling of the age trend identifies a hypothetical 8-year-old client score (raw item average = 2.8) as falling significantly below neurotypical expectation (z = −2.2), whereas a linear fit (panel c) does not accurately capture this divergence (z = −0.7).

There are several different continuous norming methods, including standard linear and polynomial regression models and newer generalized additive models (GAMS) (Sorensen et al., 2023). Linear models are useful if there is a consistent increase or decrease in a neurobehavioral process across age or where other factors have a consistent linear impact on norms. Polynomial models extend linear models to add "bends" in the curve to account for non-linear changes with age or other factors. GAMS are a flexible modeling procedure, with potential to model more nuanced developmental trends, that has previously demonstrated sensitivity to non-linear age trends in autism and other psychopathological symptoms (Uljarevic et al., 2023). Some data suggest that GAMS may be less biased relative to other linear analysis of variance or regression methods (Mundo et al., 2022). However, it remains unclear which methods may be most effective in modeling neurobehavioral data relevant to behavioral intervention outcome assessment for ASD. Ideally the statistical model chosen would provide an accurate picture of the developmental trends being assessed. Understanding the most efficient and precise norming approaches can greatly enhance behavioral intervention practice, providing clinicians the information they need to engage in initial and ongoing treatment planning.

Purpose of the present study

The primary aim of the present study was to compare continuous norming approaches in developing demographically-adjusted normative information and accurately detecting divergence from neurotypical expectation across the 12 neurobehavioral domains assessed by NET V3.0 measures. Continuous norming approaches evaluated included linear regression with age and sex predictors, polynomial regression models with sex and quadratic or cubic age terms, and GAMS with and without interactions between age and sex. We anticipated that continuous norming with GAMS would show better model fit and would better approximate neurotypical distributions in a medium-sized sample of neurotypical individuals recruited across a wide age span (ages 2 to 80). Further, we anticipated that GAMS would more precisely identify deviations from these distributions in ASD and other DD participants relative to other methods as evidenced by a higher proportion of ASD+DD participants falling in the elevated symptoms or low skills/functioning range and higher correlations among impaired classifications.

Materials and methods

Participants

Parent/caregiver informants were recruited using the Prolific online data collection service (https://prolific.co/), following our previous procedures (Frazier, et al., 2023a, 2023b, 2023c, 2023d) and the recruitment strategy was designed to approximate population sampling, with the exception that younger ages were intentionally over-sampled to generate more accurate estimates of early developmental trajectories. Inclusion criteria for the Prolific panel included: informant was a native English speaker, close contact with a child aged 2 to 17 years or adult (ages 18 to 80), and detailed knowledge of the behavior of the person being rated. The goal was to collect normative information that would be broadly applicable to an English-language audience and, thus, inclusion criteria permitted participants from the USA and UK. Electronic informed consent was obtained from all participants and the procedures of this study were reviewed and approved by the local Institutional Review Board. Data were collected from July 18, 2024 to August 15, 2024.

Measures

Demographic and health information

Parent/caregiver informants provided their age, sex, relationship to the participant, and household income. Informants also reported the participant's age, sex, race/ethnicity, estimated cognitive level, level of speech/language, and all prior clinical developmental disorder diagnoses. Informant-reported clinical diagnoses coded into three groups: no clinical diagnosis (neurotypical norm group), autism spectrum disorder diagnosis with or without additional developmental diagnoses (ASD), and other developmental disorder (DD) diagnoses.

Neurobehavioral evaluation tool – version 3.0 (NET V3.0)

The NET V3.0 is the second revision of a set of symptom and skill or functioning measures that were developed specifically for use as outcome measures in future intervention studies for ASD and/or rare neurodevelopmental genetic syndromes (Frazier, et al., 2023a, 2023b, 2023c, 2023d). The original version has demonstrated good psychometric characteristics but weak coverage of some key areas, including social motivation, response inhibition, caregiver quality of life, caregiver coping, set shifting, property destruction, obsessive/compulsive anxiety, and specific fears. As a result, an initial revision (NET V2.0) was undertaken, including stakeholder input regarding relevance, applicability, appropriateness, and readability of newly included items. The resulting scales were only slightly longer but had improved coverage of the above domains and retained excellent psychometric properties (Frazier et al., 2022a; Frazier, et al., 2023a, 2023b, 2023c, 2023d; Uljarevic et al., 2022; Uljarevic et al., 2022). Usage of the NET V2.0 measures in clinical behavioral intervention settings resulted in substantial user feedback. In particular, users noted that additional items were needed to capture social communication/interaction, motor, and practical living skills in younger and more significantly delayed or cognitively-impaired children. Further, several key areas remained under-sampled, including initiation and set shifting, and the need to permit caregivers to identify additional severe challenging behaviors not listed was noted. As a result, an additional revision (NET V3.0) was undertaken using the same stakeholder involvement described above to ensure that additional items and subscales had strong applicability, relevance, appropriateness, and readability. This revised NET V3.0 version had increased total length (V1.0 – 226 total items; V2.0 – 307 total items; V3.0 – 360 total items) relative to prior versions but better coverage of relevant constructs for comprehensive evaluation of intervention benefits and superior psychometric characteristics in several areas (e.g. better content and construct validity, stronger conditional reliability and low and high score levels). NET V3.0 covers 6 symptom domains (anxiety, ADHD, mood, restricted/repetitive behavior, severe challenging behavior, and sleep problems) and 5 skill or functional domains (social communication/interaction, executive functioning, practical living skills, motor skills, and quality of life). Combining the social communication/interaction and restricted/repetitive behavior allows for computing an autism symptom total score, resulting in 12 total scores spanning these domains. Additionally, each domain has 3 to 9 sub-domains, resulting in 55 subscale scores. For all scales and subscales, item average scores were computed with at least 75 % of item data present. Missing data were rare (<0.1 % of all cells). In the present study, median NET V3.0 completion time was 46.0 min (IQR = 24.7 min).

Procedure

Parent/caregiver informants completed the demographic and clinical information questionnaire and NET V3.0 measures using the Qualtrics-XM online platform. Participants were required to complete six attention checks using the Conscientious Responders Scale (Marjanovic et al., 2019). Failure of the first check ("I was born on February 30, 1915") or failure on more than one of the subsequent five checks resulted in removal from the final sample. Additionally, each NET V3.0 measure was checked for consistency in responding (no exclusions) and individuals who did not complete the survey were removed. IRB approval was obtained for all study procedures. Parents/legally-authorized representatives provided informed consent prior to completing any study procedures. Compensation was provided for all participants after study completion (US$10).

Statistical analyses

Sample characterization

Descriptive statistics for demographic and clinical factors were computed to characterize the sample, including neurotypical (NT: n = 536), autism spectrum disorder (ASD: n = 91), and other developmental disability (DD: n = 291) sub-samples. Chi-square or univariate ANOVA or independent samples *t*-test (or non-parametric equivalents) were used to compare demographic and clinical factors across study groups.

Model comparisons

Five distinct norming methods were compared across all 67 NET V3.0 measurements (12 total scores, 55 subscale scores). These models included a standard linear regression model (LM) with least squares fit that included main effects for age and sex, a quadratic polynomial model (QM) that added the quadratic term for age, a cubic polynomial model (CM) that added the quadratic and cubic terms for age, a GAMS model using the default thin plate regression splines (GAMS), and a GAMS model that added the interaction between age and sex (GAMSint). The QM and CM models allow for only 1 or 2 turns, respectively, in the age trend. The GAMS models allow for more complex and nuanced age trends, with the GAMSint model permitting those trends to differ across males and females. Model fit was estimated using the Akaïke Information Criterion (AIC) (Akaike, 1987), generalized cross-validation (GCV) (Reiss & Ogden, 2009), and adjusted R^2, with the lowest AIC and GCV values and the highest R^2 values being indicators of best fit (S. Wood, 2017). Models were fit using a linear link function for all measures except challenging behavior total and subscale scores which showed significant positive skew and kurtosis and where models using a log link function fit consistently better (Supplemental Table 1). To examine the effects of model (LM, QM, CM, GAM, and GAMint) and measure type (symptom or skill/functioning) on model fit indices, separate 5 (model) by 2 (measure) mixed factorial ANOVA models were computed for each model fit index across all 67 sets of analyses (335 total analyses). A significant interaction between model and measure would suggest that the pattern of model fit differs across symptom and skill/functioning measures.

Correlations among norming methods

Each model was used to compute a predicted item average score for each participant. To examine agreement in predicted scores across norming methods, intraclass correlation coefficients (ICCs) for a single indicator (Model 3,1) were computed (Shrout & Fleiss, 1979). These models accounted for rank order and mean differences (absolute agreement) (Youngstrom et al., 2019). ICCs were computed with all five model predictions as indicators (all models), with only the CM and GAMS model as indicators, and with only GAMS and GAMSint as indicators. The CM-GAMS analysis is useful for examining whether the most complex polynomial model is approximating a more complex non-linear model fit. The GAMS-GAMSint analysis examines whether adding the age by sex interaction (estimating age effects separately by sex) produces meaningful differences in predicted scores. Although high ICCs are anticipated for all analyses, ICCs< .90 likely reflect meaningful differences in predicted scores across models.

Differences in classifications across norming methods

To evaluate whether norming methods produce different classifications, z-scores derived from each method were computed by subtracting the model predicted item average score from the observed item average score and dividing by the standard error of the regression estimate. Z-scores were then classified into three categories: below the 10th percentile ($z < -1.28$), within the 10th and 90th percentiles ($z = -1.28$ to $+1.28$), and above the 90th percentile ($z > +1.28$). Differences in classifications were computed across by comparing increasingly complex models (LM vs. QM, QM vs. CM, CM vs. GAMS, GAMS vs. GAMSint). Higher percentage classification differences between model comparisons reflect greater differences in how models fit extreme scores. Lower percentage differences between model comparisons suggest that models are performing similarly in fitting extreme scores.

Divergence from normative expectation in ASD and DD

The proportion of ASD and DD participants diverging from normative expectation (scoring greater than the 90th percentile for symptom measures or less than the 10th percentile for skills/functioning measures) was calculated for all norming methods and across all 12 NET V3.0 total scores. A repeated measures analysis of variance was used to examine whether norming methods produce different average proportions. To further test whether more complex norming models yield more accurate identification of deviation from neurotypical distributions, correlations among domain total scores were computed for all 12 NET V3.0 domains, separately for each norming method. The absolute value of each correlation was Fisher z-transformed prior to averaging correlations within each method. Norming methods were then compared using the test for significance of the difference in dependent correlations.

Statistical power

Using the NT sample size ($n = 536$) and male and female sub-samples within the NT group (n's = 255 and 281, respectively), statistical power to detect a significant R^2 was estimated using a standard linear regression model. Sensitivity power analysis results indicated excellent power ($1-\beta \geq .95$) to detect $R^2 \geq .03$ in the total sample ($\alpha = .05$), assuming up to 3 predictors (age, age^2, and sex). Power was maintained for $R^2 \geq .06$ in sex-specific sub-samples, suggesting good reserve power for subsamples and more complex models. Power to detect differences across norming models, measure type, and their interaction was estimated using a mixed factorial ANOVA model. Sensitivity results indicated excellent power ($1-\beta \geq .95$) to detect medium-sized effects ($f \geq .20$) for the main effects and the interaction ($\alpha = .05$). Statistical power to detect bivariate correlations among z-scores generated across norming methods was at least adequate ($1-\beta \geq .80$) for small correlations ($r \geq .12$; $\alpha = .05$, two-tailed). Power to detect differences in norming method classifications (<10th percentile, 10th to 90th percentiles, and >90th percentile) was excellent even for small-to-medium-sized differences ($w \geq .19$; $\alpha = .05$) in the neurotypical sample. Statistical power to detect differences in the proportion of ASD + DD cases deviating from neurotypical expectation across the five norming methods was at least adequate ($1-\beta \geq .80$) if a medium-to-large effect size ($f = .33$, Cohen's $d = .66$) is observed, even if the correlations between norming methods are moderate ($r = .50$). Data preparation, ANOVA models, ICCs, and bivariate correlations, used SPSS v29 (IBM Corp, 2023). Two-tailed statistical tests were used unless otherwise specified. Linear and polynomial regression models and GAM analyses were computed using the R package *mgcv* (S. Wood, 2017; S. N. Wood, 2003) and implemented in version 4.3.3 (R Core Team, 2024) using *R Studio* version 2024.06.5 build 764. R syntax available at https://osf.io/tc3xf/.

Results

Sample characteristics

Of the 918 participants, the majority reported no clinical developmental neuropsychiatric diagnosis (NT n = 536, 58.4 %), while a substantial minority reported a non-ASD developmental diagnosis (DD n = 291, 31.7 %) or ASD with or without a co-occurring developmental diagnosis (ASD n = 91, 9.9 %). Demographic and clinical characteristics followed patterns from similar prior online data collections (Frazier, et al., 2023a, 2023b, 2023c, 2023d; Frazier et al., 2022b) (Table 1) and the neurotypical subsample approximated US and UK census characteristics (Supplemental Tables 2 and 3). Briefly, the overall sample included a majority biological mothers as informants (55.1 %) and a substantial minority of biological father or other informants, lower household income in ASD and DD groups, higher proportions of males in the ASD and DD groups, relatively balanced race/ethnicity across groups, greater proportions of non-speaking or single word use in the ASD group, and a graded increase in symptoms / decrease in skills from NT to DD to ASD groups. Geographic trends in reported clinical ASD and DD diagnoses suggested lower identification in the UK relative to USA. Importantly, reported racial and ethnic backgrounds were consistent with recent US and UK census findings, a wide range of household income was observed with median values just above expected census values, and US and UK regional participation was also consistent with US and UK census results. Age and sex were equally distributed across US and UK regions. All NET V3.0 measures showed strong measurement invariance when comparing US and UK participants (Supplemental Tables 4 and 5), consistent with prior measurement invariance analyses of these measures across other demographic and clinical factors (Frazier & Uljarevic, 2025).

Model comparisons

For symptom measures, model fit improved significantly from LM to GAMSint (Fig. 2; Supplemental Table 6), with the average variance accounted for being approximately double for CM (.055), GAMS (.060), and GAMSint (.065) relative to LM (.030). Model fit improvements for skills / functioning measures were also significant with more than double the variance accounted for by GAMS (.205) and GAMSint (.206) relative to LM (.091).

Table 1
Demographic and clinical characteristics across neurotypical (NT), autism spectrum disorder (ASD), and other developmental disability (DD).

	NT	ASD	DD	X^2 / *F* / *t* (*p*)
	n (%)	*n (%)*	*n (%)*	
N	536	91	291	
Informant (n, %)				15.7 (.003)
Biological mother	281 (52.4 %)	54 (59.3 %)	167 (57.4 %)	
Biological father	181 (33.8 %)	20 (22.0 %)	66 (22.7 %)	
Other / not reported	74 (13.8 %)	17 (18.7 %)	58 (19.9 %)	
Informant Age in Years (M, SD)	43.2 (12.4)	46.3 (12.1)	49.4 (12.9)	23.0 (<.001)
Highest Informant Education (n, %)				21.7 (.017)
Less than HS	5 (0.9 %)	0 (0.0 %)	6 (2.1 %)	
High school or GED	70 (13.1 %)	16 (17.6 %)	43 (14.8 %)	
Some college	101 (18.8 %)	30 (33.0 %)	70 (24.1 %)	
College graduate	192 (35.8 %)	22 (24.2 %)	91 (31.3 %)	
Graduate degree or higher	167 (31.2 %	22 (24.2 %)	81 (27.8 %)	
Unknown	1 (0.2 %)	1 (1.1 %)	0 (0.0 %)	
Household Income (n, %)				53.7 (<.001)
< $25,000	83 (15.5 %)	40 (44.0 %)	77 (26.5 %)	
$25,000-$34,999	62 (11.6 %)	10 (11.0 %)	27 (9.3 %)	
$35,000-$49,999	66 (12.3 %)	6 (6.6 %)	41 (14.1 %)	
$50,000-$74,999	97 (18.1 %)	13 (14.3 %)	50 (17.2 %)	
$75,000-$99,999	83 (15.5 %)	11 (12.1 %)	33 (11.3 %)	
$100,000-$149,999	78 (14.6 %)	5 (5.5 %)	35 (12.0 %)	
$150,000-$199,999	37 (6.9 %)	3 (3.3 %)	14 (4.8 %)	
$200,000 and above	22 (4.1 %)	1 (1.1 %)	6 (2.1 %)	
Unknown	8 (1.5 %)	2 (2.2 %)	8 (2.7 %)	
Geographic Location				25.0 (.002)
USA – Northeast	61 (11.4 %)	39 (13.4 %)	9 (9.9 %)	
USA – Midwest	82 (15.3 %)	57 (19.6 %)	10 (11.0 %)	
USA – South	138 (25.7 %)	93 (32.0 %)	36 (39.6 %)	
USA – West	91 (17.0 %)	52 (17.9 %)	16 (17.6 %)	
UK	164 (30.6 %)	50 (17.2 %)	20 (22.0 %)	
Participant Age in Years (M, SD)	17.6 (17.9)	20.2 (12.9)	27.0 (19.7)	25.3 (<.001)
Biological Sex (n, % male)	255 (47.6 %)	57 (62.6 %)	154 (52.9 %)	17.5 (.002)
Race				
White / Caucasian (n, %)	386 (72.0 %)	66 (72.5 %)	222 (76.3 %)	1.8 (.405)
Black / African American (n, %)	100 (18.7 %)	12 (13.2 %)	52 (17.9 %)	1.6 (.452)
Middle Eastern (n, %)	4 (0.7 %)	0 (0.0 %)	1 (0.3 %)	1.1 (.572)
East Asian (n, %)	19 (3.5 %)	1 (1.1 %)	5 (1.7 %)	3.4 (.184)
South Asian (n, %)	17 (3.2 %)	3 (3.3 %)	1 (0.3 %)	7.2 (.027)
Pacific Islander (n, %)	1 (0.2 %)	0 (0.0 %)	0 (0.0 %)	0.7 (.700)
Native American (n, %)	3 (0.6 %)	1 (1.1 %)	2 (0.7 %)	0.4 (.837)
Multi-racial (n, %)	41 (7.6 %)	13 (14.3 %	22 (2.4 %)	4.8 (.091)
Chose not to respond (n, %)	7 (1.4 %)	0 (0.0 %)	1 (0.3 %)	2.9 (.233)
Hispanic or Latino (n, %)	42 (7.8 %)	11 (1.2 %)	28 (3.1 %)	4.5 (.347)
Level of Speech Production				69.1 (<.001)
Non-speaking	0 (0.0 %)	7 (7.7 %)	1 (0.3 %)	
Single words	46 (8.6 %)	22 (24.2 %)	33 (11.3 %)	
Fluent speech	490 (91.4 %)	62 (68.1 %)	257 (88.3 %)	
Other NDD Diagnoses (n, %)				
ID/GDD	-	11 (12.1 %)	5 (1.7 %)	18.6 (<.001)
Speech/language disorder	-	11 (12.1 %)	31 (10.7 %)	0.1 (.703)
ADHD	-	34 (37.4 %)	96 (33.0 %)	0.6 (.442)
ODD/CD	-	6 (6.6 %)	7 (2.4 %)	3.7 (.054)
Anxiety disorder	-	34 (37.4 %)	137 (47.1 %)	2.6 (.104)
Specific learning disorder	-	6 (6.6 %)	22 (7.6 %)	0.1 (.757)
Motor / coordination disorder	-	2 (2.2 %)	8 (2.7 %)	0.1 (.774)
Depressive disorder	-	17 (18.7 %)	100 (34.4 %)	8.0 (.005)
Bipolar disorder / mania	-	4 (4.4 %)	19 (6.5 %)	0.6 (.455)
Obsessive compulsive disorder	-	11 (12.1 %)	19 (6.5 %)	3.0 (.085)
Tic disorder	-	1 (1.1 %)	4 (1.4 %)	0.1 (.840)
Feeding / eating disorder	-	2 (0.5 %)	15 (5.2 %)	1.4 (.233)
Other	-	3 (3.3 %)	22 (7.6 %)	2.1 (.151)
Autism symptoms	1.9 (0.5)	3.1 (0.6)	2.2 (0.6)	193.9 (<.001)
Anxiety symptoms	1.9 (0.6)	2.8 (0.8)	2.6 (0.8)	128.1 (<.001)
ADHD symptoms	2.4 (0.7)	3.2 (0.7)	2.8 (0.8)	73.0 (<.001)
Mood symptoms	1.7 (0.6)	2.4 (0.8)	2.3 (0.8)	80.7 (<.001)
Restricted / repetitive behavior	1.9 (0.7)	3.2 (0.8)	2.3 (0.9)	15.1 (<.001)
Severe challenging behavior	1.3 (0.5)	1.8 (0.9	1.4 (0.6)	29.3 (<.001)
Sleep problems	1.6 (0.4)	2.2 (0.7)	2.0 (0.6)	76.5 (<.001)

(*continued on next page*)

Table 1 (*continued*)

	NT	ASD	DD	X^2 / *F* / *t* (*p*)
	n (%)	*n* (%)	*n* (%)	
Social communication / interaction	4.1 (0.6)	2.9 (0.7)	3.8 (0.7)	157.1 (<.001)
Executive functioning and self-regulation	3.7 (0.6)	2.9 (0.7)	3.4 (0.7)	76.8 (<.001)
Practical living skills	3.2 (0.7)	3.0 (0.7)	3.4 (0.6)	18.1 (<.001)
Motor skills	3.5 (0.5)	3.2 (0.6)	3.5 (0.6)	15.1 (<.001)
Quality of life	4.0 (0.4)	3.4 (0.6)	3.7 (0.6)	73.7 (<.001)

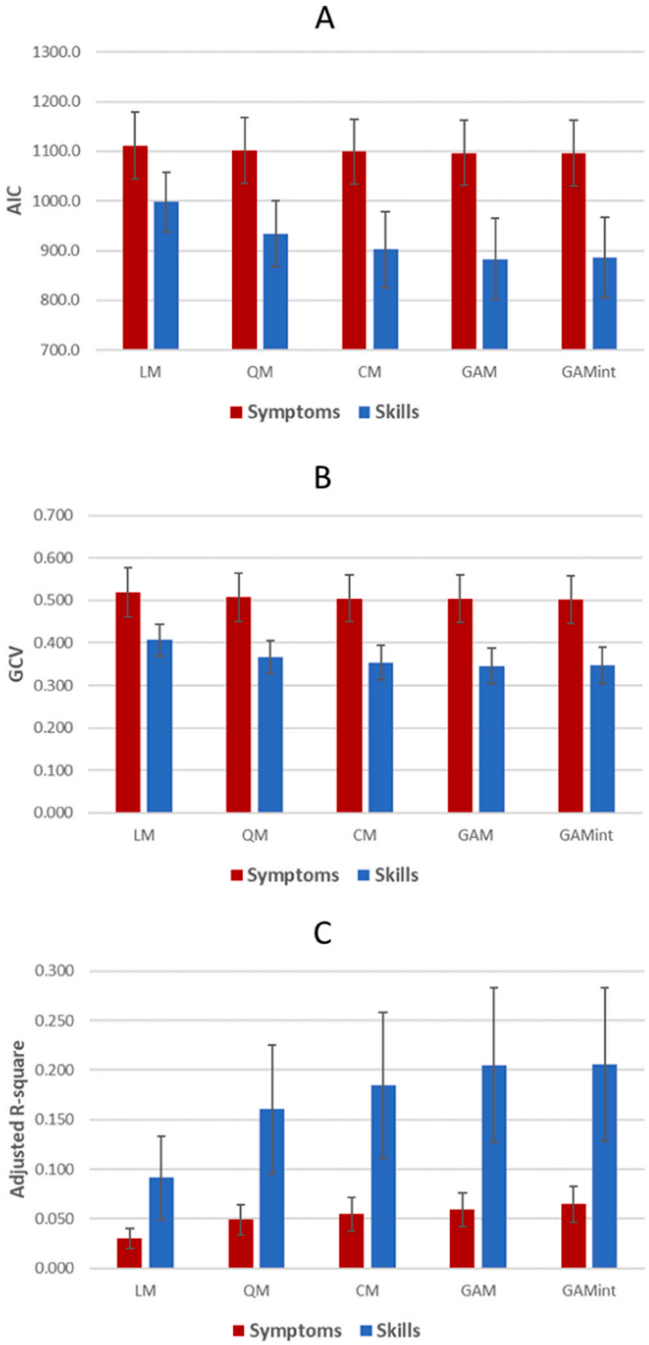

Fig. 2. Average model fit across AIC (panel a), GCV (panel b), and adjusted R^2 (panel c) statistics (+/- 95 % CI), separately for symptom and skill measures. Note. AIC=Akaike Information Criterion, GCV=Generalized Cross-Validation.

Normative patterns for age

Unique non-linear patterns emerge for many symptom (Fig. 3) and skills / functioning measures (Fig. 4) in neurotypical individuals. For example, autism, anxiety, and sleep problems tended to remain relatively stable through childhood and early adulthood while ADHD and mood symptoms showed declines through childhood and adolescence. Not surprisingly, executive functioning, practical living, and motor skills show sharp increases in early childhood with relative stability starting in early adulthood with relative stability (executive functioning and practical living skills) or slight declines (motor skills) thereafter. Social communication / interaction and quality of life remain relatively stable across age. Dramatic shifts in normative curves at older ages may reflect idiosyncratic patterns due to small sample sizes and the influence of outliers rather than true developmental change.

Normative patterns for sex

Several sex-specific developmental trends emerged across symptom and skill / functioning measures in neurotypical participants. Specifically, females show a slightly lower and more stable age pattern for ADHD symptoms but less stability across ages for social communication / interaction skills. Females also are reported to have greater mood difficulties across most of the lifespan and greater sleep difficulties across much of adulthood, relative to males.

Correlations among norming methods

While many ICCs were high (>.700), most sets of indicators suggested less than desirable agreement for normative methods (Table 2). This was particularly true when examining all normative methods, where autism, mood, sleep, motor skills, and quality of life showed lower ICCs, suggesting substantial differences in predicted values across norming methods. ICCs were generally higher when only the QM and GAMS or the GAMS and GAMSint model predictions were used as indicators. However, even in these cases, several measures (ex. mood for GAMS and GAMSint) showed low agreement levels, reinforcing that model choice is important for accurate prediction of normative expectation.

Classification differences across norming methods

Examining substantial divergence (<10th percentile or >90th percentile) from neurotypical expectation across norming models revealed meaningful differences in these classifications (Supplemental Table 7). In particular, differences in classifications were highest when comparing the LM vs. QM model (average classification difference 4.8 %) and remained elevated for the QM vs. CM model (3.2 %), reflecting nearly a quarter and a sixth of classifications being different across models. Classification differences were lower for CM vs. GAM and GAM vs. GAMint, suggesting that models estimating more sophisticated non-linear trends are converging on classifications for low and high scores.

Deviations from normative expectation in ASD+DD

There was a statistically significant and very large overall difference in the proportion of deviant scores (elevated for symptoms, reduced for skills/functioning) across norming methods (F(4, 44)= 6.53, p < .001, partial eta^2 = .372) (Supplemental Table 8). Post-

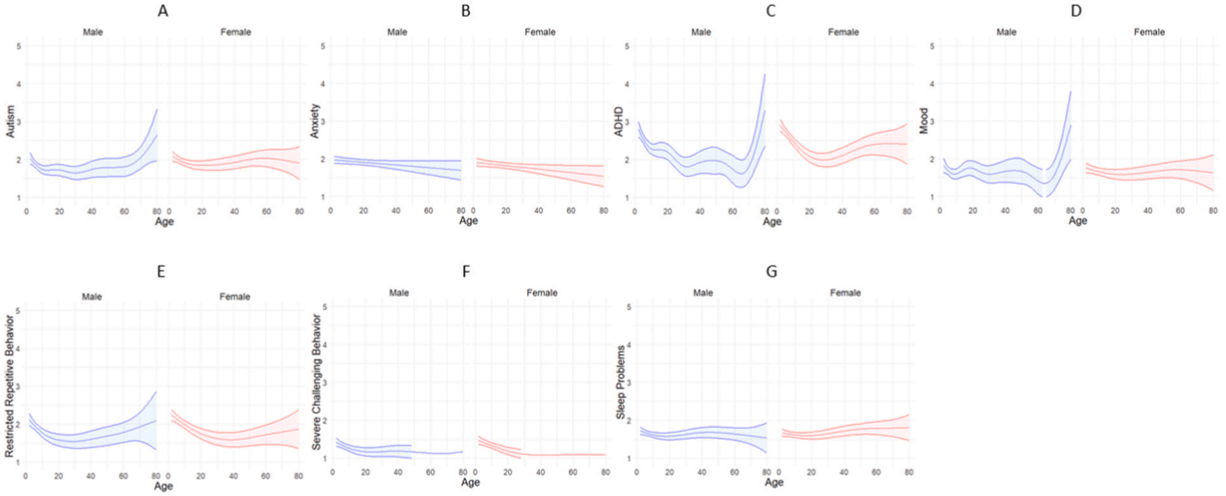

Fig. 3. GAM-estimated item average scores for each symptom measure (A-G) across ages 2 to 80, separately for males and females. Note. A=Autism, B=Anxiety, C=ADHD, D=Mood, E = Restricted / repetitive behavior, F=Severe challenging behavior, and G=Sleep problems.

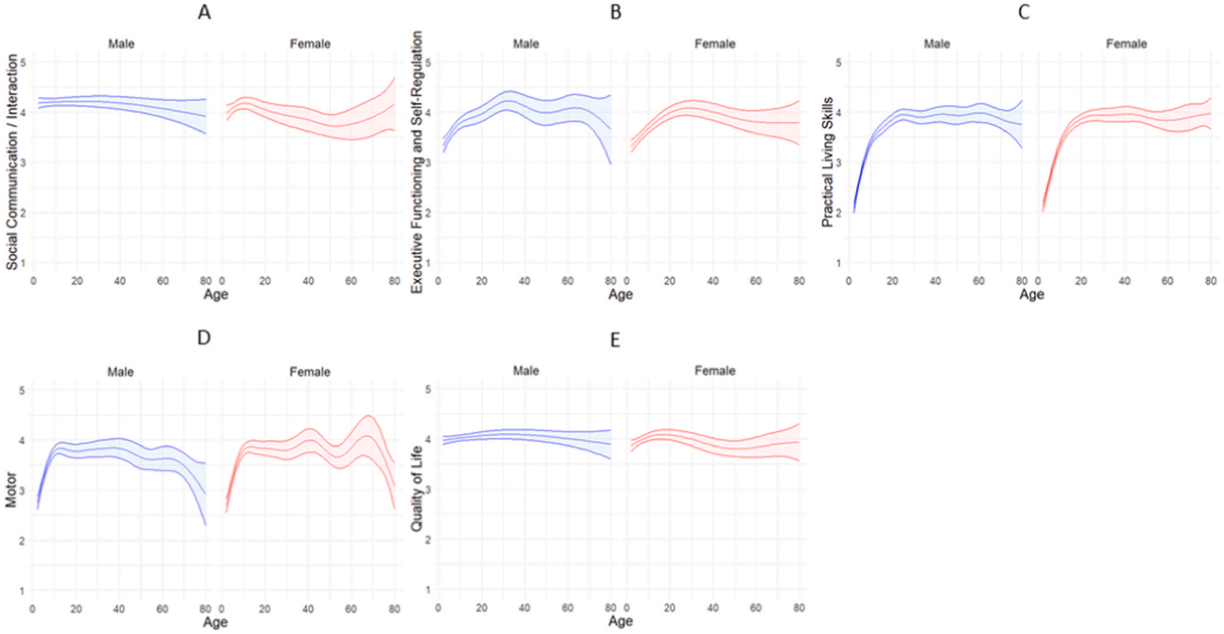

Fig. 4. GAM-estimated item average scores for each skill measure across ages 2 to 80, separately for males and females. Note. A=Social communication / interaction, B=Executive functioning and self-regulation, C= Practical living skills, D=Motor skills, and E = Quality of life.

Table 2
Intraclass correlations (ICCs) among norming model predicted NET V3.0 domain item average scores, separately for all models, cubic and GAMS models, and GAMS with and without age by sex interaction models.

	All Models	Cubic and GAMS	GAMS and GAMSint
	ICC	ICC	ICC
Symptoms			
Autism	.669	.926	.768
Anxiety	.972	.943	.993
ADHD	.758	.912	.897
Mood	.102	.659	.037
Restricted / Repetitive Behavior	.803	.982	.970
Severe Challenging Behavior	.875	.971	.961
Sleep Problems	.392	.698	.782
Skills / Functioning			
Social Communication / Interaction	.807	.892	.825
Executive Functioning and Self-Regulation	.823	.974	.985
Practical Living Skills	.845	.963	.999
Motor Skills	.678	.797	.990
Quality of Life	.563	.736	.789

Note. ICC 95 % confidence intervals were ~ + /−.01 to.03 across analyses.

hoc comparisons indicated that this effect was largely driven by the LM showing a significantly lower proportion identified than all other methods (all $p < .001$) and the QM showing a significantly lower proportion than the CM method ($p = .036$) and a non-significant trend toward a lower proportion than the GAMS method ($p = .096$). There were no significant differences in proportions identified between the CM, GAMS, and GAMSint methods.

There was a small but highly statistically significant difference in magnitude of domain inter-correlations across norming methods ($F(4, 260)= 21.99$, $p < .001$, partial eta^2 = .253). Post-hoc comparisons indicated significant differences between the LM (average $r = .354$) and QM (average $r = .379$) models versus the CM (average $r = .394$), GAMS (average $r = .386$), and GAMSint (average $r = .384$) models with the latter showing significantly larger domain inter-correlations. Interestingly, the cubic polynomial model had the highest average domain inter-correlation, reinforcing that the age effects were not linear.

Discussion

Baseline and periodic behavioral intervention outcome assessments would benefit from having a battery of measures with strong accuracy in detecting divergence from neurotypical expectation. Identifying when a youth's neurodevelopment is following an unusual path requires a precise map of what the typical course looks like. The present results suggest that outcome measures, such as the NET V3.0 informant-report scales, would benefit from using norming methods that provide optimal fit to developmental trends in males and females, particularly for measures where changes with age are substantial. For cognitive or behavioral measures where age effects are less prominent, it is possible that traditional or simpler continuous norming methods are sufficient and potentially easier to collect or implement. Having separate sex norms is commonplace for many scales assessing mental and behavioral symptoms in youths, and also is consistent with established sex differences in prevalence and presentation of ASD (Maenner et al., 2023) and other DDs (Patrick et al., 2020; Zablotsky et al., 2019). Specifically, methods such as the GAMS models evaluated in the present study appear to provide a consistently better fit than linear and quadratic models, although overfitting is possible, particularly in less dense portions of the age distribution. This is consistent with observations of nuances in developmental progress for many cognitive and behavioral processes (Best et al., 2009; Charman, 2006) and suggests that GAMS models that account for non-linear and sex-specific trends may better captures these nuances in at least some domains.

GAMS models also appear to provide a better fit to developmental trends than cubic polynomial models, although the difference is smaller and appears to be most relevant for social communication / interaction, practical living skills, motor skills, ADHD symptoms, mood symptoms, and quality of life measures. While GAMS models show better fit across most ages, it is also clear that overfitting may occur at age ranges with less normative information (Supplemental Figure 2). Further, examination of GAMS model fit with the raw data suggest that the value of GAMS models is most apparent for domains with score ranges where rapid developmental change is observed (ex. ages 2–20 for practical living skills). As with all measures and norming approaches, interpretation of domains that show significant floor or ceiling effects or clearly non-normal distributions (e.g., challenging behavior) at low or high score levels or ages where the potential ceiling and floor are most likely (e.g., practical living skills after age 30). At minimum, the present results suggest that the simple linear and quadratic models used most widely in current legacy assessments implemented in research and clinical practice are unlikely to be sufficient, and that newly-developed instruments should re-examine and determine the optimal norming approach before deployment in behavioral intervention outcome assessments.

The observation of equivalent or better performance for cubic polynomial models relative to GAMS models suggests that these models may be sufficient at modeling developmental trends for a subset of neurobehavioral domains (particularly symptom domains). However, the fact that variance accounted for (adjusted R^2) nearly always favored GAMS models supports these as a default approach to norming behavioral intervention outcome measures. This position is further supported by the flexibility of GAMS models where non-linear link functions can be modeled for measure scores that diverge from normal distributions and the fact that the models can be tuned to the optimal level of complexity. A good example of the latter feature in the present study was observed for the anxiety total score. For this measure, the most complex model (GAMSint) defaulted to a linear fit that was consistent across males and females, functionally equivalent to a standard linear regression model (LM). Thus, when the complexity is not needed, GAMS models can be estimated with simpler structure. The map they provide is no more complicated than it needs to be to chart the typical trajectory.

The choice between GAMS models with and without an age by sex interaction is likely to depend on whether significant differences in the pattern of developmental trends are anticipated in males and females. For some domains, such as mood, there is substantial a priori justification for modeling the age by sex interaction. In other cases, it may be useful to estimate both types of models and evaluate fit and sex differences in the pattern. The present results suggest that estimating sex-specific developmental patterns is likely to be most fruitful for autism, ADHD, anxiety, mood, and sleep problems as well as motor skills and possibly some aspects of social communication/interaction skills. Ultimately, if no strong a priori justification exists, it may still be useful to explore sex-specific patterns, particularly in early developmental normative data.

Unfortunately, although present data strongly suggest that normative adjustment only for age or sex is likely untenable and will result in misestimation of deviation from neurotypical expectation for at least some of the assessed constructs assessed, most legacy assessments used as behavioral intervention outcome measures utilize this approach. For instance, Vineland-3 and Behavior Rating Inventory of Executive Function, Second Edition include normative adjustment only for age (e.g., and Social Responsiveness Scale – Second Edition include normative adjustment only for sex and some measures don't adjust for any demographic information (e.g., Social Communication Questionnaire). Future revisions of these instruments would benefit from simultaneous adjustment for age and sex as well as potential inclusion of the age by sex interaction to maximize their utility as behavioral intervention outcome measures. Without these modifications, the norms are likely to misclassify people that actually are showing performance within the neurotypical range. For longitudinal monitoring applications, behavioral intervention outcome assessment might also benefit from adjustment for cognitive and language abilities to ensure that appropriate expected change scores are generated for individuals with more significantly impacted phenotypes.

The modest but highly statistically significant increase in the proportion of ASD+DD participants identified by more complex models (CM, GAMS, and GAMSint) coupled with higher domain inter-correlations in these models, supports the need for non-linear modeling of age trends in normative neurobehavioral data. These results suggest that, without careful modeling of developmental trends, a non-trivial proportion of ASD cases seen in behavioral intervention practice will be mis-identified as having scores deviating from neurotypical expectation (false positives) as well as ASD cases that are not identified as deviating but in fact are (false negatives). Thus, even though these effects are modest in the present data, the potential impact is substantial at the case level and highlights how legacy and new measurement tools must pay careful attention to accurately modeling the neurotypical score distribution. This need becomes even more clear when considering applications in intervention monitoring, where knowing when an individual who deviated from expectation at baseline is now below a cutoff derived from the normative distribution (within the 90th or 95th percentile). In these cases, inaccurate modeling can lead to meaningful differences in clinical management, including discontinuing therapy, reducing intervention target emphasis, or switching domain focus.

Limitations and future directions

There are five major limitations to the present study. First, the neurotypical group sample size was not sufficient, given the wide age range, to compare continuous norming to traditional binning. However, it is important to emphasize that the neurotypical group size obtained in the present study produced relatively small prediction intervals, particularly at young ages. This highlights one of the crucial advantages of continuous over the traditional binning approach - that good precision can be achieved in medium-sized normative samples. It is important that future research extend the present findings and conduct this type of analysis to understand how legacy measures, which frequently use a binning approach, would perform relative to GAMS models. It is possible that traditional norming with very large sample sizes that permit narrow age bins will outperform GAMS-based continuous norming in moderate samples sizes. However, even if this is observed, the magnitude of improvement may be offset by the technical and financial challenges and inefficiencies in very large sample normative collection, impeding rapid measure revision and improvement. The second limitation is that the NET V3.0 measures rely on informant-report and thus are susceptible to rater biases. Although this limitation is not specific to NET and extends across the legacy instruments (e.g., VABS-3, BRIEF, SRS-2, SCQ), it would be useful to re-examine the present findings for objective measures of cognition and behavior. We anticipate being able to conduct these analyses in the near future as normative data for the companion webcam-collected performance measures accrues (Frazier, et al., 2023a, 2023b, 2023c, 2023d). Third, the ASD and DD samples were formed based on caregiver-report of clinical diagnoses and were not independently confirmed. Prior research has indicated that online collection of ASD cases shows high correspondence with clinical diagnoses (Daniels et al., 2012; Feliciano et al., 2019), empirically-identified latent classes (Frazier et al., 2023a, 2023b, 2023c, 2023d; Frazier et al., 2012), and gold-standard assessment tools (Lee et al., 2010). However, it is likely, given the online nature of survey collection, that the ASD and DD groups are not representative of the fuller population. This particularly likely for the ASD group which includes a disproportionate number of individuals with fewer cognitive challenges – a common sampling problem in online studies of ASD. Future work is needed with broad capture of the ASD and DD population and confirmation of clinical diagnoses and with careful recruitment of the ASD and DD populations. Fourth, confidence intervals around normative values broaden after age 50 and dramatic changes in the curve are due to small numbers of cases indicating the need for additional data collection in these age ranges to develop more precise norms. Lastly, the normative population examined in the present sample focused on neurotypical participants only, excluding ASD and DD. While neurotypical norms can provide a useful comparison for clinicians who are monitoring neurobehavioral functions and want to identify when neurobehavioral functions that previously deviated from neurotypical levels return back to the neurotypical range. This approach can be helpful for informing clinical management decisions and intervention planning. However, future work should also examine whether the present results extend to normative samples that include ASD and other DD participants at population expectation levels.

As practice patterns evolve and measure development progresses, it will be useful for behavioral intervention outcome assessment guidelines and payor policy to reflect advances in all aspects of measure development, including application of modern continuous norming procedures. Measures that have accurately modeled the normative distribution are more likely to be useful in guiding clinical practice, including identifying deviations from normative expectation in a diagnostic context, facilitating greater clinician sensitivity to neurobehavioral challenges and potentially facilitating more accurate diagnostic judgments. Further, more sensitive and precise norming methods can assist clinicians in developing intervention strategy and target lists at baseline and revising the strategy and clinical management approach as progress is observed in an intervention context. Considering these advances in the development, selection, and use of behavioral intervention outcome batteries will be critical to further cementing the utility of focused and comprehensive behavioral interventions for ASD.

Conclusions

Continuous norming methods that include non-linear modeling of developmental trends appear to provide a more accurate approach to assessing patient scores on neurobehavioral domains relevant to behavioral intervention outcome assessment. Legacy and newly-developed instruments intended for use in the behavioral intervention context should use these methods to ensure that divergence from neurotypical expectation are precisely measured and monitored as intervention progresses. Deploying instruments that utilize optimal norming methods in clinical practice will be critical to further establishing the value of behavioral intervention to payors and the broader set of stakeholders.

CRediT authorship contribution statement

Frazier Thomas W: Writing – review & editing, Writing – original draft, Validation, Methodology, Investigation, Formal analysis, Conceptualization. **Hardan Antonio:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Lacey Chetcuti:** Writing – review & editing, Methodology, Conceptualization. **Mirko Uljarevic:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Formal analysis, Conceptualization. **Allison R. Frazier:** Writing – review & editing, Methodology, Conceptualization. **Katie Huba:** Writing – review & editing, Project administration, Methodology, Data curation. **Youngstrom Eric:** Writing – review & editing, Methodology, Formal analysis, Conceptualization. **Rebecca A. Womack:** Writing – review & editing, Methodology, Conceptualization.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Declaration of Competing Interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Thomas W. Frazier reports a relationship with Simons Foundation that includes: Funding grants. Thomas W. Frazier, PhD reports a relationship with MaraBio that includes: consulting or advisory and equity or stocks. Thomas W. Frazier, PhD reports a relationship with Linus Biotechnology Inc that includes: consulting or advisory. Thomas W. Frazier, PhD reports a relationship with Autism EYES that includes: equity or stocks. Thomas W. Frazier, PhD reports a relationship with I-SCAN-R that includes: equity or stocks. Thomas W. Frazier, PhD reports a relationship with AI-Measures that includes: equity or stocks. Eric A. Youngstrom, PhD reports a relationship with Helping Give Away Psychological Science that includes: employment. Eric A. Youngstrom reports a relationship with American Psychological Association that includes: consulting or advisory and equity or stocks. Eric A. Youngstrom, PhD reports a relationship with Signant Health that includes: consulting or advisory. Eric A. Youngstrom reports a relationship with Guilford Press that includes: equity or stocks. Eric A. Youngstrom reports a relationship with Joe Startup Technologies that includes: equity or stocks. Eric A. Youngstrom reports a relationship with AI-Measures that includes: equity or stocks. Antonio Y. Hardan, MD reports a relationship with PTEN Research that includes: funding grants. Antonio Y. Hardan, MD reports a relationship with Quadrant Biosciences Inc that includes: consulting or advisory and equity or stocks. Thomas W. Frazier, PhD reports a relationship with Quadrant Biosciences Inc that includes: employment and equity or stocks. Antonio Y. Hardan, MD reports a relationship with Autism Speaks that includes: funding grants. Antonio Y. Hardan, MD reports a relationship with Beaming Health that includes: equity or stocks. Antonio Y. Hardan, MD reports a relationship with F Hoffmann-La Roche Ltd that includes: consulting or advisory. Antonio Y. Hardan, MD reports a relationship with National Institutes of Health that includes: funding grants. Thomas W. Frazier, PhD reports a relationship with National Institutes of Health that includes: funding grants. Antonio Y. Hardan, MD reports a relationship with I-SCAN-R that includes: equity or stocks. Antonio Y. Hardan, MD reports a relationship with AI-Measures that includes: equity or stocks. Antonio Y. Hardan, MD reports a relationship with ParenteAI that includes: equity or stocks. Mirko Uljarevic, MD, PhD reports a relationship with I-SCAN-R that includes: equity or stocks. Mirko Uljarevic, MD, PhD reports a relationship with AI-Measures that includes: funding grants. Allison R. Frazier reports a relationship with AI-Measures that includes: equity or stocks. Allison R. Frazier reports a relationship with AI-Measures that includes: equity or stocks. Rebecca Womack reports a relationship with AI-Measures that includes: equity or stocks. None If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

The authors would like to thank the parents/caregivers participating in this study.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.reia.2025.202646.

Data availability

Data will be made available on request.

References

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317–332.

Al-Beltagi, M. (2021). Autism medical comorbidities. *World Journal of Clinical Pediatrics, 10*, 15–28.

American Psychiatric Association. (2013). *Diagnostic and Statistical Manual of Mental Disorders (Fifth ed.)*. Arlington, VA: American Psychiatric Association,.

Behavior Analyst Certification Board. (2024). BACB certificant data. In.

Best, J. R., Miller, P. H., & Jones, L. L. (2009). Executive functions after age 5: Changes and correlates. *Developmental Review, 29*, 180–200.

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. *Front Public Health, 6*, 149.

Charman, T., & Stone, W. L. (2006). *Social and communication development in autism spectrum disorders: early identification, diagnosis, and intervention*. Guilford Press,.

Cohen, I., & Sudhalter, V. (2005). *PDDBI behavior inventory: professional manual*. Lutz, FL: Psychological Assessment Resources, Inc,.

Constantino, J. N., & Gruber, C. P. (2012). *The social responsiveness scale manual, second edition (SRS-2)*. Los Angeles, CA: Western Psychological Services,.

Council of Autism Service Providers. (2024). Applied Behavior Analysis Practice Guidelines for the Treatment of Autism Spectrum Disorder: Guidelines for Healthcare Funders, Regulatory Bodies, Service Providers, and Consumers. In (Third Edition ed.).

Daniels, A. M., Rosenberg, R. E., Anderson, C., Law, J. K., Marvin, A. R., & Law, P. A. (2012). Verification of parent-report of child autism spectrum disorder diagnosis to a web-based autism registry. *Journal of Autism and Developmental Disorders, 42*, 257–265.

de Bildt, A., Kraijer, D., Sytema, S., & Minderaa, R. (2005). The psychometric properties of the Vineland Adaptive Behavior Scales in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders, 35*, 53–62.

Eckes, T., Buhlmann, U., Holling, H. D., & Mollmann, A. (2023). Comprehensive ABA-based interventions in the treatment of children with autism spectrum disorder - A meta-analysis. *BMC Psychiatry, 23*, 133.

Faja, S., Sabatos-DeVito, M., Sridhar, A., Kuhn, J. L., Nikolaeva, J. I., Sugar, C. A., Webb, S. J., Bernier, R. A., Sikich, L., Hellemann, G., Senturk, D., Naples, A. J., Shic, F., Levin, A. R., Seow, H. A., Dziura, J. D., Jeste, S. S., Chawarska, K., Nelson, C. A., 3rd, … Autism Biomarkers Consortium for Clinical, T. (2023). Evaluation of clinical assessments of social abilities for use in autism clinical trials by the autism biomarkers consortium for clinical trials. *Autism Research, 16*, 981–996.

Farmer, R. L., Floyd, R. G., & McNicholas, P. J. (2021). Is the Vineland-3 comprehensive interview form a multidimensional or unidimensional scale? Structural analysis of subdomain scores across early childhood to adulthood. *Assessment, 28*, 1848–1864.

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

Feliciano, P., Zhou, X., Astrovskaya, I., Turner, T. N., Wang, T., Brueggeman, L., Barnard, R., Hsieh, A., Snyder, L. G., Muzny, D. M., Sabo, A., Consortium, S., Gibbs, R. A., Eichler, E. E., O'Roak, B. J., Michaelson, J. J., Volfovsky, N., Shen, Y., & Chung, W. K. (2019). Exome sequencing of 457 autism families recruited online provides evidence for autism risk genes. *NPJ Genomic Medicine, 4*, 19.

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. (2023a). 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.

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., … Uljarevic, M. (2023b). Development of webcam-collected and artificial-intelligence-derived social and cognitive performance measures for neurodevelopmental genetic syndromes. *American Journal of Medical Genetics Part C, 193*, Article e32058.

Frazier, T. W., Chetcuti, L., Al-Shaban, F. A., Haslam, N., Ghazal, I., Klingemier, E. W., Aldosari, M., Whitehouse, A. J. O., Youngstrom, E. A., Hardan, A. Y., & Uljarević, M. (2023c). Categorical versus dimensional structure of autism spectrum disorder: A multi-method investigation. *JCPP Advances, 3*, Article e12142.

Frazier, T. W., Crowley, E., Shih, A., Vasudevan, V., Karpur, A., Uljarevic, M., & Cai, R. Y. (2022a). Associations between executive functioning, challenging behavior, and quality of life in children and adolescents with and without neurodevelopmental conditions. *Frontiers in Psychology*.

Frazier, T. W., Dimitropoulos, A., Abbeduto, L., Armstrong-Brine, M., Kralovic, S., Shih, A., Hardan, A. Y., Youngstrom, E. A., Uljarevic, M., & Quadrant Biosciences - As You Are Team. (2023d). The autism symptom dimensions questionnaire: Development and psychometric evaluation of a new, open-source measure of autism symptomatology. *Developmental Medicine and Child Neurology*.

Frazier, T. W., & Hardan, A. Y. (2017). Equivalence of symptom dimensions in females and males with autism. *Autism, 21*, 749–759.

Frazier, T. W., Hyland, A. C., Markowitz, L. A., Speer, L. L., & Diekroger, E. A. (2020). Psychometric evaluation of the revised child and family quality of life questionnaire (CFQL-2). *Research in Autism Spectrum Disorders, 70*.

Frazier, T. W., Khaliq, I., Scullin, K., Uljarevic, M., Shih, A., & Karpur, A. (2022b). Development and psychometric evaluation of the open-source challenging behavior scale. *Journal of Autism and Developmental Disabilities*.

Frazier, T. W., Ratliff, K. R., Gruber, C., Zhang, Y., Law, P. A., & Constantino, J. N. (2014). Confirmatory factor analytic structure and measurement invariance of quantitative autistic traits measured by the Social Responsiveness Scale-2. *Autism, 18*, 31–44.

Frazier, T.W., & Uljarevic, M. (2025). Neurobehavioral Evaluation Tool - Version 3.0 Manual. In: AI.Measures.

Frazier, T. W., Youngstrom, E. A., Speer, L., Embacher, R., Law, P., Constantino, J., Findling, R. L., Hardan, A. Y., & Eng, C. (2012). Validation of proposed DSM-5 criteria for autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*, 28–40. e23.

Gioia, G. A., Isquith, P. K., Guy, S. C., & Kenworthy, L. (2015). *Behavior rating inventory of executive function–second edition (BRIEF-2)*. Lutz, FL: Psychological Assessment Resources,.

Greiner de Magalhaes, C., Pitts, C. H., & Mervis, C. B. (2022). Executive function as measured by the Behavior Rating Inventory of Executive Function-2: Children and adolescents with Williams syndrome. *Journal of Intellectual Disability Research, 66*, 94–107.

Happe, F., Cook, J., & Bird, G. (2017). The structure of social cognition: In(ter)dependence of sociocognitive processes. *Annual Review of Psychology, 68*, 243–267.

IBM Corp. (2023). *IBM SPSS Statistics for Windows. In (29.0 ed.)*. Armonk, NY: IBM Corp,.

Joseph, A., Chong, I., Das-Gupta, Z., Bandeira de Lima, C., Dixon, D., Dovbnya, S., Fittro, E., Gerhardt, P., Huang, W., Josephson, B., Li, D., Martin, N., Mukerji, S., Rodriguez, K., Rue, H., Strunk, K., Tarbox, J., Vadgama, Y., Valentino, A., … Willis, S. (2024). Development of a standardized set of outcomes for autism spectrum disorder: The International Consortium for Health Outcomes Measurement (ICHOM). *Research in Autism Spectrum Disroders, 117*, Article 102451.

Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., Murphy, S., & Almirall, D. (2014). Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child and Adolescent Psychiatry, 53*, 635–646.

Kasari, C., Paparella, T., Freeman, S., & Jahromi, L. B. (2008). Language outcome in autism: Randomized comparison of joint attention and play interventions. *Journal of Consulting and Clinical Psychology, 76*, 125–137.

Lace, J. W., Seitz, D. J., Austin, T. A., Kennedy, E. E., Ferguson, B. J., & Mohrland, M. D. (2022). The dimensionality of the behavior rating inventory of executive function, second edition in a clinical sample. *Applied Neuropsychology: Child, 11*, 579–590.

Lai, M. C., Kassee, C., Besney, R., Bonato, S., Hull, L., Mandy, W., Szatmari, P., & Ameis, S. H. (2019). Prevalence of co-occurring mental health diagnoses in the autism population: A systematic review and meta-analysis. *Lancet Psychiatry, 6*, 819–829.

Lamsal, R., Dutton, D. J., & Zwicker, J. D. (2018). Using the ages and stages questionnaire in the general population as a measure for identifying children not at risk of a neurodevelopmental disorder. *BMC Pediatrics, 18*, 122.

Lee, H., Marvin, A. R., Watson, T., Piggot, J., Law, J. K., Law, P. A., Constantino, J. N., & Nelson, S. F. (2010). Accuracy of phenotyping of autistic children based on Internet implemented parent report. *American Journal of Medical Genetics Part B, 153B*, 1119–1126.

Lenhard, A., Lenhard, W., Suggate, S., & Segerer, R. (2018). A continuous solution to the norming problem. *Assessment, 25*, 112–125.

Lenhard, W., & Lenhard, A. (2021). Improvement of norm score quality via regression-based continuous norming. *Educational and Psychological Measurement, 81*, 229–261.

Maenner, M. J., Warren, Z., Williams, A. R., Amoakohene, E., Bakian, A. V., Bilder, D. A., Durkin, M. S., Fitzgerald, R. T., Furnier, S. M., Hughes, M. M., Ladd-Acosta, C. M., McArthur, D., Pas, E. T., Salinas, A., Vehorn, A., Williams, S., Esler, A., Grzybowski, A., Hall-Lande, J., … Shaw, K. A. (2023). Prevalence and characteristics of autism spectrum disorder among children aged 8 years - autism and developmental disabilities monitoring network, 11 Sites, United States, 2020. *Mmwr Surveillance Summaries, 72*, 1–14.

Marjanovic, Z., Bajkov, L., & MacDonald, J. (2019). The conscientious responders scale helps researchers verify the integrity of personality questionnaire data. *Psychological Reports, 122*, 1529–1549.

Markowitz, L. A., Reyes, C., Embacher, R. A., Speer, L. L., Roizen, N., & Frazier, T. W. (2016). Development and psychometric evaluation of a psychosocial quality-of-life questionnaire for individuals with autism and related developmental disorders. *Autism, 20*, 832–844.

McClain, M. B., Schwartz, S. E., Bera, J., Farmer, R. L., Serang, S., Harris, B., & Golson, M. E. (2023). Vineland-3 measurement non-invariance in children with and without intellectual and developmental disabilities. *Am J Intellect Dev Disabil, 128*, 334–343.

Micai, M., Fatta, L. M., Gila, L., Caruso, A., Salvitti, T., Fulceri, F., Ciaramella, A., D'Amico, R., Del Giovane, C., Bertelli, M., Romano, G., Schunemann, H. J., & Scattoni, M. L. (2023). Prevalence of co-occurring conditions in children and adults with autism spectrum disorder: A systematic review and meta-analysis. *Neuroscience and Biobehavioral Reviews, 155*, Article 105436.

Mundo, A. I., Tipton, J. R., & Muldoon, T. J. (2022). Generalized additive models to analyze nonlinear trends in biomedical longitudinal data using R: Beyond repeated measures ANOVA and linear mixed models. *Statistics in Medicine, 41*, 4266–4283.

National Academies of Sciences, E., and Medicine. (2024). Applied Behavior Analysis within the Department of Defense's Comprehensive Autism Care Demonstration: Proceedings of a Workshop—in Brief In. Washington, DC.

Padilla, K. L., Sarno, J., & Kazemi, E. (2024). Assessment use in applied behavior analysis. In L. M. Toby, & E. S. Ranade (Eds.), *Psychology essentials for behvior analysts*. New York: Routledge.

Padilla, K. L., Weston, R., Morgan, G. B., Lively, P., & O'Guinn, N. (2023). Validity and reliability evidence for assessments based in applied behavior analysis: A systematic review. *Behavior Modification, 47*, 247–288.

Patrick, M. E., Shaw, K. A., Dietz, P. M., Baio, J., Yeargin-Allsopp, M., Bilder, D. A., Kirby, R. S., Hall-Lande, J. A., Harrington, R. A., Lee, L. C., Lopez, M. L. C., Daniels, J., & Maenner, M. J. (2020). Prevalence of intellectual disability among eight-year-old children from selected communities in the United States, 2014. *Disabil Health J*, Article 101023.

PROMIS® Validity Standards Committee on behalf of the PROMIS® Network. (2013). PROMIS® instrument development and psychometric evaluation scientific standards. In (pp. 1–72).

R Core Team. (2024). _R: A language and environment for statistical computing. In *R Foundation for Statistical Computing*. Vienna, Austria.

Reiss, P. T., & Ogden, R. T. (2009). Smoothing parameter selection for a class of semiparametric linear models. *Journal of the Royal Statistical Society Series B: Statistical Methodology, 71*, 505–523.

Sandbank, M., Bottema-Beutel, K., Crowley LaPoint, S., Feldman, J. I., Barrett, D. J., Caldwell, N., Dunham, K., Crank, J., Albarran, S., & Woynaroski, T. (2023). Autism intervention meta-analysis of early childhood studies (Project AIM): Updated systematic review and secondary analysis. *BMJ, 383*, Article e076733.

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*, 420–428.

Simmons, G. L., Corbett, B. A., Lerner, M. D., Wofford, K., & White, S. W. (2024). Social competence in autism: A structural equation modeling approach. *Autism Research, 17*, 761–774.

Sorensen, O., Fjell, A. M., & Walhovd, K. B. (2023). Longitudinal modeling of age-dependent latent traits with generalized additive latent and mixed models. *Psychometrika, 88*, 456–486.

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland adaptive behavior scales, Third Edition (Vineland-3)*. San Antonio, TX: Pearson,.

Trump, C. E., & Ayres, K. M. (2020). Autism, insurance, and discrimination: The effect of an autism diagnosis on behavior-analytic services. *Behavior analysis in practice, 13*, 282–289.

Uljarevic, M., Cai, R. Y., Hardan, A. Y., & Frazier, T. W. (2022). Development and validation of the Executive Functioning Scale. *Frontiers in Psychiatry, 13*, 1078211.

Uljarevic, M., Frazier, T. W., Jo, B., Scahill, L., Youngstrom, E. A., Spackman, E., Phillips, J. M., Billingham, W., & Hardan, A. (2023). Dimensional assessment of restricted and repetitive behaviors: Development and preliminary validation of a new measure. *Journal of the American Academy of Child and Adolescent Psychiatry, 62*, 568–581.

Uljarevic, M., Frazier, T. W., Phillips, J. M., Jo, B., Littlefield, S., & Hardan, A. Y. (2020). Mapping the research domain criteria social processes constructs to the social responsiveness scale. *Journal of the American Academy of Child and Adolescent Psychiatry, 59*, 1252–1263. e1253.

Uljarevic, M., Spackman, E. K., Cai, R. Y., Paszek, K. J., Hardan, A. Y., & Frazier, T. W. (2022). Daily living skills scale: Development and preliminary validation of a new, open-source assessment of daily living skills. *Front Psychiatry, 13*, 1108471.

Wilkinson, E., Farmer, C., Kleiman, E., & Bal, V. H. (2023). Factor structure of the VABS-3 Comprehensive Parent/Caregiver form in autistic individuals: Poor fit of three-factor and unidimensional models. *Autism*, 13623613231179288.

Wood, S. (2017). *Generalized additive models: an introduction with R (2nd ed.)*. CRC,.

Wood, S. N. (2003). Thin-plate regression splines. *Journal of the Royal Statistical Society (B), 65*, 95–114.

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 & Adolescent Psychology, 48*, 811–824.

Zablotsky, B., Black, L. I., Maenner, M. J., Schieve, L. A., Danielson, M. L., Bitsko, R. H., Blumberg, S. J., Kogan, M. D., & Boyle, C. A. (2019). Prevalence and trends of developmental disabilities among children in the United States: 2009-2017. *Pediatrics, 144*.

Zhu, J., & Chen, H.-Y. (2011). Utility of inferential norming with smaller sample sizes. *Journal of Psychoeducational Assessment, 29*, 570–580.