Title

Extending the Usefulness of the Brief Observation of Social Communication Change (BOSCC): Validating the Phrase Speech and Young Fluent Version.

Journal

Journal of Autism and Developmental Disorders, 54(3)

Authors

Sterrett, Kyle
Holbrook, Alison
Kim, So
et al.

Publication Date

2024-03-01

DOI

10.1007/s10803-022-05877-5

Peer reviewed

Journal of Autism and Developmental Disorders
https://doi.org/10.1007/s10803-022-05877-5

ORIGINAL PAPER

Extending the Usefulness of the Brief Observation of Social Communication Change (BOSCC): Validating the Phrase Speech and Young Fluent Version

Katherine Byrne[1] · Kyle Sterrett[1] · Alison Holbrook[1] · So Hyun Kim[2] · Rebecca Grzadzinski[3] · Catherine Lord[1]

Accepted: 13 December 2022

Abstract
The current study investigated the utility of the Brief Observation of Social Communication Change-Phrase Speech Young Fluent (BOSCC-PSYF) as an outcome measure of treatment response by analyzing the measure's psychometric properties and initial validity. The BOSCC coding scheme was applied to 345 administrations from 160 participants diagnosed with autism. Participants included individuals of any age with phrase speech, or individuals under the age of 8 years with complex sentences. All were receiving behavioral intervention throughout the study. Test–retest and inter-rater reliability were good for the Early Communication and Social Reciprocity/Language domains, and fair for the Restricted and Repetitive Behavior domain. Significant changes occurred over time in the Early Communication and Social Reciprocity/Language domains, and Core Total scores. The BOSCC-PSYF may provide a low-cost, flexible, and user-friendly outcome measure that reliably measures changes in broad social communicative behaviors in a short period of time.

Keywords Autism · Social communication · Measurement · Treatment response

Introduction

Individuals with autism spectrum disorder (ASD) are involved in numerous interventions throughout their lifespan, the most common of which are aimed at improving social communicative behaviors (Fuller & Kaiser, 2020; Sandbank et al., 2020). Quantifying and measuring the effectiveness of interventions is essential to understanding and monitoring the development of skills in the individuals involved. However, the field of autism intervention research faces significant limitations in measuring effectiveness, including biases inherent in parent- or clinician-report and the reliability of available measures of change over brief periods of time (Grzadzinski et al., 2020). Furthermore, the lack of a uniform measurement approach across studies complicates the comparison of the effects of various interventions, making it unclear which may be most beneficial for whom and when (Cunningham, 2012). There is a critical need for outcome measures that adequately address the limitations discussed below and that reliably detect changes in the social communicative behaviors of individuals with autism, especially in a short period of time.

Limitations of Previously Used Outcome Measures in Intervention Research

Relying on Parent or Clinician Report has Biases

Outcome measures used in intervention research commonly rely on caregiver or clinician reports which can be problematic due to the likelihood of systematic measurement error, expectancy bias or placebo effects (Anagnostou et al., 2015; Sandbank et al., 2020). The Clinical Global Impression (CGI) rating scales (Busner & Targum, 2007), one of the most commonly used outcome measures in intervention research (Bolte & Diehl, 2013; Toolan et al., 2022), is a subjective measure of relative improvement completed by clinicians. In the case of some behavioral interventions, the CGI is completed by the clinician responsible for delivering

✉ Katherine Byrne
katherinebyrne@g.ucla.edu

[1] University of California, Los Angeles, California, USA

[2] Korea University, Seoul, South Korea

[3] University of North Carolina Chapel Hill, Chapel Hill, NC, USA

Published online: 22 December 2022

treatment, in part because this mirrors typical clinical practice. In medication trials, caregivers or physicians may begin the study blind to treatment condition, and later become "unblinded" as a result of possible side effects the child experienced throughout the course of the study (Wolery & Garfinkle, 2002).

Biases inherent in caregiver- and clinician-report measures can accentuate the appearance of treatment effects, leads to the belief that strong effects are present beyond the more subtle changes in social communicative behaviors which are truly occurring (Grzadzinski et al., 2020). For example, a number of studies have demonstrated that caregiver-rated treatment response was associated with caregiver beliefs regarding allocation of treatment condition, even when no significant differences were found between placebo and intervention groups on objective outcome measures (Guastella et al., 2015; Owley et al., 2001). Furthermore, Jones et al. (2017) found a decrease in parent-reported autism-related behaviors and problem behaviors over an eight-week period when, in fact, no treatment was employed.

Caregiver-reported biases may be attributed to the Rosenthal effect, in which expectations about the outcome of an intervention may affect caregivers' responses. Other caregiver biases include overestimating a child's abilities due to reluctance to acknowledge a child's delays, difficulty recalling a child's developmental milestones, investment in positive outcomes, and paying greater attention to challenging behaviors as opposed to prosocial behaviors, each of which could affect measurement of change (Miller et al., 2017; Nordahl-Hansen et al., 2014; Ozonoff et al., 2011; Zapolski & Smith, 2013). While caregiver report and clinician judgment are important sources of information, reliance on these measures alone limits the interpretation of intervention responses (Miller et al., 2017).

Diagnostic Tools are Not Sensitive to Change in Short Period of Time

Changes in autism-specific behaviors are often measured using diagnostic instruments (Dawson et al., 2010; Green et al., 2022). Yet, diagnostic instruments, such as the Autism Diagnostic Observation Schedule, Second Edition (ADOS-2; Lord et al., 2012), were not intended to be used as outcome measures of short-term treatment response. Rather, they were intended to measure relatively stable constructs over time (Cunningham, 2012).

Thus, diagnostic instruments are typically not sensitive enough to detect subtle changes in a short period of time (Grzadzinski et al., 2020; Owley et al., 2001). While some studies have found significant changes over time in ADOS raw scores, these changes were also evident in treatment-as-usual groups (Green et al., 2022; Gutstein et al., 2007). In other studies that found significant raw score changes, changes were usually not evident over short periods of time and were related to changes in other domains, such as language development, as opposed to changes in the severity of autism features (Esler et al., 2015; Gotham et al., 2012). Furthermore, the use of raw score changes on diagnostic instruments, such as the ADOS-2, must be interpreted with caution due to the influence of age, language level, and verbal IQ on raw scores (Kim et al., 2018). As a result, Calibrated Severity Scores (CSS) were created as a standardized metric of autism symptom severity that is less confounded by changes in general maturity or language development (Gotham et al., 2008).

The ADOS CSS has successfully measured changes in autism symptom severity (Gotham et al., 2008; Grzadzinski et al., 2020). Yet, these changes have been evident over long periods of time (i.e., years as opposed to months; Estes et al., 2015; Gotham et al., 2012; Pickles et al., 2016; Thurm et al., 2015). CSS scores demonstrate high test–retest reliability, meaning that scores are stable in a short period of time (Janvier et al., 2022). Considering that short-term intensive interventions are common for individuals with autism, CSS scores are likely not a useful outcome measure to be used to test their effectiveness.

Finally, diagnostic measures require substantial training to use reliably and are often time-consuming to administer. The amount of time needed, and the level of training required to administer and score these assessments make diagnostic instruments difficult to implement in large scale, multisite studies, especially considering that they must be administered more than once to measure changes over time. An instrument that reliably measures autism-specific behaviors and is more sensitive to subtle changes in a short period of time will be of crucial importance for autism intervention research moving forward (Matson, 2007).

Lack of Uniform Measurement Approach

Autism intervention research has utilized hundreds of disparate outcome measures to test the effectiveness of various treatments (Bolte & Diehl, 2013). There is little consensus regarding which symptoms to target or which tools to use in determining intervention effectiveness (Anagnostou et al., 2015; Sandbank et al., 2020). This is due, at least in part, to the heterogeneity of the type and severity of autism features. For example, deficits in social communication could be considered to include verbal or nonverbal communication delays, trouble developing or maintaining relationships, among many other possible areas of difficulty (Volkmar et al., 2004).

The lack of uniform measurement approach is also the result of the use of study-specific outcome measures that are used to measure specific behaviors, such as joint attention (Green et al., 2022; Kasari et al., 2012; Rogers et al., 2012;

Yoder et al., 2014). In a large-scale review of 195 prospective intervention trials for individuals with ASD, 289 outcome measures were identified (Bolte & Diehl, 2013). Of the 289 measures, 62% were found in only one publication over a 10-year period and 21% of these measures were designed or modified by the research investigator specifically for use in that study.

Study-specific outcome measures are often limited to quantifying the frequency of highly specific behaviors (Kaale et al., 2012), as opposed to capturing changes in broad social communicative behaviors (Spence & Thurm, 2010). These measures are often proximal to the treatment and may reflect learning a task in a specific context, though they are targeted with the hope that improvement of these behaviors will have positive cascading effects on other domains, such as language development (Green et al., 2022). While identifying changes in specific behaviors is important, whether these context-specific behaviors result in generalized gains across broad social communication strategies often goes unmeasured (Sandbank et al., 2020; Yoder et al., 2013). Moreover, behaviors can be operationalized differently across studies, making the comparison of results across outcome measures nearly impossible, even when they appear to measure the same behavior (Wolery & Garfinkle, 2002).

Call for Novel Measures

Reliably measuring changes in social communicative behaviors as a result of intervention has proven especially difficult. These behaviors are often subtle, meaning their measurement must be sensitive enough to capture small, but clinically meaningful changes that indicate measurable improvement and, ideally, predict other positive gains (Anagnostou et al., 2015; Grzadzinski et al., 2020). Expert panels have concluded that existing outcome measures widely used in ASD intervention research are not appropriate intervention response measures without certain modifications (e.g., only appropriate for specific populations, such as young children or those with average or greater IQ), making the use of a uniform outcome measurement approach difficult (Anagnostou et al., 2015; McConachie et al., 2015; Scahill et al., 2015). Moreover, few measures are flexible enough to be available for use across studies or sites. There is currently a call by intervention researchers for novel outcome measures that can reliably detect change, be used across studies, and fill the gaps left behind by the limitations of existing measures (Fletcher-Watson & McConachie, 2017; McConachie et al., 2015).

The Brief Observation of Social Communication Change (BOSCC)

The Brief Observation of Social Communication Change (BOSCC) was developed to provide a blinded, standardized, and efficient method of measuring subtle changes in the social communicative behaviors of individuals with ASD and other neurodevelopmental conditions over relatively short periods of time (i.e., as short as 8–12 weeks). It is a play-based assessment conducted with a participant and play partner, such as a parent or research staff member. It was initially developed using codes from the ADOS, which apply specifically to ASD symptoms, but these codes were modified and expanded upon to examine and measure more subtle social communicative behaviors (Grzadzinski et al., 2016; Lord et al., 2012).

The BOSCC is flexible and standardized, which allows for its use across sites and studies. It is observation-based, using interactions with play partners who may be blinded to treatment or not, and coded by individuals who must be blind to treatment condition and goals, lessening the possibility of bias or placebo effects. The goal of the BOSCC is to address the challenges that intervention research faces in measuring intervention effectiveness by providing a novel, standardized outcome measure that is minimally biased, sensitive to change in short periods of time, easy to code, and flexible enough to be used in a variety of settings, with a variety of populations and research contexts, and by people of all skill levels.

Current State of the BOSCC

The BOSCC has been validated for use with minimally verbal (MV) children with ASD, called the BOSCC-MV (Grzadzinski et al., 2016; Kim et al., 2018; Kitzerow et al., 2016; Nordahl-Hansen et al., 2016). Using a sample of 56 children between the ages of 1–5 years, results demonstrated statistically significant changes in the "Core Total" items as compared to a no-change alternative; ADOS CSS scores over the same period showed no statistically significant changes (Grzadzinski et al., 2016). Furthermore, psychometric properties of the BOSCC-MV showed high to excellent inter-rater reliability and test–retest reliability. Exploratory Factor Analyses (EFA) revealed two underlying factors: Social Communication (SC) and Restricted and Repetitive Interests and Behaviors (RRB). This two-factor structure mapped onto well-known behavioral diagnostic assessments of ASD, such as the ADOS, and generally fits well with diagnostic features of ASD as specified in DSM-5 and ICD-11 (American Psychiatric Association, 2013; World Health Organization, 2019). Since its original

validation study was published, several other studies have corroborated the strong psychometric properties of the BOSCC-MV, its ability to detect changes in a short period of time, and be conducted in a variety of settings with various play partners (Gengoux et al., 2019; Green et al., 2022; Kim & Lord, 2010; Kitzerow et al., 2016; Nordahl-Hansen et al., 2016).

During the development of the BOSCC, pilot testing was conducted to determine whether the BOSCC-MV administration and coding scheme could be used with older and more verbal children. This developmental stage revealed that the BOSCC-MV was unable to identify changes in children over the age of 8; scores were variable over time and not related to intervention status in children receiving a range of treatments from four different sites. Furthermore, changing the coding scheme without modifying the administration was insufficient. Both the context in which the interaction occurred, and the codes used for scoring, needed to be altered for changes to be detected. We used this pilot data collected during the developmental stage of the BOSCC to extend the work already conducted on the BOSCC-MV by creating new contexts (e.g., administration materials appropriate to developmental level) and a new coding scheme more appropriate for older and more verbal children with ASD.

Current Study

The aim of the present study is to determine the utility of the BOSCC as an outcome measure of intervention response in a sample of young autistic children who are verbally fluent, or autistic individuals of any age who consistently and spontaneously use phrase speech. This version is called the BOSCC Phrase Speech/Young Fluent (PSYF). More specifically, this paper will (1) determine items for inclusion in the final BOSCC-PSYF coding scheme and its algorithm, (2) analyze the factor structure of the measure by exploring the relationships between items, (3) examine the initial psychometric properties, including inter-rater and test–retest reliability, and (4) provide evidence of its utility as a measure of change by examining changes in scores over time in autistic individuals receiving various behavioral interventions and compare these changes to changes in scores over the same period of time in other behavioral and parent-report measures.

Method

Participants

Participants included 160 English-speaking children between the ages of 2–10 years (M_{age} = 4.9 years) with a documented diagnosis of autism spectrum disorder (ASD). Sex at birth was reported by caregivers in which 82% were male and 18% were female. The caregiver-reported racial and ethnic identities and parental education can be found in Supplement Table 1. All participants had language abilities suitable for the PSYF administration. Specifically, the BOSCC-PSYF is appropriate for: (1) individuals of any age who use phrase speech, (defined as spontaneous, non-rote two-word phrases which include both a noun and a verb, such as "want ball"), or (2) children with fluent language (defined as multiclausal sentences with flexible grammatical and sentence structures) who are younger than 8 years of age. All administrations and scoring were completed in English.

All participants were actively receiving behavioral intervention at the time of participation, though the types of intervention varied. For example, some children were enrolled in a short-term intensive day program (approximately 35 h per week for 16 weeks), while others were enrolled in various less-intensive (at least once per week) long-term, ABA-style programs (approximately 10–20 h per week). For the purposes of this validation study, comparison of specific treatment effects across the various interventions will not be explored.

Procedure

Participants (n = 160) were recruited from three sources. The first was a short term, intensive partial hospitalization program for children with ASD (n = 30). Two other sources (n = 25, n = 105) were research studies that took place on UCLA's and Weill Cornell Medicine's campuses. All participating families signed informed consent forms approved by the participating institutions' Institutional Review Board before participating in this study.

Whenever possible, each source administered the BOSCC at (at least) two timepoints, along with other diagnostic, cognitive, and adaptive behavior measures. Between one and six videos were available for each child (M = 2.05 videos, SD = 0.57). Sixteen participants were lost to follow up; thus, two or more videos were available for 144 participants. All BOSCCs were conducted in an in-person laboratory setting with a research assistant. Participants with only a single BOSCC datapoint available were retained for purposes of psychometric analyses of validity (e.g., factor analyses) and reliability (e.g., inter-rater), but not in analyses of change.

Measures

Brief Observation of Social Communication Change (BOSCC-PSYF)

The BOSCC was developed as an intervention response measure of social communicative and other behaviors

associated with autism. The BOSCC is a 12-min, videotaped play interaction between an individual and a play partner. The BOSCC can be administered in a lab, clinic, or home setting, though it is essential that this context and the type of play partner remain consistent across each observation. The play interaction is conducted with a standardized set of toys that is designed to offer opportunities for active participation and various levels of play between the participant and play partner. It was designed to be easy to administer and, thus, can be implemented with caregivers, research assistants or clinicians who receive minimal instruction, as long as someone of the same role administers the BOSCC at all time points for a given participant. For purposes of this study, all BOSCC administrations were implemented by clinicians (not administering treatment) or research assistants.

Coding Procedures The BOSCC videos are split into two 6-min segments (Segment A and Segment B) which are each watched and coded twice. The BOSCC-PSYF includes 17 items that are scored on a 6-point scale ranging from 0 ("atypical behavior not present") to 5 ("atypical behavior present and significantly impairs functioning"). The PSYF items are comprised of 10 items from the BOSCC-MV (Grzadzinski et al., 2016) modified to fit the social communicative behaviors of children with greater language skills, and three novel items (i.e., verbal exchanges, offering information, and stereotyped speech). These 13 items are averaged across the two segments and summed to create a total BOSCC score. The final four items, which are not included in the scoring process, are used as indicators of mood/disposition and other co-occurring behaviors sometimes seen in ASD (i.e., social engagement in play activities/interaction, activity level, disruptive behaviors, anxious behaviors). These items are scored to determine the validity of the administration and to provide quantitative information about non-ASD specific behaviors that are seen frequently in children which may impact social communicative behaviors (e.g., oppositional behaviors). High scores suggest that difficulties in the assessment may be exacerbated by issues other than those related to ASD symptoms.

The BOSCC coding scheme employs empirically based decision trees for ease of use (Grzadzinski et al., 2016). Each decision tree contains detailed information regarding the frequency and quality of specific behaviors. At each branch, the coders answer a specific question (e.g., yes or no, frequency, consistency) concerning the child's behavior on that specific item (e.g., eye contact) until they arrive at a numerical code. Videos were coded by one psychologist, one postdoctoral researcher, four graduate students, and one research assistant. All coders obtained reliability before beginning the coding process (as described in Grzadzinski et al., 2016) and were blind to timepoint and treatment status. A random sub-sample of 54 videos were chosen to determine inter-rater reliability.

Additional Measures

Other diagnostic, cognitive, and adaptive functioning assessments were collected from all participants as part of their involvement in various intervention programs. The battery of assessments each participant received varied depending upon which source the participant was recruited from; however, whenever possible, all participants were administered at least one measure of ASD symptom severity, one cognitive test, and one measure of adaptive functioning. As a result of COVID restrictions, this was not possible for everyone. Thus, participation within each measure (described below) was variable. The results of these assessments were included in this study for purposes of comparing rates of change in these measures to that of the BOSCC.

ASD Symptom Severity ASD symptom severity was measured in two ways: The Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012) and the Social Responsiveness Scale (SRS-2; Constantino & Gruber, 2012). The ADOS is a standardized diagnostic measure comprised of both structured and semi structured tasks used to assess symptoms of ASD. The ADOS-2 provides a total Calibrated Severity Scores (CSS) that indicates severity of autism symptoms during the assessment and can be used to compare symptom severity levels across individuals of varying developmental levels. Domain severity scores are also provided for social affect (CSS SA) and restricted and repetitive behaviors (CSS RRB) domains (Gotham et al., 2008). The ADOS-2 was administered to 88 of our participants at one time point. Twenty-five participants received Module 2, which is appropriate for individuals of any age who speak in phrases but are not verbally fluent. The remaining 63 participants received Module 3, which is appropriate for verbally fluent children and adolescents. ADOS-2 scores were collected for 61 individuals at a second timepoint, which allowed for analysis of change in scores over time. None of the clinicians who administered the ADOS-2 were involved in the coding of the BOSCC, allowing coders to be completely blind to the participant and timepoint.

The SRS is a parent-report measure that identifies the presence and severity of social impairment in individuals with ASD. The SRS was collected for 51 participants at one time point, and 18 participants at two time points.

Cognitive Functioning Verbal and nonverbal cognitive functioning were assessed using a variety of measures, including the Mullen Scales of Early Learning (MSEL; Mullen, 1995), the Differential Ability Scales (DAS-2; Elliot et al., 2018), the Wechsler Preschool and Primary Scale of Intel-

ligence (WPPSI-IV; Wechsler, 2012), the Peabody Picture Vocabulary Test (PPVT-IV; Dunn & Dunn, 2007), and the Ravens Progressive Matrices (Raven et al., 2000). The MSEL was collected for 30 children, the DAS-2 for 62 children, the WPPSI-IV for 19 children, and the PPVT-IV and Ravens for 25 children. Cognitive measures were only collected at one timepoint, so analysis of change in scores over time was not conducted.

Adaptive Functioning The Vineland Adaptive Behavioral Scales (VABS-3; Sparrow et al., 2016) is a measure of adaptive functioning that provides standard scores in the domains of Communication, Daily Living Skills, Socialization, and Motor Skill, as well as an overall Adaptive Behavior Composite. Only the Communication and Socialization domains were used for purposes of this study. A combination of the comprehensive interview form and the caregiver report form was used. The VABS was administered to 151 participants at one timepoint and 73 at a second timepoint, allowing for analysis of change in scores over time.

Data Analysis

All analyses were carried out using R version 4.0.2 (R Core Team, 2021)–the Lavaan package was used to estimate all factor analysis models (Rosseel, 2012).

Item Level Descriptive Information

Several versions of the BOSCC-PSYF item level coding schemes were tested with the goal of achieving as close as possible to uniform distribution of codes across all items, though a normal distribution was also acceptable. Item level codes were re-written over several iterations until near-flat distributions were achieved (Grzadzinski et al., 2016). Consistent with the BOSCC-MV, we did not expect a uniform distribution for items related to restricted and repetitive interests and behaviors (RRBs) because the presentation of these behaviors is extremely heterogeneous across individuals and the short duration of the BOSCC assessment may not allow for consistent presentation of these behaviors (Kim & Lord, 2010).

Factor Structure

A multi-step process was undertaken to systematically evaluate the factor structure of the BOSCC-PSYF. While the factor structure of the BOSCC-MV has been validated, an exploratory approach was taken here due to differences between the coding schemes and the intended populations of the two versions (Grzadzinski et al., 2016).

Model fit was determined using the Tucker-Lewis Index (TLI), Comparative Fit Index (CFI), and root mean squared error of approximation (RMSEA). Values closer to "1" using the TLI and CFI and values closer to "0" using the RMSEA indicate better model fit. Recommended cutoffs for well-fitting models are typically greater than 0.95 for the TLI and CFI and ≤ 0.06 for the RMSEA, though these cutoffs may be overly exclusive in small samples (Hu & Bentler, 1999).

Using baseline data, scree and parallel plots were generated on which to base decisions of the number of factors to extract. Subsequently, exploratory structural equation models (ESEM) were fit to the data. This involved fitting an Exploratory Factor Analysis (EFA) model with an oblimin rotation using maximum likelihood estimation, and testing one-, two-, three-, and four-factor solutions. This was followed by a Confirmatory Factor Analysis (CFA) using the cross loadings from the EFA as the starting point for estimation. Factors were allowed to covary in these models. ESEM was chosen to balance the drawbacks of overly restrictive CFA models (e.g., cross loadings between factors are typically set to zero) while allowing for modifications and extensions (Asparouhov & Muthen, 2009; Marsh et al., 2014).

Lastly, a traditional CFA model was fit to the full dataset based on the observed factor structure from the ESEM to confirm that the factor structure holds up using the full dataset. Factors were allowed to covary. TLI, CFI, and RMSEA were used to evaluate model fit.

Longitudinal Measurement Invariance

Four steps were taken to evaluate the measurement invariance of the BOSCC-PSYF over time. These four steps were: (1) configural invariance, which tests whether the factor structure is comparable across entry and exit; (2) metric invariance, which tests whether the items load onto the same factors across entry and exit; (3) scalar invariance, which compares the intercepts across entry and exit; and (4) strict invariance, which tests whether the residual variances are comparable across entry and exit. Nested models were tested using chi-square difference tests; non-significant tests indicate invariance across the models that were tested.

Reliability

Test–retest and inter-rater reliability were analyzed. Test–retest reliability was estimated from 26 participants who had a second BOSCC conducted within one-month of each other. Inter-rater reliability was estimated from 54 double coded videos. Absolute agreement was assessed using two-way random effects models. Inter-rater and test–retest reliability results are described for each domain of the BOSCC-PSYF derived from the factor analyses and the Core Total.

Change Analyses

Following procedures from the analyses of the BOSCC-MV (Grzadzinski et al., 2016; Kim & Lord, 2010), first, paired sample *t*-tests were used to compare BOSCC Core Total and domain scores and other behavioral measures (i.e., VABS Communication and Socialization, ADOS CSS and SRS Total *T*-Scores) from the first available timepoint to the last available timepoint. This raw score difference was also standardized as a Cohen's *d* effect size. Next, individual growth models were fit separately using BOSCC Core Total and domain scores, as well as the other behavioral measures with sufficient data (i.e., VABS Communication and Socialization, ADOS CSS and SRS Total *T*-Scores). This involved fitting a linear regression separately for each participant using participant's age at the time of assessment as the independent variable to generate an average rate of change per month. Rate of change over 4.5 months, the average length of time between the entry and exit appointments in our sample, was reported. To be consistent with prior analyses, these rates of change were also converted to expected change scores over 6 months (Grzadzinski et al., 2016; Kim & Lord, 2010). These rates were then divided by the standard deviation of the measure at entry to generate an effect size comparable to a Cohen's *d*. Due to the wide range of age and cognitive abilities of the participants, we also ran a linear mixed effect model to evaluate whether age, verbal IQ (VIQ) and nonverbal IQ (NVIQ) at entry were related to or moderated change in BOSCC scores over time.

Lastly, again following the procedures of Grzadzinski et al. (2016) and Kim & Lord (2010), change of greater than or equal to 8 points on VABS Standard Scores and SRS Total *T*-Scores (half a standard deviation) and greater than or equal to 1 point on ADOS CSS Scores (1 standard deviation) were used to classify "responders" on each of the behavioral measures (i.e., VABS Communication and Socialization, ADOS CSS and SRS Total *T*-Scores). After response status was determined, independent samples *t*-tests were used to determine whether "responders" and "non-responders" for each measure differed in the amount of change on BOSCC domain and Core Total Scores.

Results

Item Level Descriptive Information

The distribution of BOSCC-PSYF codes (averaged for Segment A and B) across 14 out of the 17 items in the final version of the BOSCC-PSYF can be found in Supplement 1. Activity Level, Disruptive Behavior/Irritability, and Anxious Behaviors are not depicted because these items were rarely observed and scored; however, these items provide useful information in determining whether the BOSCC administration is a representative sample of the child's behavior. Thus, these codes were retained in the final coding scheme.

Factor Structure

Two sets of ESEM models were tested; the first included all items. The scree and parallel plots indicated a four-factor solution best fit the data (RMSEA = 0.045, CFI = 0.987, TLI = 0.971). These factors could be described as: (1) Early Communication, (2) Social Reciprocity/Language, (3) Play and (4) Restricted and Repetitive Behaviors and Interests (RRBs).

Due to concerns about over-specifying and mis-specifying the model driven by substantive and statistical concerns, such as a negative variance estimate for the "Play with Objects" item, a second ESEM model was fit excluding the "Play with Objects" item. The scree and parallel plots suggested a three-factor solution best fit the data. One- two- and three-factor solutions were tested. The best fitting solution was the three-factor solution (RMSEA = 0.074, CFI = 0.962, TLI = 0.929). Parameter estimates for the three-factor model are included in Table 1. These factors could be described as: (1) Early Communication, (2) Social Reciprocity/Language, and (3) RRBs. The "Engagement in Play with Others" item loaded onto the Social Reciprocity/Language factor in the absence of the "Play with Objects" item. The fit of the final model was adequate without the inclusion of the "Play with Objects" item. However, due to the clinical value that the item provides and the possibility of play skills improving with intervention, a decision was made to keep this item in the final codes but not include it in the measure's factor

Table 1 Exploratory structural equation model factor loadings without play

	Standardized loading (SE)		
	Factor 1	Factor 2	Factor 3
Eye contact	− 0.058	**0.768**	0.013
Facial expressions	0.151	**0.569**	0.160
Gesture	0.320	**0.392**	− 0.161
Integration of non-verbal communication	0.009	**0.985**	− 0.018
Quality of social overtures	**0.672**	0.141	0.069
Quality of social responses	**0.725**	0.073	0.109
Verbal exchanges	**0.933**	− 0.060	− 0.015
Offering information	**0.758**	0.094	− 0.076
Engagement in play with others	**0.599**	0.058	0.125
Stereotyped speech	0.223	− 0.043	**0.267**
Sensory behaviors	0.104	0.049	**0.340**
Mannerisms	− 0.214	0.013	**0.411**
Repetitive behaviors	0.086	0.032	**0.667**

Bolded items indicate which factor the items load onto

structure. Figure 1 depicts the final items, domains, and Core Total.

The CFA model adequately fit the data (CFI = 0.937, TLI = 0.921 and RMSEA = 0.076). Item loadings across the three factors were high apart from some RRB items. The factor loadings of the items onto the Early Communication factor ranged from 0.63 to 0.95, 0.71 to 0.84 for the Social Reciprocity/Language factor, and 0.25 to 0.58 for the RRB factor. Model parameters across items are include in Table 2.

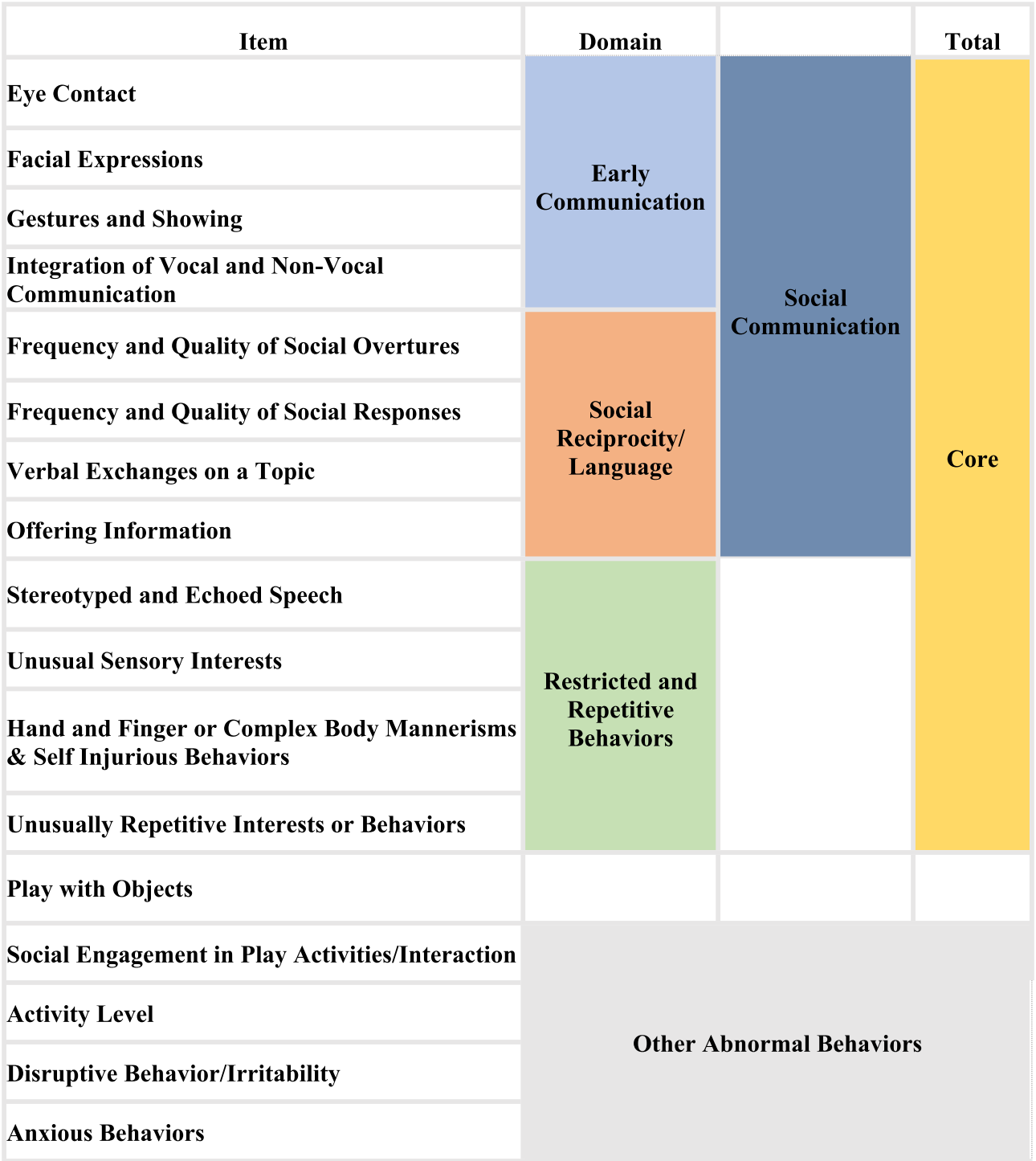

Fig. 1 Visual depiction of BOSCC items, domains, and totals

Table 2 Final confirmatory factor analysis parameter estimates

	Standardized loading (SE)		
	Factor 1	Factor 2	Factor 3
Eye contact	.790		
Facial expressions	.741		
Gesture	.630		
Integration of non-verbal communication	.951		
Quality of social overtures		.825	
Quality of social responses		.837	
Verbal exchanges		.839	
Offering information		.749	
Engagement in play with others		.707	
Stereotyped speech			.414
Sensory behaviors			.488
Mannerisms			.253
Repetitive behaviors			.577

Measurement Invariance

Across time, there was evidence of configural, metric, scalar, and strict invariance. This suggests that the factor structure, item loadings, intercepts and residuals do not change substantially when measuring individuals across time. This is an indication that the BOSCC-PSYF measures the same factor structure across timepoints; thus, comparing mean scores across time is appropriate. Comparisons of the model fit statistics are provided in Table 3.

Test–Retest Reliability

Test–retest reliability was estimated from 26 videos. Adequate test–retest reliability for one item, "Engagement in Play with Others" was not able to be reached, likely due to the variable nature of the construct it measures. Additionally, when coding this item, coders reported anecdotally that the score seemed to fluctuate based on the child's mood and disposition during the assessment. As a result, this item was removed from the algorithm and added as an "Other Abnormal Behaviors" code.

Overall, test–retest reliability was good for the Early Communication and Social Reciprocity/Language domains and fair for the RRB domain. The Intraclass Correlation Coefficient (ICC) value for the Early Communication domain was 0.74, 95% CI [0.50, 0.87], 0.68, 95% CI [0.40, 0.84] for the Social Reciprocity/Language domain and 0.51, 95% CI [0.15, 0.75] for the RRB domain. The ICC value for the combined Social Communication domain was 0.75, 95% CI [0.51, 0.88] and 0.71, 95% CI [0.45, 0.86] for the Core Total score.

Inter-Rater Reliability

Inter-rater reliability was estimated from 54 double-coded videos. Inter-rater reliability was good for the Early Communication domain and Social Reciprocity/Language domain and fair for the RRB domain. The ICC value for the Early Communication domain was 0.85, 95% CI [0.75, 0.91], 0.89, 95% CI [0.81, 0.94] for the Social Reciprocity/Language domain and 0.60, 95% CI [0.38, 0.75] for the RRB domain. The ICC value for the combined Social Communication domain and Core Total was 0.91, 95% CI [0.85, 0.95] and 0.90, 95% CI [0.83, 0.94], respectively.

Change Analysis and Validity

Paired *t*-tests indicated statistically significant decreases (improvement in symptoms) in scores from entry to exit on the Early Communication domain (M = − 0.71 [0.17, 1.25], $t(111), = 2.61$, $p = 0.01$, Cohen's $d = -0.25$), the Social Reciprocity/Language domain (M = − 1.17 [0.55, 1.78], $t(111), = 3.77$, $p < 0.01$, Cohen's $d = -0.36$), the combined Social Communication domain (M = − 1.87 [0.94, 2.81], $t(111), = 3.97$, $p < 0.01$, Cohen's $d = -0.38$), and the Core Total (M = − 1.87 [0.82, 2.92], $t(111), = 3.53$, $p < 0.01$, Cohen's $d = -0.33$). There were no statistically significant changes in the RRB domain. Over the same length of time, there was no statistically significant change in ADOS CSS scores (M = − 0.486 [− 0.17, 1.15], $t(36) = 1.48$, $p = 0.49$, Cohen's $d = -0.24$), in VABS Communication scores (M = 0.46 [− 2.67, 3.61], $t(29) = 0.30$, $p = 0.76$, Cohen's $d = 0.05$) or in VABS Socialization scores (M = − 1.23 [1.54, 4.02], $t(29) = 0.91$, $p = 0.37$, Cohen's $d = 0.17$). There was statistically significant change in SRS Total *T*-Scores (M = − 6.16 [1.75, 10.58], $t(11) = 3.07$, $p = 0.01$, Cohen's $d = 0.89$).

Results from the individual growth models indicated that the average rates of change on the BOSCC over 4.5 months

Table 3 Measurement invariance model fit

	DF	AIC	BIC	X^2	X^2 Difference	p-value
Configural invariance	124	11977	12299	240.80		
Metric invariance	134	11965	12249	249.15	8.3455	.5951
Scalar invariance	144	11953	12199	257.01	7.8650	.6420
Strict invariance	157	11946	12143	276.76	19.7447	.1018

was small in the Early Communication domain (Cohen's $d = -0.28$, 95% CI [− 0.65, 0.09], see Fig. 2a), and greater for the Social Reciprocity/Language domain (Cohen's $d = -0.40$, 95% CI [− 0.65, − 0.15], see Fig. 2b), the combined Social Communication domain (Cohen's $d = -0.38$, 95% CI [− 0.63, − 0.13], see Fig. 2c), and the Core Total (Cohen's $d = -0.39$, 95% CI [− 0.65, − 0.13], see Fig. 2d), with larger changes when comparisons were made for 6 months. The average rate of change over 4.5 months was smaller in the VABS Communication Standard Score (Cohen's $d = -0.07$, 95% CI [− 0.23, − 0.08]), the VABS Socialization Standard Score (Cohen's $d = -0.09$, 95% CI [− 0.25, 0.06] and ADOS CSS (Cohen's $d = -0.32$, 95% CI [− 0.80, 0.16]) than the BOSCC-PSYF. The average rate of change over 4.5 months was larger for the SRS Total score (Cohen's $d = -1.33$, 95% CI [− 2.31, − 0.36]) than the average rate of BOSCC-PSYF change (though, as mentioned previously, this finding should be interpreted with caution due to the small sample of SRS scores at exit).

Moderating Variables

The results of the linear mixed models suggested that children's chronological age ($t = -2.00$, $p = 0.049$) though not VIQ ($t = -1.78$, $p = 0.079$) nor NVIQ ($t = 0.203$, $p = 0.84$) was related to Early Communication scores, where younger children had higher (more impaired) scores on the Early Communication domain. Children's chronological age ($t = -2.49$, $p = 0.015$) and VIQ ($t = -3.62$, $p < 0.001$), though not NVIQ ($t = 1.46$, $p = 0.149$) were related to Social Reciprocity/Language scores, where younger children and children with lower VIQ's had higher Social Reciprocity/Language scores. The Combined Social Communication Total scores were related to both children's age ($t = -2.44$, $p = 0.02$) and VIQ ($t = -2.97$, $p = 0.004$), though not NVIQ ($t = 0.93$, $p = 0.35$). VIQ ($t = -4.11$, $p < 0.001$), though neither chronological age ($t = -1.84$, $p = 0.07$) nor NVIQ ($t = 0.99$, $p = 0.32$), was related to Core Total scores. Most relevant to the use of the BOSCC as a measure of intervention response, change over time in all domains and for Core Total scores was not moderated by age, VIQ or NVIQ.

Response Status

T-tests comparing the amount of change in BOSCC domain and Core Total scores by response status indicated that the individuals who were considered "responders" on the SRS Total Score demonstrated significantly more change in the BOSCC Early Communication ($t(16) = 2.34$, $p = 0.03$) and combined Social Communication domains than SRS "non-responders" ($t(17) = 2.21$, $p = 0.04$). There were no statistically significant differences on any BOSCC domain score for VABS Communication, VABS Socialization or ADOS CSS responders.

Discussion

Results from our analyses confirm prior literature that the BOSCC is a promising outcome measure of intervention response (Grzadzinski et al., 2016). The BOSCC-PSYF, which is intended to be used with individuals of all ages who speak in flexible phrases or children under the age of 8 who speak in complex sentences, has been demonstrated to be sensitive to subtle changes in social communicative behaviors over a brief period of time. To the best of our knowledge, the BOSCC is the first brief, observation-based outcome measure of intervention response which measures a range of broad social communicative behaviors that is sensitive to changes in a short period of time. The BOSCC can be conducted by individuals of any skill level, including caregivers, therapists, naïve research assistants or highly trained clinicians.

A three-factor structure proved to be the best fit to the data. The items relating to broad social communicative behaviors were split into two domains—one including nonverbal and early communicative behaviors, the second including behaviors that relate to social reciprocity and are mostly based in language skills. The three-factor structure of the BOSCC-PSYF diverges from the two-factor structure evident in the BOSCC-MV (Grzadzinski et al., 2016), but is similar to the factor structure in the ADOS Module 3 described in Zheng et al. (2021). The three-factor structure of the BOSCC allows for researchers to decide whether to examine the two Social Communication domains separately or together, depending on the goals of treatment.

Using individual growth models, the Social Reciprocity/Language domain demonstrated statistically significant changes over time, whereas the Early Communication domain did not (though paired *t*-tests showed significant changes in both domains). This finding diverges from the BOSCC-MV, in which young children are most likely to demonstrate changes in items such as eye contact, facial expressions, and gestures (similar to the BOSCC-PSYF Early Communication domain) whereas older and more verbal children in our sample were most likely to demonstrate changes in the Social Reciprocity/Language domain (Grzadzinski et al., 2016). As mentioned previously, it is possible that interpreting the domains separately may prove most useful in identifying changes, depending on the type of intervention children are involved in. Researchers should make a-priori decisions regarding which skills are most likely to demonstrate change depending on the age of the child and the goals of the specific intervention.

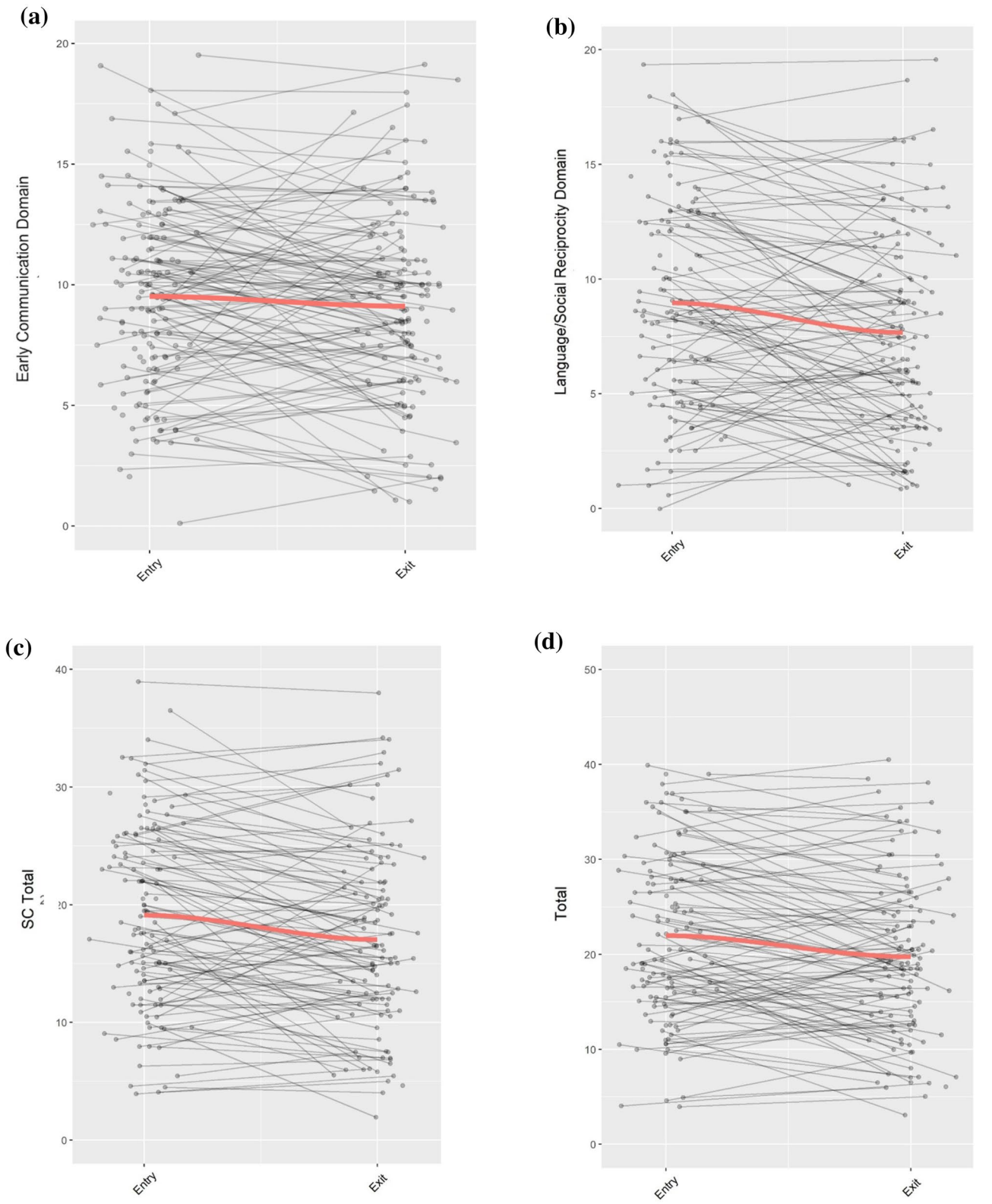

Fig. 2 **a** Decrease in Early Communication domain scores (improvement in behavior) over 4.5-month period. **b** Decrease in Language/Social Reciprocity domain scores (improvement in behavior) over 4.5-month period. **c** Decrease in combined Social Communication domain scores (improvement in behavior) over 4.5-month period. **d** Decrease in combined Core Total scores (improvement in behavior) over 4.5-month period

Over a 4.5-month period, the BOSCC-PSYF demonstrated small, statistically significant effect sizes in the Social Reciprocity/Language, combined Social Communication, and Core Total domains. In contrast, the VABS Communication and Socialization scores over the same 4.5-month period demonstrated much smaller effect sizes. The effect size of the ADOS CSS was slightly smaller. While the SRS demonstrated large effect sizes over the 4.5-month period, this measure is a parent report measure which may allow for the possibility of bias because parents were aware of their child's participation in an intervention at the time. These results suggest that the BOSCC-PSYF may be more sensitive to changes over brief periods of time than the VABS and possibly the ADOS CSS, two commonly used outcome measures (Grzadzinski et al., 2020), but should be used in conjunction with other measures, such as parent report measures, to achieve a comprehensive understanding of treatment response.

Similar to the results of the BOSCC-MV, the RRB domain of the BOSCC-PSYF demonstrated lower inter-rater and test–retest reliability and did not achieve a uniform or normal distribution of codes. While this was an expected outcome based on previous literature and from initial analyses of the ADOS (from which the BOSCC items were developed), it nonetheless indicates that the BOSCC RRB domain may not prove useful in identifying changes over short periods of time (Grzadzinski et al., 2016; Kitzerow et al., 2016; Lord et al., 2012). It may be that subtle changes in RRBs are difficult to capture in a brief observational measure, or that these behaviors do not vary greatly over time (e.g., they are either present or not). Nonetheless, the Core Total scores (i.e., RRB domain combined with the Social Communication domains) demonstrated significant amounts of change, indicating that RRBs are worth considering in conjunction with social communicative behaviors. Parent report data on RRBs should continue to be collected and used in combination with observation-based measures, such as the BOSCC.

Psychometric analyses indicate that the BOSCC has good inter-rater reliability. This is a promising finding, considering that individuals of all levels of experience coded these BOSCC-PSYF videos (e.g., undergraduate research assistants, graduate students, and post-doctorate level scholars). As is mentioned in Grzadzinski et al. (2016), because the BOSCC measures social communication changes within an individual, inter-rater reliability between individuals at one site is crucial, whereas reliability across sites is less important (unlike common diagnostic measures, such as the ADOS).

Psychometric analyses indicate that the BOSCC-PSYF also has good test–retest reliability for each domain and total (except for RRBs), though it is still lower than would be preferred. There are a few plausible reasons for this. First, due to collecting data during the COVID-19 pandemic, the number of test–retest cases we had available to analyze ($n = 26$) was less than ideal. Future test–retest data will continue to be collected. Furthermore, conducting test–retest reliability of brief observational measures has proven challenging. The mood and disposition of children can fluctuate easily, changing the behaviors that arise during the assessment (which is true for many observational assessments, but especially challenging for brief assessments). Thus, the BOSCC should be conducted at a time in which the child is in a neutral or positive mood and should be discontinued if the mood or behavior of the child is not representative of their usual behavior. Nonetheless, even in controlled settings, children demonstrate varying behaviors across the BOSCC administrations over short periods of time which we hope will not mask the measurement of intervention response.

The results garnered from this study are promising. However, there are several limitations to consider, the most prominent of which is the lack of a control group. All participants in this study participated in treatment of some kind, meaning that there was no "true" control group to which we could compare change scores. Future work will include the use of control groups to compare social communication changes across groups. Additionally, we would not expect that all participants would change in response to a particular treatment. Changes may be variable across individuals and across treatments. Larger sample sizes would allow consideration of individual differences in response to treatment, which we did not do here. In addition to larger sample sizes, future work on the BOSCC should also include racially diverse samples. Finally, the BOSCC is a measure of the generalization of changes in social communication to a standard set of activities; it is possible that some treatments result in proximal changes that, in the end, yield more general improvements that are not measured by the BOSCC.

Conclusion

Results from this study provide initial validation of the BOSCC-PSYF as an outcome measure of treatment response for individuals of all ages who have phrase speech and for fluent speaking children under the age of 8. The BOSCC provides a standardized, reliable, and valid measure of social communication changes over a short period of time. Its flexible nature allows for individuals of varying skill levels to administer and code the assessment. While telehealth administration was not used for the current study, it can be conducted through telehealth (by providing kits to families and videotaping caregiver-implemented BOSCC administrations through videoconferencing platforms), increasing accessibility across communities, and meeting the needs of the changing environment during the COVID-19 pandemic and beyond (Zwaigenbaum et al., 2021). The BOSCC coding

scheme can be applied to videos that do not implement the standardized BOSCC administration, including caregiver child interactions or segments of ADOS-2 administration (Kim et al., 2018). The way the BOSCC is conducted (i.e., videotaped and later scored) allows for "truly blinded" coders who are unaware of participant characteristics, timepoint, or treatment status and, if carried out with "blinded" interaction partners, they too can remain unbiased. This measure, used in conjunction with existing measures of treatment response, including caregiver report, has the potential to fill an important gap in currently available outcome measures used in autism intervention research.

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10803-022-05877-5.

Author Contributions All authors contributed to the study conception and design. Data collection and analysis were performed by AH, SHK, KS, and KB. The first draft of the manuscript was written by KB and KS and all authors commented on all iterations of the manuscript. All authors read and approved the final manuscript.

Funding This work was supported by two grants from the Simons Foundation (624965, CL; 345327, SK & CL). This work was also supported by a Dennis Weatherstone Predoctoral Fellowship from Autism Speaks.

Declarations

Conflict of interest Financial Interests: CL acknowledges the receipt of royalties from the sale of the Autism Diagnostic Observation Schedule (ADOS) and the Autism Diagnostic Interview-Revised (ADI-R). KB, KS, AH, SK, and RG have no potential financial conflicts to disclose. Non-financial Interests: KB, KS, AH, SK, RG, and CL have no relevant non-financial interests to disclose.

Ethical Approval All procedures performed in studies involving human participants were in accordance with the ethical standards of the University of California, Los Angeles and Weill Cornell Medicine Institutional Review Boards and with the 1964 Helsinki Declaration and its later amendments or comparable ethical standards.

Consent to Participate Informed consent was obtained from all participants.

Consent to Publish Participants signed informed consent regarding publishing their data.

References

American Psychiatric Association. (2013). *Diagnostic and statistical manual of mental disorders* (5th ed.). American Psychiatric Association.

Anagnostou, E., Jones, N., Huerta, M., Halladay, A. K., Wang, P., Scahill, L., Horrigan, J. P., Kasari, C., Lord, C., Choi, D., Sullivan, K., & Dawson, G. (2015). Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder. *Autism, 19*(5), 622–636. https://doi.org/10.1177/1362361314542955

Asparouhov, T., & Muthén, B. (2009). Exploratory structural equation modeling. *Structural Equation Modeling: A Multidisciplinary Journal, 16*(3), 397–438. https://doi.org/10.1080/10705510903008204

Bolte, E. E., & Diehl, J. J. (2013). Measurement tools and target symptoms/skills used to assess treatment response for individuals with autism spectrum disorder. *Journal of Autism and Developmental Disorders, 43*(11), 2491–2501. https://doi.org/10.1007/s10803-013-1798-7

Busner, J., & Targum, S. D. (2007). The clinical global impressions scale: Applying a research tool in clinical practice. *Psychiatry (Edgmont), 4*(7), 28.

Constantino, J. N., & Gruber, C. P. (2012). *Social responsiveness scale: SRS-2*. Western Psychological Services.

Cunningham, A. B. (2012). Measuring change in social interaction skills of young children with autism. *Journal of Autism and Developmental Disorders, 42*(4), 593–605. https://doi.org/10.1007/s10803-011-1280-3

Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenson, J., Donaldson, A., & Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the early start denver model. *Pediatrics*. https://doi.org/10.1542/peds.2009-0958

Dunn, L. M., Dunn, D. M., Pearson Assessments. (2007). *PPVT-4: Peabody picture vocabulary test*. NY: Pearson Assessments.

Elliott, C. D., Salerno, J. D., Dumont, R., & Willis, J. O. (2018). The differential ability scales—Second edition. In D. P. Flanagan & E. M. McDonough (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 360–382). The Guilford Press.

Esler, A. N., Bal, V. H., Guthrie, W., Wetherby, A., Weismer, S. E., & Lord, C. (2015). The autism diagnostic observation schedule, toddler module: Standardized severity scores. *Journal of Autism and Developmental Disorders, 45*(9), 2704–2720. https://doi.org/10.1007/S10803-015-2432-7

Estes, A., Munson, J., Rogers, S. J., Greenson, J., Winter, J., & Dawson, G. (2015). Long-term outcomes of early intervention in 6-year-old children with autism spectrum disorder. *Journal of the American Academy of Child and Adolescent Psychiatry, 54*(7), 580–587. https://doi.org/10.1016/j.jaac.2015.04.005

Fletcher-Watson, S., & McConachie, H. (2017). The search for an early intervention outcome measurement tool in autism. *Focus on Autism and Other Developmental Disabilities, 32*(1), 71–80. https://doi.org/10.1177/1088357615583468

Fuller, E. A., & Kaiser, A. P. (2020). The effects of early intervention on social communication outcomes for children with autism spectrum disorder: A meta-analysis. *Journal of Autism and Developmental Disorders, 50*(5), 1683–1700. https://doi.org/10.1007/S10803-019-03927-Z

Gengoux, G. W., Abrams, D. A., Schuck, R., Millan, M. E., Libove, R., Ardel, C. M., Phillips, J. M., Fox, M., Frazier, T. W., & Hardan, A. Y. (2019). A pivotal response treatment package for children with autism spectrum disorder: An RCT. *Pediatrics*. https://doi.org/10.1542/PEDS.2019-0178/-/DCSUPPLEMENTAL

Gotham, K., Pickles, A., & Lord, C. (2008). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(5), 693–705. https://doi.org/10.1007/S10803-008-0674-3

Gotham, K., Pickles, A., & Lord, C. (2012). Trajectories of autism severity in children using standardized ADOS scores. *Pediatrics*. https://doi.org/10.1542/peds.2011-3668

Green, J., Leadbitter, K., Ellis, C., Taylor, L., Moore, H. L., Carruthers, S., James, K., Taylor, C., Balabanovska, M., Langhorne, S., Aldred, C., Slonims, V., Grahame, V., Parr, J., Humphrey, N., Howlin, P., McConachie, H., Couteur, A. L., Charman, T., … Pickles, A. (2022). Combined social communication therapy at home and ineducation for young autistic children in England (PACT-G): a parallel, single-blind, randomised controlled trial.

Thelancet. Psychiatry, 9(4), 307–320. https://doi.org/10.1016/S2215-0366(22)00029-3

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

Grzadzinski, R., Janvier, D., & Kim, S. H. (2020). Recent developments in treatment outcome measures for young children with autism spectrum disorder (ASD). *Seminars in Pediatric Neurology, 34*, 100806. https://doi.org/10.1016/J.SPEN.2020.100806

Guastella, A. J., Gray, K. M., Rinehart, N. J., Alvares, G. A., Tonge, B. J., Hickie, I. B., Keating, C. M., Cacciotti-Saija, C., & Einfeld, S. L. (2015). The effects of a course of intranasal oxytocin on social behaviors in youth diagnosed with autism spectrum disorders: A randomized controlled trial. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 56*(4), 444–452. https://doi.org/10.1111/jcpp.12305

Gutstein, S., Burgess, A. F., & Montfort, K. (2007). Evaluation of the relationship development intervention program. *Autism, 11*(5), 397–411. https://doi.org/10.1177/1362361307079603

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

Jones, R. M., Carberry, C., Hamo, A., & Lord, C. (2017). Placebo-like response in absence of treatment in children with autism. *Autism Research: Official Journal of the International Society for Autism Research, 10*(9), 1567–1572. https://doi.org/10.1002/AUR.1798

Kaale, A., Smith, L., & Sponheim, E. (2012). A randomized controlled trial of preschool-based joint attention intervention for children with autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 53*(1), 97–105. https://doi.org/10.1111/j.1469-7610.2011.02450.x

Kasari, C., Gulsrud, A., Freeman, S., Paparella, T., & Hellemann, G. (2012). Longitudinal follow-up of children with autism receiving targeted interventions on joint attention and play. *Journal of the American Academy of Child and Adolescent Psychiatry, 51*(5), 487–495. https://doi.org/10.1016/j.jaac.2012.02.019

Kim, S. H., Grzadzinski, R., Martinez, K., & Lord, C. (2018). Measuring treatment response in children with autism spectrum disorder: Applications of the brief observation of social communication change to the autism diagnostic observation schedule. *Autism, 23*(5), 1176–1185. https://doi.org/10.1177/1362361318793253

Kim, S. H., & Lord, C. (2010). Restricted and repetitive behaviors in toddlers and preschoolers with autism spectrum disorders based on the Autism Diagnostic Observation Schedule (ADOS). *Autism research : official journal of the International Society for Autism Research, 3*(4), 162–173. https://doi.org/10.1002/aur.142

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 Research: Official Journal of the International Society for Autism Research, 9*(9), 940–950. https://doi.org/10.1002/AUR.1588

Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. (2012). *Autism diagnostic observation schedule: ADOS-2*. Western Psychological Services Torrance.

Marsh, H. W., Morin, A. J., Parker, P. D., & Kaur, G. (2014). Exploratory structural equation modeling: An integration of the best features of exploratory and confirmatory factor analysis. *Annual Review of Clinical Psychology, 10*, 85–110. https://doi.org/10.1146/annurev-clinpsy-032813-153700

Matson, J. L. (2007). Determining treatment outcome in early intervention programs for autism spectrum disorders: A critical analysis of measurement issues in learning based interventions. *Research in Developmental Disabilities, 28*(2), 207–218. https://doi.org/10.1016/j.ridd.2005.07.006

McConachie, H., Parr, J., Glod, M., Hanratty, J., Livingstone, N., Oono, I., Robalino, S., Baird, G., Beresford, B., Charman, T., Garland, D., Green, J., Gringras, P., Jones, G., Law, J., Le Couteur, A. S., Macdonald, G., McColl, E. M., Morris, C., & Williams, K. (2015). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment (Winchester, England), 19*(41), 1–538. https://doi.org/10.3310/HTA19410

Miller, L. E., Perkins, K. A., Dai, Y. G., & Fein, D. A. (2017). Comparison of parent report and direct assessment of child skills in toddlers. *Research in Autism Spectrum Disorders, 41–42*, 57. https://doi.org/10.1016/J.RASD.2017.08.002

Mullen, E. M. (1995). *Mullen scales of early learning* (AGS). American Guidance Service Inc.

Nordahl-Hansen, A., Fletcher-Watson, S., McConachie, H., & Kaale, A. (2016). Relations between specific and global outcome measures in a social-communication intervention for children with autism spectrum disorder. *Research in Autism Spectrum Disorders, 29–30*, 19–29. https://doi.org/10.1016/J.RASD.2016.05.005

Nordahl-Hansen, A., Kaale, A., & Ulvund, S. E. (2014). Language assessment in children with autism spectrum disorder: Concurrent validity between report-based assessments and direct tests. *Research in Autism Spectrum Disorders, 8*(9), 1100–1106. https://doi.org/10.1016/J.RASD.2014.05.017

Owley, T., Mcmahon, W., Cook, E. H., Laulhere, T., South, M., Zellmer Mays, L., Shernoff, E. S., Lainhart, J., Modahl, C. B., Corsello, C., Ozonoff, S., Risi, S., Lord, C., Leventhal, B. L., & Filipek, P. A. (2001). Multisite, double-blind, placebo-controlled trial of porcine secretin in autism. *Journal of the American Academy of Child and Adolescent Psychiatry, 40*(11), 1293–1299. https://doi.org/10.1097/00004583-200111000-00009

Ozonoff, S., Iosif, A. M., Young, G. S., Hepburn, S., Thompson, M., Colombi, C., Cook, I. C., Werner, E., Goldring, S., Baguio, F., & Rogers, S. J. (2011). Onset patterns in autism: Correspondence between home video and parent report. *Journal of the American Academy of Child and Adolescent Psychiatry*. https://doi.org/10.1016/J.JAAC.2011.03.012

Pickles, A., Le Couteur, A., Leadbitter, K., Salomone, E., Cole-Fletcher, R., Tobin, H., Gammer, I., Lowry, J., Vamvakas, G., Byford, S., Aldred, C., Slonims, V., McConachie, H., Howlin, P., Parr, J. R., Charman, T., & Green, J. (2016). Parent-mediated social communication therapy for young children with autism (PACT): Long-term follow-up of a randomised controlled trial. *The Lancet, 388*(10059), 2501–2509. https://doi.org/10.1016/S0140-6736(16)31229-6

Raven, J., Raven, J. C., & Court, J. H. (2000). *Manual for Raven's progressive matrices and vocabulary scales: The standard progressive matrices*. Oxford University Press.

Rogers, S. J., Estes, A., Lord, C., Vismara, L., Winter, J., Fitzpatrick, A., Guo, M., & Dawson, G. (2012). Effects of a brief Early Start Denver Model (ESDM)–based parent intervention on toddlers at risk for autism spectrum disorders: A randomized controlled trial. *Journal of the American Academy of Child & Adolescent Psychiatry, 51*(10), 1052–1065.

Rosseel, Y. (2012). lavaan: An R package for structural equation modeling. *Journal of Statistical Software, 48*(2), 1–36. https://doi.org/10.18637/jss.v048.i02

Sandbank, M., Bottema-Beutel, K., Crowley, S., Cassidy, M., Dunham, K., Feldman, J. I., Crank, J., Albarran, S. A., Raj, S., Mahbub, P., & Woynaroski, T. G. (2020). Project AIM: Autism intervention meta-analysis for studies of young children. *Psychological Bulletin, 146*(1), 1. https://doi.org/10.1037/BUL0000215

Scahill, L., Aman, M. G., Lecavalier, L., Halladay, A. K., Bishop, S. L., Bodfish, J. W., Grondhuis, S., Jones, N., Horrigan, J. P., Cook, E. H., Handen, B. L., King, B. H., Pearson, D. A., McCracken, J. T., Sullivan, K. A., & Dawson, G. (2015). Measuring repetitive behaviors as a treatment endpoint in youth with autism spectrum disorder. *Autism: The International Journal of Research and Practice, 19*(1), 38–52. https://doi.org/10.1177/1362361313510069

Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland adaptive behavior scales (Vineland-3)* (3rd ed.). NCS Pearson.

Spence, S. J., & Thurm, A. (2010). Testing autism interventions: Trials and tribulations. *The Lancet, 375*(9732), 2124–2125. https://doi.org/10.1016/S0140-6736(10)60757-X

Thurm, A., Manwaring, S. S., Swineford, L., & Farmer, C. (2015). Longitudinal study of symptom severity and language in minimally verbal children with autism. *Journal of Child Psychology and Psychiatry and Allied Disciplines, 56*(1), 97–104. https://doi.org/10.1111/jcpp.12285

Toolan, C., Holbrook, A., Schlink, A., Shire, S., Brady, N., & Kasari, C. (2022). Using the Clinical Global Impression scale to assess social communication change in minimally verbal children with autism spectrum disorder. *Autism Research, 15*(2), 284–295. https://doi.org/10.1002/AUR.2638

Volkmar, F. R., Lord, C., Bailey, A., Schultz, R. T., & Klin, A. (2004). Autism and pervasive developmental disorders. *Journal of Child Psychology and Psychiatry, 45*(1), 135–170. https://doi.org/10.1046/J.0021-9630.2003.00317.X

Wechsler, D. (2012). *Wechsler preschool and primary scales of intelligence (WPPSI-IV)* (4th ed.). Psychological Corporation.

Wolery, M., & Garfinkle, A. N. (2002). Measures in intervention research with young children who have autism. *Journal of Autism and Developmental Disorders, 32*(5), 463–478. https://doi.org/10.1023/A:1020598023809

World Health Organization. (2019). *International statistical classification of diseases and related health problems* (11th ed.). https://icd.who.int/

Yoder, P. J., Bottema-Beutel, K., Woynaroski, T., Chandrasekhar, R., & Sandbank, M. (2013). Social communication intervention effects vary by dependent variable type in preschoolers with autism spectrum disorders. *Evidence-Based Communication Assessment and Intervention, 7*(4), 150–174. https://doi.org/10.1080/17489539.2014.917780

Yoder, P., Woynaroski, T., Fey, M., & Warren, S. (2014). Effects of dose frequency of early communication intervention in young children with and without down syndrome. *American Journal on Intellectual and Developmental Disabilities, 119*(1), 17–32. https://doi.org/10.1352/1944-7558-119.1.17

Zapolski, T. C. B., & Smith, G. T. (2013). Comparison of parent versus child-report of child impulsivity traits and prediction of outcome variables. *Journal of Psychopathology and Behavioral Assessment, 35*(3), 301–313. https://doi.org/10.1007/S10862-013-9349-2

Zheng, S., Kaat, A., Farmer, C., Kanne, S., Georgiades, S., Lord, C., Esler, A., & Bishop, S. L. (2021). Extracting latent subdimensions of social communication: A cross-measure factor analysis. *Journal of the American Academy of Child and Adolescent Psychiatry, 60*(6), 768-782.e6. https://doi.org/10.1016/J.JAAC.2020.08.444

Zwaigenbaum, L., Bishop, S., Stone, W. L., Ibanez, L., Halladay, A., Goldman, S., Kelly, A., Klaiman, C., Lai, M. C., Miller, M., Saulnier, C., Siper, P., Sohl, K., Warren, Z., & Wetherby, A. (2021). Rethinking autism spectrum disorder assessment for children during COVID-19 and beyond. *Autism Research, 14*(11), 2251–2259. https://doi.org/10.1002/AUR.2615