Reliability of the Commonly Used and Newly-Developed Autism Measures

Thomas W. Frazier

Department of Psychology, John Carroll University, Cleveland, OH, USA

Departments of Pediatrics and Psychiatry, SUNY Upstate Medical University, Syracuse, NY, USA

(ORCID: 0000-0002-6951-2667)

Andrew J. O. Whitehouse

Telethon Kids Institute, University of Western Australia, Australia

Susan R. Leekam

School of Psychology, College of Biomedical and Life Sciences, Cardiff University, UK

Sarah J. Carrington

School of Psychology, College of Health and Life Sciences, Aston University, Birmingham, UK

(ORCID: 0000-0001-5548-8793)

Gail A. Alvares

Telethon Kids Institute, University of Western Australia, Australia

David W. Evans

Department of Psychology, Program in Neuroscience, Bucknell University, Lewisburg, PA, USA

Antonio Y. Hardan

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

Mirko Uljarević

Department of Psychiatry and Behavioral Sciences, Stanford University, Stanford, CA, USA

Author Note

Thomas W. Frazier, https://orcid.org/0000-0002-6951-2667

Correspondence concerning this article should be addressed to Thomas W. Frazier, Department of Psychology, John Carroll University, 1 John Carroll Boulevard, University Heights, OH 44118. Email: tfrazier@jcu.edu.

Financial support for this project was provided by Autism Speaks.

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, Roche Pharma, MaraBio, Scioto Biosciences, 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. Uljarevic has equity options and an advisory role in Quadrant Biosciences and has an investor stake in iSCAN-R. Dr. Hardan has equity options and an advisory role in Quadrant Biosciences, has an investor stake in iSCAN-R, and is a consultant/advisor for Jazz Pharmaceuticals, Beaming Health, and IAMA Therapeutics.

Drs. Frazier, Uljarevic, and Hardan were involved in the development and validation of the Autism Symptom Dimensions Questionnaire, the Dimensional Assessment for Restricted and Repetitive Behaviors, and the Stanford Social Dimensions Scale. Dr. Evans was involved in the development of the Childhood Routines Inventory-Revised. Drs. Leekam and Carrington were involved in the development of the Diagnostic Interview for Social and Communication Disorders and Dr. Leekam was involved in the development of the Repetitive Behavior Questionnaire-2.

Abstract

Although existing data suggest adequate to excellent reliability for existing, commonly-used autism measures, few studies have had sufficient sample sizes to compare scale and conditional reliability across instruments. The aim of the present study was to compare scale reliability (internal consistency, average corrected item-total correlations, and model reliability) and conditional reliability derived from item response theory analyses among the most commonly used, as well as several newly developed, observation, interview, and parent-report autism instruments. When available, data sets were combined to facilitate large sample evaluation. Scale and conditional reliability estimates were computed for total scores and for subscales. Although results indicated generally good to excellent scale reliability for total scores for all measures, scale reliability was weaker for RRB subscales of the ADOS and ADI-R, reflecting the relatively small number of items for these measures. For diagnostic measures, conditional reliability tended to be very good (>.80) in the regions of the latent trait where ASD and non-ASD developmental disability cases would be differentiated. For parent-report scales, conditional reliability of total scores tended to be excellent (>.90) across very wide ranges of autism symptom levels, with a few notable exceptions. In general, these findings support the use of all of the clinical observation, interview, and parent-report autism symptom measures examined but also suggest specific limitations that warrant consideration when choosing measures for specific clinical or research applications.

Keywords: autism; reliability; item response theory; observation; interview; questionnaire

Reliability of the Commonly Used and Newly-Developed Autism Measures

Comprehensive and sensitive capture of autism symptoms is crucial for screening, diagnosis, and longitudinal monitoring, including tracking symptom change with development and during psychosocial or medical interventions (Charman & Gotham, 2013). A number of childhood autism measures have been developed over the last three decades, including observational, clinical interview, and informant(parent)-report measures (Lord et al., 2020). Although many of these were initially developed for the screening or diagnostic context, most have since been used to monitor change over time in autism symptom presentation. Careful evaluation of each measure's psychometric properties is crucial for deciding their utility and appropriateness for specific applications, particularly in situations where it is not feasible to include multiple autism symptom measures due to limited time or resources.

A number of psychometric properties are relevant to instrument development, selection and use (Boateng et al., 2018). Reliability is a particularly important characteristic to consider as measurement error constrains validity. More precisely, the maximum validity coefficient is a function of the square root of the reliability coefficients of the variables being examined (Streiner & Norman, 2008). Thus, lower reliability reduces power and can inhibit the ability to detect significant relationships in research (Leon et al., 1995). Instrument reliability is particularly important to understand prior to clinical use, as the error around an individual score can greatly influence interpretation (Streiner & Norman, 2008). While reliability is dependent on context, in general, measures with higher reliability (lower measurement error) in diverse samples are likely to produce more precise estimates of autism symptom level in most clinical and research contexts and are thus preferable, especially for making clinical decisions that have important implications for individuals and their families.

Internal consistency and model reliability of autism measures have been evaluated relatively frequently, including in US (Frazier et al., 2023; Taylor et al., 2022) and international samples (Murray et al., 2017; Nguyen et al., 2019). Although it is encouraging that conditional reliability derived from item response theory analyses has gained more attention in the last five years, it remains underutilized (Janvier et al., 2022; Taylor et al., 2022). Conditional reliability is particularly important to examine because it

permits and evaluation of measurement precision across different score ranges, including ranges most relevant for clinical assessment (moderately high to very high symptom scores) and score ranges relevant to intervention settings (average to high scores). In addition, the majority of previous studies that have focused on the reliability of autism measures have relied on samples of small to moderate size.

To address the relatively limited view of reliability of the autism symptom literature, the present study aimed to evaluate the scale and conditional reliability across existing autism symptom measures. This investigation focused on observation, interview, and informant(parent) report measures and aimed to include both the most widely used and well-established measures and newly developed autism symptom measures where data was accessible, either through publicly available datasets and data sharing initiatives or the internal collaborative networks. We hypothesized that there would be significant differences in scale and conditional reliability across measures, with newly-developed measures yielding higher reliability estimates than existing commonly-used measures, given that they were developed and validated based on modern measurement development frameworks. We further expected that informant-report measures would show higher reliability than observation or interview measures given that informant-report instruments are easier to collect, can rely on parent knowledge of the child – leading to larger item pools per instrument, – and have tended to focus on poly-ordinal (Likert) scales which tend to yield higher reliability (Simms et al., 2019).

Methods

Data Sets

For each measure, available data sets were obtained from the following sources: Simons Simplex Collection (SSC) (Fischbach & Lord, 2010), the Autism Genetic Resource Exchange (AGRE) (Geschwind et al., 2001), National Database for Autism Research (NDAR) (Hall et al., 2012), DISCO clinic data (Leekam et al., 2000; S. Leekam et al., 2007; Leekam et al., 2002b; S. R. Leekam et al., 2007; Wing et al., 2002), Social Responsiveness Scale normative data (SRS Norm) (Constantino & Gruber, 2012), Healthy Brain Network (HBN) (Alexander et al., 2017), ASDQ Norms (Frazier et al., 2023), SSDS Norms (Phillips et al., 2019), CRI-R Norms (Evans et al., 2017), and DARB Norms (Uljarević et

al., 2022). When measures were present across datasets, samples were combined to create aggregate datasets while removing (where known) duplicate cases. Specifically, three samples were combined to create the Autism Diagnostic Observation Schedule (ADOS) dataset (SSC, AGRE, NDAR), five samples were combined to create the Social Responsiveness Scale (SRS) dataset (HBN, SSC, NDAR, SRS Norm, and AGRE), three samples were combined to create the Social Communication Questionnaire (SCQ) and Repetitive Behavior Scale – Revised (RBS-R) datasets (HBN, NDAR, and SSC). Information on the cohorts included for each measure is included in Supplemental Table 1.

Measures

Autism Diagnostic Observation Schedule (ADOS)

The ADOS/ADOS-2 (first and second editions) is a clinician-observation measure of autism symptoms (Lord et al., 2002; Lord et al., 2012). The measure includes five modules (toddler and modules 1-4) that are administered dependent on age and speech/language status. For the present study, only data from modules 1-4 were available. Each module was analyzed separately using only items included in the respective ADOS-2 algorithm scores. Although not typically interpreted, SCI (social affect) and RRB subscales were also scored and analyzed to independently evaluate reliability for these domains and for comparison to other measures where SCI and RRB domain scores are computed.

Autism Diagnostic Interview-Revised (ADI-R)

The ADI-R (Lord et al., 1994) is a standardized, semi-structured clinical interview for caregivers of children and adults. For the present study, item mapping to DSM-5 was used to identify total and SCI and RRB subscales (Huerta et al., 2012). For the total and subscales, item scores of 3 were recoded to 2 to be consistent with instrument scoring.

Diagnostic Interview for Social and Communication Disorders (DISCO)

The DISCO is a 320-item semi-structured interview used by clinicians to elicit information from caregivers about the individual's profile of development and behavior. Across different samples, it has been shown to have good sensitivity and specificity (Carrington et al., 2015; Carrington et al., 2014; Kent et al., 2013; Maljaars et al., 2012), and interrater reliability ($\kappa \geq 0.7$) (Wing et al., 2002) and criterion

validity (Leekam et al., 2002a; Maljaars et al., 2012; Nygren et al., 2009). It has a DSM-5 algorithm item set (Kent et al., 2013), also published in abbreviated form (Carrington et al., 2019; Carrington et al., 2014). The abbreviated item set (48 items) was used for this analysis.

Social Responsiveness Scale (SRS)

The SRS/SRS-2 (first and second editions) is 65-item, parent-report, ordinally-scaled (1= "not true" to 4= "almost always true") quantitative assessment of the severity of autism traits. It is one of the most frequently used quantitative measures of autism symptoms (Constantino & Gruber, 2012). SCI and RRB subscales were derived from SRS-2 scoring.

Social Communication Questionnaire (SCQ)

The lifetime version of the SCQ is a parent-report dichotomously-keyed (yes/no) rating scale that consists of 40 questions many of which tap DSM-IV-TR symptom domains (Rutter et al., 2003). Lifetime ratings reference the child's behavior throughout their developmental history, increasing diagnostic validity (Lord et al., 1997). Items 2-39 were summed for the total score and SCI and RRB subscales were determined by the authors using item content and based on prior factor analyses (Uljarevic et al., 2020; Uljarević et al., 2021).

Autism Symptom Dimensions Questionnaire (ASDQ)

The ASDQ is a newly-created 39-item autism symptom measure informed by DSM-5 criteria and recent factor analyses of autism symptom data, with input from informant caregivers and autism clinicians (Frazier et al., 2023). Items are rated using a 5-point Likert scale (1=Never, 2=Rarely, 3=Sometimes, 4=Often, 5=Very Often). The SCI and RRB subscales include 17 and 18 items, respectively.

Stanford Social Dimensions Scale (SSDS)

The SSDS is a 58-item dimensional measure designed to provide parental perspective on their child's social abilities (Phillips et al., 2019). Factor analyses have suggested a five-factor solution with factors interpreted as Social Motivation (SM), Social affiliation (SA), Expressive Social Communication

(ESC), Social Recognition (SR), and Unusual Approach (UA). Each of these factors is positively correlated and therefore, items were treated as a single set for evaluation of the SCI domain.

Repetitive Behavior Scale – Revised (RBS-R)

The RBS-R is a 43-item parent-report rating scale for measuring the presence and severity of a variety of forms of restricted, repetitive behavior that are characteristic of individuals with ASD (Lam & Aman, 2007). The RBS-R consists of 6 subscales: stereotyped behavior, self-injurious behavior, compulsive behavior, routine behavior, sameness behavior, and restricted behavior. For the present study, items from all subscales except self-injurious behavior were analyzed as a single scale to evaluate measurement of the RRB domain.

Repetitive Behavior Questionnaire – 2 (RBQ-2)

The RBQ-2 is a 20-item parent-report questionnaire. Factor analytic studies across ASD (Barrett et al., 2015; Lidstone et al., 2014) and normative development (S. Leekam et al., 2007; Uljarevic et al., 2017) have suggested that the RBQ-2 has a stable two-factor structure encompassing repetitive sensory-motor and insistence on sameness factors. For the present study, all items were analyzed as a single scale to evaluate the RRB domain.

Childhood Routines Inventory – Revised (CRI-R)

The CRI-R is a 62-item, parent-report measure rated on a five-point Likert scale (Evans et al., 2017). Items evaluate stereotypies, tics, compulsions, habits and routines, rigidity, insistence on sameness, and sensory sensitivities. While the instrument assesses two broad domains of RRB, these domains are highly correlated. For the present study, items were treated as a single scale for evaluation of measurement precision for the RRB domain.

Dimensional Assessment for Restricted and Repetitive Behaviors (DARB)

The DARB is a new measure of RRB that was developed and refined through the iterative series of steps described by the PROMIS framework. Concepts guiding item development included (i) good coverage of the full range of symptom severity and presentations, (ii) applicability across the cognitive functioning range, and (iii) applicability across the lifespan. Measure development was informed by

recent factor analyses of RRB symptoms and the final measure includes 98 items rated using a 5-point Likert scale. For the present study, only 96 items with significant loadings on 7 of the 8 subscales were combined into a single RRB scale; self-injury items were excluded.

Statistical Analyses

Descriptive statistics were computed separately for ASD, developmental delay (DD), and neurotypical (NT) groups across demographic (age, sex) and clinical factors (IQ), when available, to characterize the combined cohorts for each measure.

Missing Data Handling

Missing data was modest across measures (0%-5.7%; Supplement 1). Thus, for each measure with missing data, five imputed datasets were generated using fully conditional Monte Carlo Markov Chain specification with 10 iterations. Classical test theory reliability analyses were computed on the original and each imputed dataset. In each case, deviations across datasets were very small (<.01). Mean values across imputed datasets are presented. Item response theory analyses were computed with the original data assuming that data were missing at random.

Classical Test Theory (CTT) - Scale Reliability

To evaluate measurement precision for each measure, CTT reliability coefficients (internal consistency and average corrected item-total correlations) (Streiner & Norman, 1995) and model reliability (MacDonald's coefficient ω) (Revelle & Condon, 2019) were computed using all items as inputs (total scales), only SCI items as inputs (SCI scales), and only RRB items as inputs (RRB scales). Model reliability was computed using factor loadings from a single factor confirmatory factor analysis using an SPSS macro (Hayes & Coutts, 2020). Confidence intervals (95%) were also calculated for internal consistency reliability estimates. Internal consistency and model reliability estimates falling in the ranges <.70, .70 to .79, .80 to .89, and >.90 were considered poor, fair, good, and excellent (Nunnally & Bernstein, 1994). Average corrected item-total correlations >=.30 were considered at least adequate (Streiner & Norman, 1995). To evaluate the association between original publication year and CTT

reliability, bivariate non-parametric Spearman's rho correlations were computed between internal consistency reliability coefficients transformed to Fisher's z and publication year.

Item Response Theory (IRT) - Conditional Reliability

IRT analyses (Embretson & Reise, 2000; Hambleton et al., 1991; Reise et al., 1993) were conducted using maximum likelihood estimation with robust standard errors and a logit link with the single factor mean and variance fixed to 0 and 1, respectively. Principal components analyses were first conducted to ensure that each measure had a large first principal component indicating that a substantial proportion of the variance in items scores reflected a general dimension consistent with scoring. After checking dimensionality, unifactorial IRT analyses were completed for each measure total and subscale score. Scale information estimates were converted to conditional reliability using the formula: reliability = 1 – [1/Information(theta)] (Thissen, 2000) from theta= -6 to +6.

Comparisons between measures within each category (clinical observation, parent interview, parent-report) were conducted using repeated measures analysis of variance with conditional reliability coefficients (after conversion to Fisher's z) as the dependent variable and specific measure (e.g., ADOS module 1 vs. module 2, etc.) as the independent variable. Comparisons across categories were computed using repeated measures analysis of variance by first averaging conditional reliability estimates across different measures within each category. This analysis examines whether observational, interview, or parent-report total scores show different levels of conditional reliability.

Statistical Power

Scale and conditional reliability analyses are considered over-powered given the large sample sizes for each measure. The only exception is for the newly-developed SSDS, where the sample is much smaller (N=170), but still adequate as an initial evaluation of reliability. Repeated measures comparisons of conditional reliability estimates were expected to have at least adequate power ($\geq$.80) to detect a small-to-medium effect size or larger ($d \geq .36$), assuming 61 observations along the information curve (from theta -6 to +6) for a two-measure comparison (α=.05, two-tailed).

Data preparation, descriptive analyses, internal consistency reliability, corrected item-total correlations used SPSS v28 (IBM Corp, 2021). Item response theory analyses were computed in MPlus version 8.5 (Muthén & Muthén, 1998-2017).

Results

Measure Cohorts

Relevant to the present analyses, sample composition varied widely, in terms of the proportion of ASD, non-ASD DD, and neurotypical participants (Supplement 1). However, all samples contained the full range of scores on all items, supporting the inclusion of a wide range of symptom levels regardless of specific diagnostic composition. Age and sex also varied widely but was consistent with generally younger ages, often present in diagnostic clinic samples, and with a high proportion of males often seen in ASD-diagnosed samples. When available, average sample IQs for autistic participants ranged from very low (SS=50) to average (SS=102), with generally average sample average IQs in the DD and neurotypical groups (Supplement 1). Missing data rates were low (0% to 5.7%), indicating that selective data attrition is not likely to influence results.

Clinical Observation Measures

Total scale reliability fell in the good to excellent range across ADOS modules (Table 1). Conditional reliability was generally adequate (>.70) in the middle of the score range (theta -1.5 to +1.1) for most ADOS modules (Figure 1), with a shift toward better conditional reliability at lower scores for module 1 – “few to no words” and a shift upward toward better conditional reliability at higher scores for module 4. These deviations may reflect the inclusion of “easier” items for module 1 and “harder” items for module 4 or may reflect slight differences in the population as module 1 is often administered to young children with early developing language levels and module 1 – “few to no words” had the lowest number of neurotypical cases in that cohort. Comparisons across modules indicated stronger conditional reliability for modules 2 and 3 relative to modules 1 (both forms) and 4 [$F(4, 240) = 7.67$, $p<.001$].

Scale reliability remained good to excellent for ADOS SCI scales across modules (Table 2). However, scale reliability was poor for ADOS RRB scales, consistent with the small number of items (4-

5 per module). Similarly, conditional reliability for ADOS SCI scales was good to excellent and only slightly weaker than for total scales (Supplement 2), while conditional reliability was generally inadequate for ADOS RRB scales (Supplement 3), with only fair to good levels in very narrow score ranges, which is not surprising given that the ADOS was not developed to provide an in-depth RRB assessment. Comparisons across modules indicated small but significant differences for SCI [$F(4, 240) = 4.65$, $p=.001$] with higher conditional reliability for modules 2 and 3 relative to other modules. Large significant differences were observed for RRB scales, with the highest conditional reliability observed for module 3 and the lowest for module 1 (both forms) [$F(4, 240) = 15.14$, $p<.001$].

Parent Interview Measures

Total scale reliability fell in the good to excellent range for both interview measures (Table 1). Conditional reliability was generally adequate (>.70) in the middle of the score range (theta -2.0 to +1.8; Figure 2), with better conditional reliability for the DISCO than the ADI-R [$F(1, 60) = 14.35$, $p<.001$]. Subscale reliability was excellent for the ADI-R SCI and good for the DISCO SCI, but RRB subscale values fell in the poor and fair ranges, respectively (Table 2). Conditional reliability for SCI subscales was at least adequate (≥.70) in the average score range (theta -1.8 to 2.5) with no significant difference between the ADI-R and DISCO [$F(1, 60) = 1.59$, $p=.212$] (Supplement 4). However, for RRB subscales, conditional reliability was substantially better for the DISCO relative to the ADI-R [$F(1, 60) = 58.81$, $p<.001$], with adequate levels extending from extremely low to very high scores (theta -3.4 to +2.1) (Supplement 5).

Parent-Report Questionnaire Measures

Scale reliability was excellent (≥.90) for all parent-report total scales (Table 1). While the SRS and ASDQ total scores had excellent conditional reliability coverage from very low to extremely high scores (theta -2.8 to +4.5), the SCQ total score only maintained good reliability from low to very high scores (theta -1.4 to +3.0) (Figure 3). Not surprisingly, given differences in content coverage, average conditional reliability varied substantially across instruments [$F(2, 120) = 123.57$, $p<.001$], The SRS

(r_{xx}=.86) and ASDQ (r_{xx}=.85; p<.001) did not significantly differ (p=.142), while the SCQ had much lower average conditional reliability (r_{xx}=.64) than both measures (both p<.001).

Parent-report SCI and RRB subscales also had excellent scale reliability for most instruments; the exceptions being for the SCQ-RRB scale and RBQ-2, which fell in the good range (Table 2). Conditional reliability coverage was at least adequate (≥.70) from very low to extremely high (theta -2 to +4.6) scores for the SRS, ASDQ, and SSDS SCI subscales (Table 2 and Supplement 6). Similarly, conditional reliability for RRB subscales was at least adequate (≥.70) from low to extremely high (theta -1.6 to +2.9) scores for the SRS, ASDQ, RBS-R, CRI-R, DARB, and RBQ-2 (Supplement 7). For both the SCI and RRB scales, the SCQ had weaker conditional reliability coverage.

Average conditional reliability varied significantly across the SCI [F(3, 180) = 118.89, p<.001] and RRB [F(6, 360) = 163.05, p<.001] subscales. For SCI scales, the SSDS (r_{xx}=.90) had the highest average conditional reliability followed by the SRS (r_{xx}=.85), ASDQ (r_{xx}=.77), and SCQ (r_{xx}=.56). For RRB scales, the DARB (r_{xx}=.91) and CRI (r_{xx}=.88) had the highest conditional reliability followed by the RBS-R (r_{xx}=.78), ASDQ (r_{xx}=.77), and RBQ-2 (r_{xx}=.72). Average conditional reliability estimates for the SRS (r_{xx}=.60) and SCQ (r_{xx}=.41) were weaker.

A significant positive correlation was observed between publication year and internal consistency reliability across all instruments (r=.66, p=.007; Supplement 8).

Informant-report total scores tended to have higher average conditional reliability (r_{xx}=.80) than interview total scores (r_{xx}=.73; p<.001), which were higher than the average conditional reliability for ADOS observation total scores (r_{xx}=.64; p<.001) [F(2, 120) = 55.12, p<.001].

Discussion

High levels of reliability across nearly all total scales and most SCI and RRB subscales indicate that for many applications, the choice of measure should not be driven primarily by reliability, but rather by information source (clinician, parent interview, parent report) as well as by content, construct, and predictive (diagnostic) validity considerations. This is particularly true since the measures evaluated in this study all tend to cover a broad age range but have very different item content, coverage of autism

symptom domains, levels of diagnostic differentiation (Sanchez & Constantino, 2020), and different demographic and clinical factors influencing observed scores (Frazier et al., 2014).

It is crucial to highlight certain exceptions where reliability becomes crucial to consider. For example, the ADOS modules were not designed for symptom monitoring and have only a small number of RRB items. Thus, using ADOS modules to assess change over time is not warranted since they have smaller score ranges with the measurement precision needed to accurately evaluate individual differences and present with limited scale and conditional reliability for the RRB domain. It is, however, important to emphasize that the ADOS was not developed with the intention to provide a comprehensive and dimensional characterization of RRB or to track autism symptom levels across a wide range of trait levels, thus, identified limitations are not surprising and do not detract from the utility of the instrument when used for its primary purpose. Recent development efforts at creating a brief observation instrument for tracking symptoms in toddlers may address this limitation (Grzadzinski et al., 2016). Similarly, the SCQ was clearly inferior in both scale and conditional reliability to the SRS and ASDQ. The latter two measures had highly similar total scale and conditional reliability estimates, with the SRS showing slightly stronger reliability for SCI and the ASDQ showing stronger reliability for RRB symptoms. The equivalence in total scale reliability is notable given that the SRS is a commercial measure with 65 items and limited symptom domain coverage, while the ASDQ is a 39-item, free, open-access measure with strong coverage of symptom domains consistent with DSM-5 criteria and recent factor analyses as well as good coverage of constructs aligned with dimensional frameworks such as the National Institute of Mental Health's Research Domain Criteria initiative (Frazier & Hardan, 2017; Uljarevic et al., 2019, 2020; Uljarevic et al., 2021).

In general, newly developed parent-report measures had better reliability than older, widely used measures. However, it is important to highlight that RBQ-2, although developed almost two decades ago, also showed excellent properties. This is not too surprising, given that these measures were designed to have better content coverage of autism symptom domains. This suggests that clinicians and researchers should focus on the intended purpose when choosing a measure. For example, if the purpose is simply

diagnostic differentiation or broad coverage of autism symptom level, then existing measures are likely to be more efficient, while if the purpose is more detailed coverage of specific symptom domains to generate more detailed treatment recommendations, newer measures should be considered. In this regard, the ASDQ may present a good balance, as it has strong diagnostic differentiation for a parent-report instrument, includes good SCI and RRB coverage, is free, has only 39 items, and assesses well-replicated autism symptom sub-domains. For applications where a detailed assessment of SCI or RRB domains is required, the SSDS, RBS-R, and DARB are the best choices among the parent-report scales.

Reliability was stronger for newer measures and for informant-report measures relative to interview and observation measures. These observations are likely due to greater attention to psychometrics in recent measure development efforts, especially content and construct coverage, and the fact that informant-report questionnaires are often able to include higher numbers of items with poly-ordinal (Likert) response scales. While not surprising, this does emphasize the need to focus on measures that have an adequate number of items and that response scales are chosen to provide a useful range of information about specific behaviors or symptoms while also making sure that response choices are relevant and easy to rate.

Based on the present findings, the following recommendations are made for considering reliability in measure selection and future measure development. 1) ADOS modules all have good reliability for diagnostic evaluation use. However, reliability is insufficient in situations where monitoring a wide range of symptom levels is desired. Future revisions of the ADOS might consider adding RRB items to enhance the reliability of the assessment of this domain. Using a broader Likert scale may also be useful for reliably capturing individual differences across a wide range of autism symptom presentations. 2) The ADI-R had good total scale reliability and adequate conditional reliability for measuring individual differences in the range important for diagnostic differentiation. However, conditional reliability was generally lower than the DISCO, and the ADI-R appears less able to capture more significant social communication/interaction symptoms or subtler restricted/repetitive behaviors. For measuring a wider range of autism symptom presentations, the DISCO appears to be a better choice when a parent-interview

measure is desired. 3) The SCQ, given a weaker scale and conditional reliability, should only be chosen in situations where other comparable parent-report measures (SRS and ASDQ) are not feasible. The measure has been attractive to some users based on the simplicity of the dichotomous (yes/no) response scale, but the present analyses suggest that this scaling approach substantially reduces conditional reliability. 4) If only a single parent-report measure can be implemented and a measure that covers SCI and RRB domains is required, either the SRS or ASDQ should be considered. The choice between these two measures should be largely based on content coverage, construct validity, and predictive validity. 5) Dedicated parent-report SCI and RRB measures, such as the SSDS, DARB, CRI-R, RBQ-2, have very strong reliability profiles. Similarly, the choice among these measures should be largely based on content coverage, construct validity, predictive validity, and practical considerations.

Limitations and Future Directions

Samples for each measure were selected based on the availability of large datasets. For observation and interview measures, these samples generally reflect at-risk populations rather than the full population with neurotypical individuals. While all measures had the full score range represented, these sampling differences may shift the conditional reliability curves toward the lower score range (leftward shift). Thus, the middle of the latent trait (and the peaks of the conditional reliability curves) may actually be in ranges where ASD and non-ASD developmental disability cases are differentiated rather than representing measurement precision across the full population. Further, although every attempt was made to include a broad set of both widely used and well-established measures as well as newly developed measures, this study only included the measures where data was accessible to authors. For instance, of the observation measures, only the ADOS modules were available. Future studies including other observation measures, such as the Childhood Autism Rating Scale (Schopler et al., 2010) or the Autism Observation Scale for Infants (Bryson et al., 2008), are needed. Furthermore, future studies should attempt to evaluate additional commonly used screening and diagnostic instruments, such as the Autism Spectrum Quotient (Baron-Cohen et al., 2001), Modified Checklist for Autism in Toddlers (Robins et al., 2014), Communication and Symbolic Behavior Scales (Wetherby & Prizant, 2003), Screening Tool for Autism

in Toddlers (Stone et al., 2008), TELE-ASD-PEDS (Corona et al., 2021), Autism Impact Measure (Kanne et al., 2014), and Brief Observation of Social Communication Change (Grzadzinski et al., 2016; Kitzerow et al., 2016).

Finally, the present study did not evaluate measurement invariance or differential item functioning. These parameters are also key to understanding whether scales are measuring consistently across relevant demographic and clinical groups. Unfortunately, the datasets aggregated for the present study did not include sufficient information on race/ethnicity and all data were from US samples. While prior work suggests that many measures show good invariance, at least across age, sex, and race/ethnicity in US and UK populations, these findings merit replication in large datasets and would be important considerations in future measure choice in clinical and research settings. Nor did this study evaluate test-retest or inter-rater reliability, key consideration, particularly for observation and interview measures and longitudinal monitoring applications.

Conclusion

In summary, the present study demonstrated good to excellent scale and conditional reliability for nearly all measures evaluated, with the notable exceptions of weaker conditional reliability for the ADOS, ADI-R, and SCQ. Strong reliability estimates extended to measures of the SCI and RRB domains with the exception that ADOS and ADI-R measurement precision was weaker for the RRB domain. Future studies should consider scale and conditional reliability in measure selection, although, for cases where reliability is roughly equivalent, choices should be based on intended use, cost and availability, applicability to the target population, and validity considerations.

References

Alexander, L. M., Escalera, J., Ai, L., Andreotti, C., Febre, K., Mangone, A., Vega Potler, N., Langer, N., Alexander, A, , Kovacs, M., Litke, S, O'Hagan, B, , Andersen, J., Bronstein, B, , Bui, A., Bushey, M., Butler, H., Castagna, V, , Camacho, N., Chan, E., Citera, D., Clucas, J, , Cohen, S., Dufek, S., Eaves, M., . . . Gregory, C. (2017). *The Healthy Brain Network Biobank: An open resource for transdiagnostic research in pediatric mental health and learning disorders.* Cold Spring Harbor Laboratory. https://doi.org/https://doi.org/10.1101/149369

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Clubley, E. (2001). The autism-spectrum quotient (AQ): evidence from Asperger syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disorders*, *31*(1), 5-17. https://doi.org/10.1023/a:1005653411471

Barrett, S. L., Uljarevic, M., Baker, E. K., Richdale, A. L., Jones, C. R., & Leekam, S. R. (2015). The Adult Repetitive Behaviours Questionnaire-2 (RBQ-2A): A Self-Report Measure of Restricted and Repetitive Behaviours. *Journal of Autism and Developmental Disorders*, *45*(11), 3680-3692. https://doi.org/10.1007/s10803-015-2514-6

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. https://doi.org/10.3389/fpubh.2018.00149

Bryson, S. E., Zwaigenbaum, L., McDermott, C., Rombough, V., & Brian, J. (2008). The Autism Observation Scale for Infants: scale development and reliability data. *Journal of Autism and Developmental Disorders*, *38*(4), 731-738. https://doi.org/10.1007/s10803-007-0440-y

Carrington, S., Leekam, S., Kent, R., Maljaars, J., Gould, J., Wing, L., Le Couteur, A., Van Berckelaer-Onnes, I., & Noens, I. (2015). Signposting for diagnosis of Autism Spectrum Disorder using the Diagnostic Interview for Social and Communication Disorders (DISCO). *Research in Autism Spectrum Disorders*, *9*, 45-52. https://doi.org/10.1016/j.rasd.2014.10.003

Carrington, S. J., Barrett, S. L., Sivagamasundari, U., Fretwell, C., Noens, I., Maljaars, J., & Leekam, S. R. (2019). Describing the Profile of Diagnostic Features in Autistic Adults Using an Abbreviated Version of the Diagnostic Interview for Social and Communication Disorders (DISCO-Abbreviated). *Journal of Autism and Developmental Disorders*, *49*(12), 5036-5046. https://doi.org/10.1007/s10803-019-04214-7

Carrington, S. J., Kent, R. G., Maljaars, J., Le Couteur, A., Gould, J., Wing, L., Noens, I., Van Berckelaer-Onnes, I., & Leekam, S. R. (2014). DSM-5 Autism Spectrum Disorder: In search of essential behaviours for diagnosis. *Research in Autism Spectrum Disorders*, *8*(6), 701-715. https://doi.org/10.1016/j.rasd.2014.03.017

Charman, T., & Gotham, K. (2013). Measurement Issues: Screening and diagnostic instruments for autism spectrum disorders - lessons from research and practice. *Child Adolesc Ment Health*, *18*(1), 52-63. https://doi.org/10.1111/j.1475-3588.2012.00664.x

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

Corona, L. L., Wagner, L., Wade, J., Weitlauf, A. S., Hine, J., Nicholson, A., Stone, C., Vehorn, A., & Warren, Z. (2021). Toward Novel Tools for Autism Identification: Fusing Computational and Clinical Expertise. *Journal of Autism and Developmental Disorders*, *51*(11), 4003-4012. https://doi.org/10.1007/s10803-020-04857-x

Embretson, S. E., & Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Evans, D. W., Uljarevic, M., Lusk, L. G., Loth, E., & Frazier, T. (2017). Development of Two Dimensional Measures of Restricted and Repetitive Behavior in Parents and Children. *Journal of the American Academy of Child and Adolescent Psychiatry*, *56*(1), 51-58. https://doi.org/10.1016/j.jaac.2016.10.014

Fischbach, G. D., & Lord, C. (2010). The Simons Simplex Collection: a resource for identification of autism genetic risk factors. *Neuron*, *68*(2), 192-195. https://doi.org/10.1016/j.neuron.2010.10.006

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, T. (2023). The Autism Symptom Dimensions Questionnaire: Development and psychometric evaluation of a new, open-source measure of autism symptomatology. *Developmental Medicine and Child Neurology*. https://doi.org/10.1111/dmcn.15497

Frazier, T. W., & Hardan, A. Y. (2017). Equivalence of symptom dimensions in females and males with autism. *Autism*, *21*(6), 749-759. https://doi.org/10.1177/1362361316660066

Frazier, T. W., Youngstrom, E. A., Embacher, R., Hardan, A. Y., Constantino, J. N., Law, P., Findling, R. L., & Eng, C. (2014). Demographic and clinical correlates of autism symptom domains and autism spectrum diagnosis. *Autism*, *18*(5), 571-582. https://doi.org/10.1177/1362361313481506

Geschwind, D. H., Sowinski, J., Lord, C., Iversen, P., Shestack, J., Jones, P., Ducat, L., & Spence, S. (2001). The autism genetic resource exchange: A resource for the study of autism and related neuropsychiatric conditions. *American Journal of Human Genetics*, *69*, 463-466.

Grzadzinski, R., Carr, T., Colombi, C., McGuire, K., Dufek, S., Pickles, A., & Lord, C. (2016). Measuring Changes in Social Communication Behaviors: Preliminary Development of the Brief Observation of Social Communication Change (BOSCC). *Journal of Autism and Developmental Disorders*, *46*(7), 2464-2479. https://doi.org/10.1007/s10803-016-2782-9

Hall, D., Huerta, M. F., McAuliffe, M. J., & Farber, G. K. (2012). Sharing heterogeneous data: the national database for autism research. *Neuroinformatics*, *10*(4), 331-339. https://doi.org/10.1007/s12021-012-9151-4

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of item response theory*. Sage.

Hayes, A. F., & Coutts, J. J. (2020). Use omega rather than Cronbach's alpha for estimating reliability. But... *Communication Methods and Measures*, *14*(1), 1-24.

Huerta, M., Bishop, S. L., Duncan, A., Hus, V., & Lord, C. (2012). Application of DSM-5 criteria for autism spectrum disorder to three samples of children with DSM-IV diagnoses of pervasive developmental disorders. *American Journal of Psychiatry*, *169*(10), 1056-1064. https://doi.org/10.1176/appi.ajp.2012.12020276

IBM Corp. (2021). *IBM SPSS Statistics for Windows*. In (Version 28.0) IBM Corp.

Janvier, D., Choi, Y. B., Klein, C., Lord, C., & Kim, S. H. (2022). Brief Report: Examining Test-Retest Reliability of the Autism Diagnostic Observation Schedule (ADOS-2) Calibrated Severity Scores (CSS). *Journal of Autism and Developmental Disorders*, *52*(3), 1388-1394. https://doi.org/10.1007/s10803-021-04952-7

Kanne, S. M., Mazurek, M. O., Sikora, D., Bellando, J., Branum-Martin, L., Handen, B., Katz, T., Freedman, B., Powell, M. P., & Warren, Z. (2014). The Autism Impact Measure (AIM): initial development of a new tool for treatment outcome measurement. *Journal of Autism and Developmental Disorders*, *44*(1), 168-179. https://doi.org/10.1007/s10803-013-1862-3

Kent, R. G., Carrington, S. J., Le Couteur, A., Gould, J., Wing, L., Maljaars, J., Noens, I., van Berckelaer-Onnes, I., & Leekam, S. R. (2013). Diagnosing autism spectrum disorder: who will get a DSM-5 diagnosis? *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *54*(11), 1242-1250. https://doi.org/10.1111/jcpp.12085

Kitzerow, J., Teufel, K., Wilker, C., & Freitag, C. M. (2016). Using the brief observation of social communication change (BOSCC) to measure autism-specific development. *Autism Res*, *9*(9), 940-950. https://doi.org/10.1002/aur.1588

Lam, K. S., & Aman, M. G. (2007). The Repetitive Behavior Scale-Revised: independent validation in individuals with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*(5), 855-866. https://doi.org/10.1007/s10803-006-0213-z

Leekam, S., Libby, S., Wing, L., Gould, J., & Gillberg, C. (2000). Comparison of ICD-10 and Gillberg's criteria for Asperger syndrome. *Autism*, *4*(1), 11-28.

Leekam, S., Tandos, J., McConachie, H., Meins, E., Parkinson, K., Wright, C., Turner, M., Arnott, B., Vittorini, L., & Le Couteur, A. (2007). Repetitive behaviours in typically developing 2-year-olds. *Journal of Child Psychology and Psychiatry*, *48*(11), 1131-1138. https://doi.org/10.1111/j.1469-7610.2007.01778.x

Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002a). The Diagnostic Interview for Social and Communication Disorders: algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(3), 327-342. https://doi.org/10.1111/1469-7610.00024

Leekam, S. R., Libby, S. J., Wing, L., Gould, J., & Taylor, C. (2002b). The Diagnostic Interview for Social and Communication Disorders: algorithms for ICD-10 childhood autism and Wing and Gould autistic spectrum disorder. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(3), 327-342. http://www.ncbi.nlm.nih.gov/entrez/query.fcgi?cmd=Retrieve&db=PubMed&dopt=Citation&list_uids=11944875

Leekam, S. R., Nieto, C., Libby, S. J., Wing, L., & Gould, J. (2007). Describing the sensory abnormalities of children and adults with autism. *Journal of Autism and Developmental Disorders*, *37*(5), 894-910. https://doi.org/10.1007/s10803-006-0218-7

Leon, A. C., Marzuk, P. M., & Portera, L. (1995). More reliable outcome measures can reduce sample size requirements. *Archives of General Psychiatry*, *52*(10), 867-871. https://doi.org/10.1001/archpsyc.1995.03950220077014

Lidstone, J., Uljarevic, M., Sullivan, J. P., Rodgers, J., McConachie, H., Freeston, M., Le Couteur, A., Prior, M., & Leekam, S. (2014). Relations among restricted and repetitive behviors, anxiety and sensory features in children with autism spectrum disorders. *Research in Autism Spectrum Disorders*, *8*(2), 82-92.

Lord, C., Brugha, T. S., Charman, T., Cusack, J., Dumas, G., Frazier, T., Jones, E. J. H., Jones, R. M., Pickles, A., State, M. W., Taylor, J. L., & Veenstra-VanderWeele, J. (2020). Autism spectrum disorder. *Nat Rev Dis Primers*, *6*(1), 5. https://doi.org/10.1038/s41572-019-0138-4

Lord, C., Pickles, A., McLennan, J., Rutter, M., Bregman, J., Folstein, S., Fombonne, E., Leboyer, M., & Minshew, N. (1997). Diagnosing autism: analyses of data from the Autism Diagnostic Interview. *Journal of Autism and Developmental Disorders*, *27*(5), 501-517.

Lord, C., Rutter, M., DiLavore, P. C., & Risi, S. (2002). *Autism Diagnostic Observation Schedule: ADOS manual*. Western Psychological Services.

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Manual (Part 1): Modules 1-4*. Western Psychological Services.

Lord, C., Rutter, M., & LeCouteur, A. (1994). ADI-Revised: a revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *Journal of Autism and Developmental Disorders*, *24*, 569-685.

Maljaars, J., Noens, I., Scholte, E., & van Berckelaer-Onnes, I. (2012). Evaluation of the criterion and convergent validity of the Diagnostic Interview for Social and Communication Disorders in young and low-functioning children. *Autism*, *16*(5), 487-497. https://doi.org/10.1177/1362361311402857

Murray, A. L., McKenzie, K., Kuenssberg, R., & Booth, T. (2017). Do the Autism Spectrum Quotient (AQ) and Autism Spectrum Quotient Short Form (AQ-S) Primarily Reflect General ASD Traits or Specific ASD Traits? A Bi-Factor Analysis. *Assessment*, *24*(4), 444-457. https://doi.org/10.1177/1073191115611230

Muthén, L. K., & Muthén, B. O. (1998-2017). *Mplus User's Guide. Eighth Edition.* Muthén & Muthén.

Nguyen, P. H., Ocansey, M. E., Miller, M., Le, D. T. K., Schmidt, R. J., & Prado, E. L. (2019). The reliability and validity of the social responsiveness scale to measure autism symptomology in Vietnamese children. *Autism Res*, *12*(11), 1706-1718. https://doi.org/10.1002/aur.2179

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

Nygren, G., Hagberg, B., Billstedt, E., Skoglund, A., Gillberg, C., & Johansson, M. (2009). The Swedish version of the Diagnostic Interview for Social and Communication Disorders (DISCO-10). Psychometric properties. *Journal of Autism and Developmental Disorders*, *39*(5), 730-741. https://doi.org/10.1007/s10803-008-0678-z

Phillips, J. M., Uljarevic, M., Schuck, R. K., Schapp, S., Solomon, E. M., Salzman, E., Allerhand, L., Libove, R. A., Frazier, T. W., & Hardan, A. Y. (2019). Development of the Stanford Social Dimensions Scale: initial validation in autism spectrum disorder and in neurotypicals. *Mol Autism*, *10*, 48. https://doi.org/10.1186/s13229-019-0298-9

Reise, S. P., Widaman, K. F., & Pugh, R. H. (1993). Confirmatory factor analysis and item response theory: Two approaches for exploring measurement invariance. *Psychological Bulletin*, *114*(3), 552-566.

Revelle, W., & Condon, D. M. (2019). Reliability from alpha to omega: A tutorial. *Psychol Assess*, *31*(12), 1395-1411. https://doi.org/10.1037/pas0000754

Robins, D. L., Casagrande, K., Barton, M., Chen, C. M., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for Autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics*, *133*(1), 37-45. https://doi.org/10.1542/peds.2013-1813

Rutter, M., Bailey, A., & Lord, C. (2003). *The Social Communication Questionnaire Manual*. Western Psychological Services.

Sanchez, M. J., & Constantino, J. N. (2020). Expediting clinician assessment in the diagnosis of autism spectrum disorder. *Developmental Medicine and Child Neurology*, *62*(7), 806-812. https://doi.org/10.1111/dmcn.14530

Schopler, E., Van Bourgondien, M., Wellman, G., & Love, S. (2010). *Childhood Autism Rating Scale – 2nd Edition*. Western Psychological Services.

Simms, L. J., Zelazny, K., Williams, T. F., & Bernstein, L. (2019). Does the Number of Response Options Matter? Psychometric Perspectives Using Personality Questionnaire Data. *Psychological Assessment*, *31*(4), 557-566. https://doi.org/10.1037/pas0000648

Stone, W. L., McMahon, C. R., & Henderson, L. M. (2008). Use of the Screening Tool for Autism in Two-Year-Olds (STAT) for children under 24 months: an exploratory study. *Autism*, *12*(5), 557-573. https://doi.org/10.1177/1362361308096403

Streiner, D. L., & Norman, G. R. (1995). *Health Measurement Scales: A Practical Guide To Their Development and Use* (2nd ed.). Oxford University Press.

Streiner, D. L., & Norman, G. R. (2008). *Health measurement scales: A practical guide to their use* (4th ed.). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199231881.001.0001

Taylor, B. P., Liu, J., Mowrey, W., Eule, E., Bolognani, F., & Hollander, E. (2022). The Montefiore-Einstein Rigidity Scale-Revised (MERS-R): Development, administration, reliability, and validity in child and adult Autism Spectrum Disorder (ASD). *Journal of Psychiatric Research*, *147*, 142-147. https://doi.org/10.1016/j.jpsychires.2021.12.055

Thissen, D. (2000). *Reliability and measurement precision* (2nd ed.). Lawrence Erlbaum Associates.

Uljarevic, M., Arnott, B., Carrington, S. J., Meins, E., Fernyhough, C., McConachie, H., Le Couteur, A., & Leekam, S. R. (2017). Development of restricted and repetitive behaviors from 15 to 77 months: Stability of two distinct subtypes? *Developmental Psychology*, *53*(10), 1859-1868. https://doi.org/10.1037/dev0000324

Uljarević, M., Frazier, T. W., Jo, B., Scahill, L., Youngstrom, E. A., Spackman, E., Phillips, J. M., Billingham, W., & Hardan, A. Y. (2022). Dimensional Assessment of Restricted and Repetitive Behaviors: Development and Preliminary Validation of a New Measure. *Journal of American Academy of Child and Adolescent Psychiatry.* https://doi.org/https://doi.org/10.1016/j.jaac.2022.07.863

Uljarevic, M., Frazier, T. W., Phillips, J. M., Jo, B., Littlefield, S., & Hardan, A. Y. (2019). Mapping the Research Domain Criteria Social Processes Constructs to the Social Responsiveness Scale. *Journal of the American Academy of Child and Adolescent Psychiatry*, *58*(10S), S311. https://doi.org/10.1016/j.jaac.2019.07.938

Uljarevic, M., Frazier, T. W., Phillips, J. M., Jo, B., Littlefield, S., & Hardan, A. Y. (2020). Quantifying Research Domain Criteria Social Communication Subconstructs Using the Social Communication Questionnaire in Youth. *J Clin Child Adolesc Psychol*, 1-11. https://doi.org/10.1080/15374416.2019.1669156

Uljarevic, M., Jo, B., Frazier, T. W., Scahill, L., Youngstrom, E. A., & Hardan, A. Y. (2021). Using the big data approach to clarify the structure of restricted and repetitive behaviors across the most commonly used autism spectrum disorder measures. *Mol Autism*, *12*(1), 39. https://doi.org/10.1186/s13229-021-00419-9

Uljarević, M., Jo, B., Frazier, T. W., Scahill, L., Youngstrom, E. A., & Hardan, A. Y. (2021). Using the big data approach to clarify the structure of restricted repetitive behaviors across the most commonly used autism spectrum disorder measures. *Molecular Autism*.

Wetherby, A. M., & Prizant, B. M. (2003). *Communication and Symbolic Behavior Scales (CSBS), Normed Edition.* Paul A. Brookes Publishing Co.

Wing, L., Leekam, S. R., Libby, S. J., Gould, J., & Larcombe, M. (2002). The Diagnostic Interview for Social and Communication Disorders: background, inter-rater reliability and clinical use. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(3), 307-325. https://doi.org/10.1111/1469-7610.00023

Table 1. Total scale reliability estimates for autism diagnostic and symptom measures.

	Number of Items	Internal Consistency Cronbach's α (95% CI)	Average Corrected Item-Total r	Model Reliability McDonald's ω	IRT Theta Range (reliability ≥.70)
Diagnostic Observation					
ADOS – Module 1 ("few to no words")	14	.83 (.81-.85)	.50	.83	-3.2 to +1.1
ADOS – Module 1 ("some words")	14	.92 (.91-.92)	.67	.93	-2.0 to +2.2
ADOS – Module 2	14	.87 (.86-.88)	.56	.88	-2.1 to +3.4
ADOS – Module 3	14	.84 (.83-.85)	.51	.85	-2.1 to +4.5
ADOS – Module 4	15	.90 (.89-.91)	.60	.91	-1.5 to +2.9
Diagnostic Interview					
ADI-R	42	.92 (.91-.92)	.48	.91	-2.2 to +1.6
DISCO	48	.85 (.83-.88)	.35	.83	-2.8 to +3.2
Parent-Report					
SRS	65	.97 (.96-.97)	.59	.97	-3.0 to +4.5
SCQ	39	.90 (.90-.91)	.43	.90	-1.4 to +3.0
ASDQ	39	.95 (.94-.95)	.56	.97	-2.8 to +5.6

Note. ADOS module 1 "few to no words" N=1299; ADOS module 1 "some words" N=2481; ADOS module 2 N=1620; ADOS module 3 N=4932; ADOS module 4 N=1706; ADI-R N=1929, DISCO N=272, SRS N=16755; SCQ N=6214; ASDQ N=1467; RBS-R N=5299; CRI-R N=3031.

Table 2. Social communication / interaction (SCI) and restricted / repetitive behavior (RRB) subscale reliability estimates for autism diagnostic and symptom measures.

	Number of Items	Internal Consistency Cronbach's α (95% CI)	Average Corrected Item-Total r	Model Reliability McDonald's ω	IRT Theta Range (reliability ≥.70)
Social Communication / Interaction					
ADOS – Module 1 ("few to no words")	10	.84 (.82-.86)	.56	.84	-3.2 to +1.0
ADOS – Module 1 ("some words")	10	.93 (.92-.93)	.71	.92	-1.9 to +2.0
ADOS – Module 2	10	.86 (.85-.87)	.59	.87	-1.9 to +3.1
ADOS – Module 3	10	.86 (.85-.87)	.58	.87	-1.7 to +4.2
ADOS – Module 4	10	.90 (.89-.91)	.66	.90	-1.2 to +2.5
ADI-R	26	.94 (.94-.95)	.66	.95*	-2.1 to +1.5
DISCO	24	.83 (.80-.86)	.45	.84	-1.8 to +2.8
SRS	53	.96 (.96-.97)	.56	.96	-3.0 to +4.6
SCQ	27	.90 (.90-.91)	.50	.91	-1.3 to +2.4
ASDQ	17	.93 (.92-.94)	.65	.93	-2.0 to +4.8
SSDS	40	.94 (.92-.95)	.52	.94	-4.9 to +5.3
Restricted, Repetitive Behavior					
ADOS – Module 1 ("few to no words")	4	.56 (.54-.58)	.35	.56	-1.8 to -0.4
ADOS – Module 1 ("some words")	4	.67 (.66-.69)	.45	.67	-0.9 to +1.1
ADOS – Module 2	4	.69 (.68-.70)	.47	.69	-0.9 to +1.3
ADOS – Module 3	4	.56 (.55-.57)	.35	.57	-2.5 to +3.5
ADOS – Module 4	5	.68 (.67-.69)	.48	.71	-0.4 to +2.8
ADI-R	16	.69 (.67-.71)	.29	.62	-0.9 to +2.0
DISCO	24	.76 (.72-.80)	.30	.84	-3.4 to +2.1
SRS	12	.93 (.92-.93)	.69	.93	-1.6 to +2.9
SCQ	11	.84 (.84-.85)	.53	.85	-0.5 to +2.2
ASDQ	18	.94 (.93-.94)	.65	.94	-2.0 to +4.8
RBS-R*	35	.94 (.94-.95)	.56	.95	-1.7 to +4.8
DARB	96	.96 (.96-.97)	.46	.96	-3.7 to +6.0

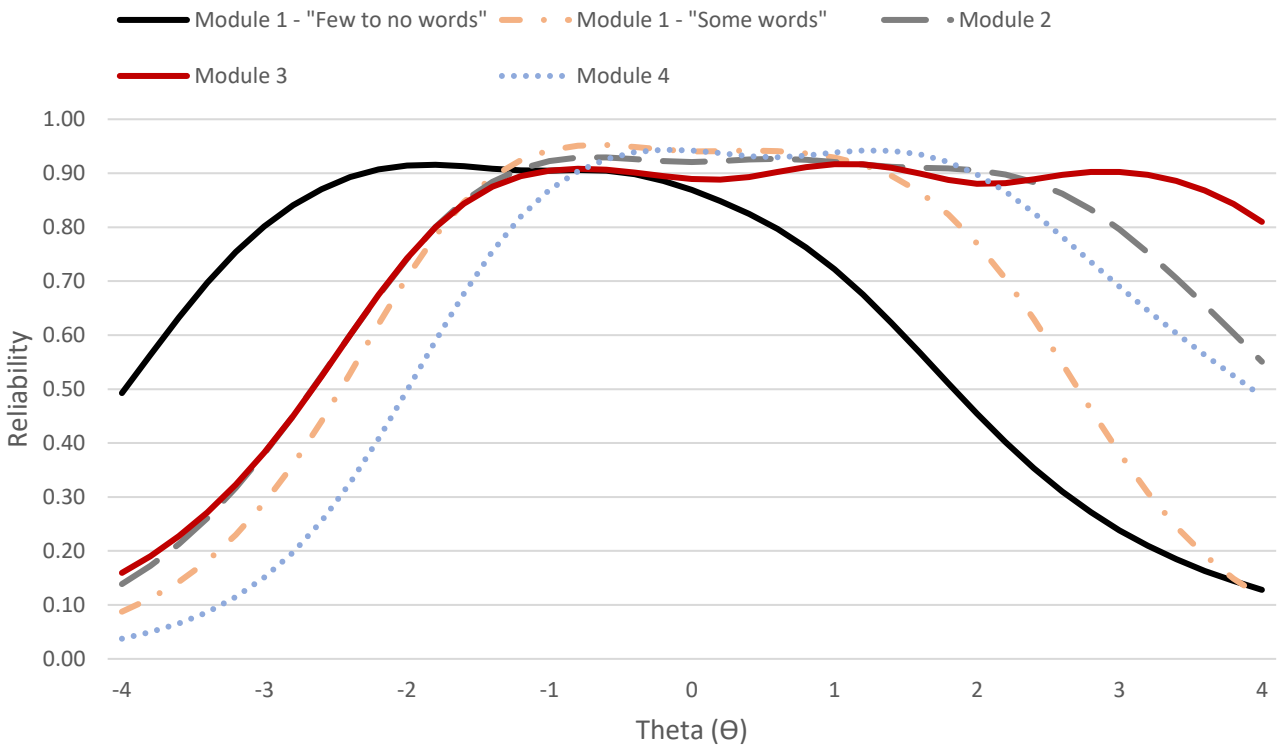
CRI-R	62	.97 (.97-.98)	.60	.97	-3.1 to +4.9
RBQ-2	19	.89 (.88-.89)	.51	.89	-2.3 to +4.2

Note. Sample sizes were the same as for total scales. *Self-injury items were excluded from RBS-R and DARB calculations. MacDonald's Omega for ADI-R total symptoms and SCI symptoms was computed without items 35 (conversation), 46 (attention to voice), and 66 (social distance) due to estimation difficulties when these items were included.

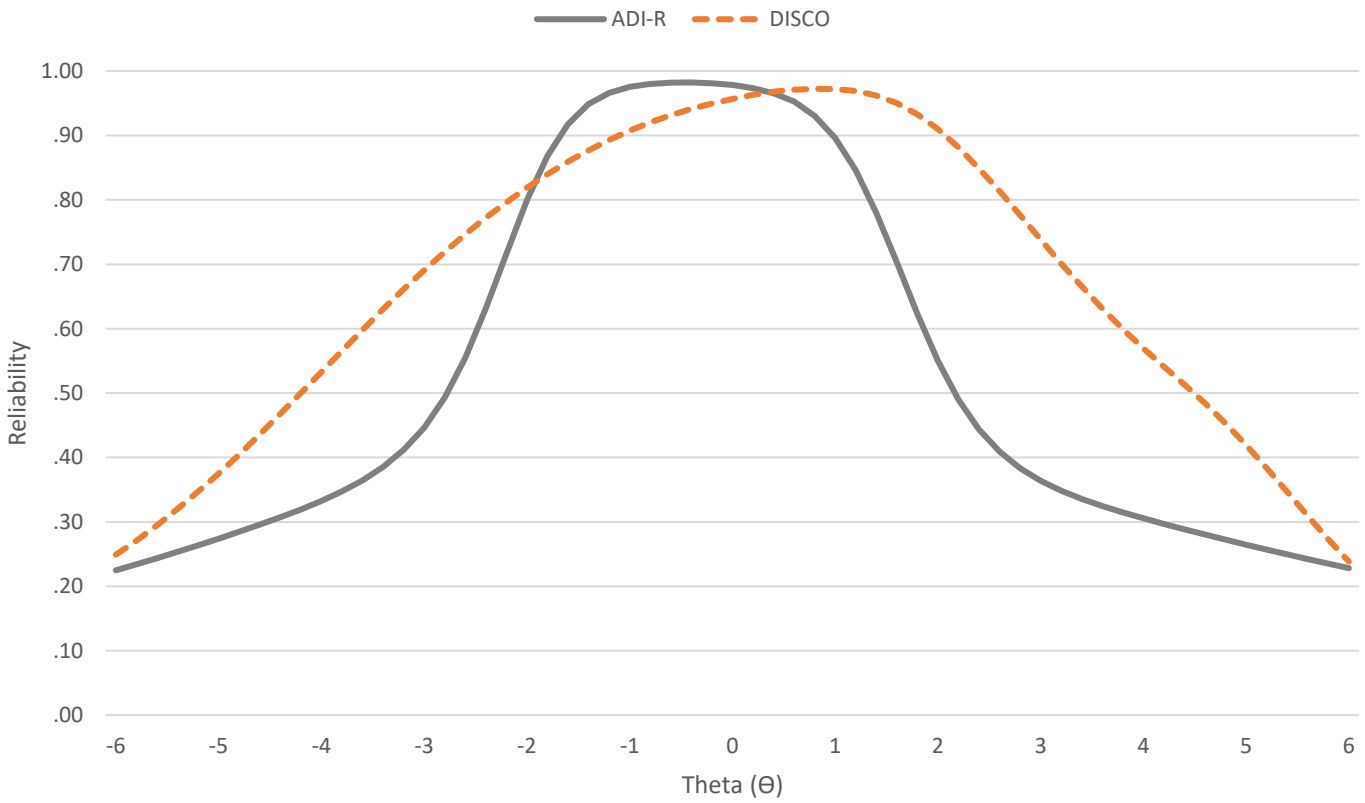
Figure 1. Conditional reliability for total scores across ADOS modules.

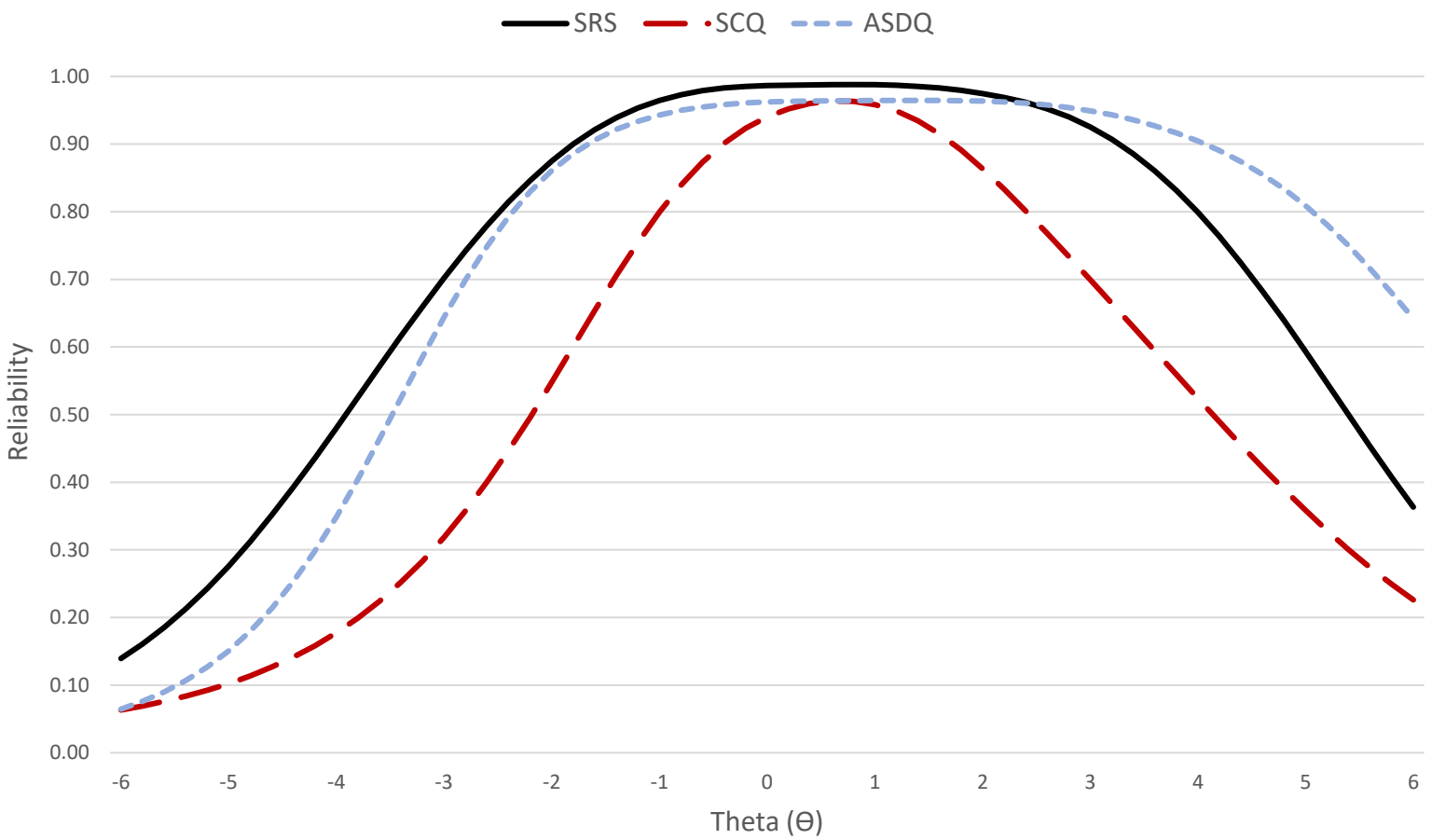
Figure 2. Conditional reliability for total scale scores across parent-interview measures.

Figure 3. Conditional reliability for total scale scores across parent-report questionnaires.



Note. Module 1 – "Few to no words" is shifted left relative to other modules. This could in part reflect the fact that 93% of the children administered this module met criteria for ASD (n=1209). For the other ADOS modules there was a better balance of ASD versus non-ASD/TD cases (% ASD module 1 "some words" = 73%; module 2=81%; module 3=85%; module 4=59%).


ADI-R
DISCO
1.00
.90
.80
.70
.60
.50
.40
.30
.20
.10
.00
Reliability
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Theta (Θ)


SRS
SCQ
ASDQ
1.00
0.90
0.80
0.70
0.60
0.50
0.40
0.30
0.20
0.10
0.00
Reliability
-6
-5
-4
-3
-2
-1
0
1
2
3
4
5
6
Theta (Θ)