UCLA
UCLA Electronic Theses and Dissertations

Title

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

Permalink

https://escholarship.org/uc/item/3151h6vn

Author

Byrne, Katherine

Publication Date

2022

Peer reviewed|Thesis/dissertation

eScholarship.org Powered by the California Digital Library
University of California

UNIVERSITY OF CALIFORNIA

Los Angeles

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

A thesis submitted in partial satisfaction of the requirements for the degree Master of Arts in Education

by

Katherine Anne Byrne

2022

© Copyright by

Katherine Anne Byrne

2022

ABSTRACT OF THE THESIS

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

by

Katherine Anne Byrne

Master of Arts in Education

University of California, Los Angeles, 2022

Professor Catherine Lord Morrison, Chair

Individuals with autism spectrum disorder (ASD) are commonly involved in interventions aimed at improving communication or other social communicative behaviors. However, the field of ASD intervention research faces significant limitations in its current practices of measuring treatment effectiveness. There is a need for the development of outcome measures that adequately address the limitations of the measures historically used and that can reliably detect changes in the social communicative behaviors of individuals with ASD, especially in a short period of time. The aim of the present study was to determine the utility of the BOSCC-Phrase Speech Young Fluent (PSYF) as an outcome measure of treatment response. Specifically, this study analyzed the factor structure of the measure, examined its initial psychometric properties, and provided evidence of its utility as a measure of change. The BOSCC coding scheme was

applied to 345 video administrations from 160 participants diagnosed with ASD. Participants included individuals of any age with consistent and flexible phrase speech, or individuals under the age of 8 years with fluent, complex sentences. The BOSCC-PSYF has an underlying three-factor structure. Test-Retest reliability was good for the Early Communication domain, moderate for the Social Reciprocity/Language domain, and poor for the RRB domain. Inter-rater reliability was good for the Early Communication and Social Reciprocity/Language domains and fair for the RRB domain. Significant changes occurred over time in the Early Communication and Social Reciprocity/Language domains, and Core Total scores. Standardized effect sizes of change were larger in the BOSCC domains than in ADOS CSS and VABS Communication Standard Scores. The BOSCC provides a standardized, flexible, and minimally biased assessment of social communication changes in response to treatment. Its validation would have important implications for ASD intervention research, including the possibility of a low-cost measure that reliably measures changes in broad social communicative behaviors in a short period of time, can be conducted and coded by individuals of various skill levels, and is flexible enough to be used across various sites/studies.

The thesis of Katherine Anne Byrne is approved.

Connie L. Kasari

Jeffrey J. Wood

Catherine Lord Morrison, Committee Chair

University of California, Los Angeles

2022

Table of Contents

List of Figures

List of Tables

List of Abbreviations

Abbreviation	Meaning
ADOS-2	Autism Diagnostic Observation Schedule, *Second Edition*
ASD	Autism spectrum disorder
BOSCC	Brief Observation of Social Communication Change
CSS	Calibrated Severity Score
CGI	Clinical Global Impression
CFI	Comparative Fit Index
CFA	Confirmatory factor analysis
DSM-5	Diagnostic and Statistical Manual of Mental Disorders, *Fifth Edition*
DAS-II	Differential Ability Scales, *Second Edition*
EFA	Exploratory factor analysis
ESEM	Exploratory structural equation modeling
ICD-11	International Classification of Disease 11th Revision
MV	Minimally Verbal
MSEL	Mullen Scales of Early Learning
PPVT-4	Peabody Picture Vocabulary Test, *4th Edition*
PSYF	Phrase Speech/Young Fluent
RRB	Restricted and Repetitive Interests and Behaviors (RRB)
RMSEA	Root mean square error of approximation
SC	Social Communication
SRS-2	Social Responsiveness Scale, *Second Edition*
TLI	Tucker-Lewis Index
VABS-3	Vineland Adaptive Behavior Scales, *Third Edition*
WPPSI-IV	Wechsler Preschool and Primary Scale of Intelligence, *Fourth Edition*

Individuals with autism spectrum disorder (ASD) are involved in numerous treatments and interventions throughout their lifespan, the most common of which are aimed at improving social communicative behaviors (Fuller & Kaiser, 2020; Sandbank et al., 2020; Rogers & Vismara, 2008). 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 ASD intervention research faces significant limitations in measuring treatment 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 treatments may be optimal for whom and when (Cunningham, 2012; Magiati et al., 2011). 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 ASD, 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 report. This can be problematic due to the likelihood of systematic measurement error, expectancy bias or placebo effects (Anagnostou et al., 2015; Bolte & Diehl, 2013; Sandbank et al., 2020). This is due, in part, to "unblinding," which occurs when caregivers or clinicians are aware of or suspect a treatment condition in an intervention trial. For example, the Clinical Global Impression (CGI) rating scales (Busner & Targum, 2007) is one of the most commonly used outcome measures in intervention research (Bolte & Diehl, 2013; Toolan et al., 2022). The

CGI is a subjective measure of relative improvement completed by clinicians, which cannot always control for expectancy or placebo biases. For example, in the case of some behavioral interventions, the CGI is sometimes completed by the clinician responsible for delivering the treatment to the participant, in part because this mirrors typical clinical practice. The clinician's involvement in the delivery of treatment indicates that he/she is not blind to condition, yet he/she is responsible for collecting the treatment response data. In medication trials, caregivers or physicians may begin the study blind to treatment condition, and later become aware of treatment condition 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 in intervention research, which may lead one to believe that strong treatment 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 ASD-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 a treatment may affect caregivers' responses. Other caregiver biases include overestimating a child's abilities due to reluctance to acknowledge a child's delays, difficulty recalling and reporting a child's developmental milestones, caregivers' investment in positive outcomes, and caregivers' likelihood to pay 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; Sandbank et al., 2020; Zapolski & Smith, 2013). While caregiver report and clinician judgment are important sources of information regarding a child's skills and deficits, reliance on these measures alone limits the interpretation of treatment responses (Miller et al., 2017).

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

Changes in ASD-specific symptoms are often measured using diagnostic instruments (Aldred et al., 2004; Dawson et al., 2010; Green et al., 2010). However, 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 responses to short-term treatments. Rather, these instruments 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 (Owley et al., 2001; Grzadzinski et al., 2020). While some studies have found significant changes over time in ADOS raw scores, these changes were also evident in the treatment-as-usual groups (Green et al., 2010; Gutstein et al., 2007). In other studies that have 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 ASD symptoms (Estes 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

ASD symptom severity that is less confounded by changes in general maturity or language development (Gotham et al., 2008).

The use of the ADOS CSS has been successful in measuring changes in ASD symptom severity over time (Gotham et al., 2008; Grzadzinski et al., 2020). Yet, these changes have only 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; Shumway et al., 2012; Thurm et al., 2015). Considering that short-term intensive interventions are common for individuals with ASD, 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 have to be administered more than once to measure changes over time. While the use of diagnostic instruments as treatment response measures was once encouraged, an instrument that reliably measures ASD-specific symptoms and is more sensitive to subtle changes in a short period of time will be of crucial importance for ASD intervention research moving forward (Matson, 2007).

Lack of Uniform Measurement Approach

ASD intervention research has utilized hundreds of disparate outcome measures in order 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. This is due, at least in part, to the heterogeneity of the type and severity of ASD symptoms present within individuals. For example, deficits in social communication could be

considered to include verbal or nonverbal communication delays or 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 in research to measure specific behaviors, such as joint attention (Green et al., 2010; 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 disparate outcome measures were identified (Bolte & Diehl, 2013). Of the 289 measures, 61.6% were found in only one publication over a 10-year period and 20.8% 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 specific task in a specific context, although they are targeted in interventions and outcome measurement in the hopes that improvement of these behaviors will have positive cascading effects on other domains, such as language development or better peer relations (Green et al., 2010; Mundy et al., 1990). While identifying changes in specific behaviors is important, whether these context-specific behaviors resulted in more generalized gains across broad social communication strategies often goes unmeasured (Sandbank et al., 2020). It is necessary to understand whether context-specific behaviors generalize to other aspects of social communication and result in broader positive effects on development (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 quite subtle, meaning their measurement must be sensitive enough to capture small, but clinically meaningful changes that indicate measurable improvement and ideally predict more positive outcomes (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 treatment response measures without certain modifications (e.g., use only for specific populations, such as young children or those with average or greater IQ), making the use of a uniform measurement approach of treatment response 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., at least 8 to 12 weeks). The BOSCC is a play-based assessment conducted with the participant and a play partner, such as a parent or research staff member.

The BOSCC was initially developed using codes from the ADOS-2, 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). This measure is flexible and standardized, which allows for its use across sites and studies. Finally, the BOSCC is observation-based, using interactions with partners who may be completely 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 problems that intervention research has historically faced 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 by people of all skill levels as well as with a variety of populations and research contexts.

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 in the treatment group as compared to a no-change control group; ADOS CSS scores over the same period of time 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 fit 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 and its ability to detect changes in a short period of time (Gengoux et al., 2019; Kim et al., 2019; Kitzerow et al., 2016; Nordahl-Hansen et al., 2016).

Pilot testing was conducted to test whether the BOSCC-MV could be used with older or more verbal children. Results 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 treatment status in children receiving a range of treatments in four different sites. This same pilot testing also revealed that changing the coding scheme for these older children without modifying the administration was insufficient. Rather, the context in which the interaction occurred needed to be altered so there was less variability if changes were to be detected. We used this pilot data to extend the work already conducted on the BOSCC-MV to create new contexts 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 treatment 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.

Method

Participants

Participants included 160 English-speaking children between the ages of 2-18 years with a documented diagnosis of autism spectrum disorder (ASD). Eighty-four percent of the sample identified as being male, and 16% identified as being female. The self-reported racial identities of the participants in this study were as follows: 9% Black, 9% Asian American, 70% White, and 12% biracial. Twenty-one percent identified as being Hispanic. All participants had language abilities suitable for the PSYF administration. Specifically, the PSYF module is appropriate for 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 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 hours 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 hours 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 (though some participants were lost to follow-up), along with collecting 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 will be available for 144 participants. 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 analysis of change.

Measures

Brief Observation of Social Communication Change (BOSCC-PSYF)

The BOSCC was developed as a treatment response measure of social communicative and other behaviors associated with an autism spectrum disorder (ASD). The BOSCC is a 12-minute, videotaped play interaction between an individual and a play partner (e.g., clinician, teacher, caregiver). 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 are 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 or research assistants.

Coding Procedures. The BOSCC videos are split into two 6-minute 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 modified to fit the social communicative behaviors of children with phrase speech, and 3 novel items (i.e., verbal exchanges, offering information, 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; 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. Each decision tree contains detailed information regarding the frequency and quality of specific behaviors. At each branch, the coder answers a yes or no question 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 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 investigating the convergent validity 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 individuals 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 was assessed using a variety of measures, including the Mullen Scales of Early Learning (MSEL; Mullen, 1995), the Differential Ability Scales (DAS-II; Elliot et al., 2018), the Wechsler Preschool and Primary Scale of Intelligence (WPPSI-IV; Wechsler, 2012), the Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007), and the Ravens Progressive Matrices (Raven et al., 2000). The MSEL was collected for 30 children, the DAS-II for 62 children, the WPPSI-IV for 19 children, and the PPVT-4 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 Communication, Daily Living Skills, Socialization, and Motor Skill domains, as well as an overall Adaptive Behavior Composite score. The VABS-3 was administered to 151 participants at one timepoint, and 73 of these same participants at a second timepoint, which allowed for analysis of change in scores over time. A combination of the comprehensive interview form and the caregiver report form was used.

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 either uniform or normal distribution of codes across all items. Item level codes were re-written over several versions until near-flat distributions were achieved. We did not expect a uniform distribution for items related to restricted and repetitive interests and behaviors because the presentation of these behaviors is extremely heterogeneous across individuals. Furthermore, 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 minimally verbal version of the BOSCC 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 square 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 .95 for the TLI and CFI and ≤ .06 for the RMSEA, although these cutoffs tend to be overly exclusive in small samples (Hu & Bentler, 2009).

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. Again, factors were allowed to covary. TLI, CFI, and RMSEA were used to evaluate the fit of the CFA model.

Longitudinal Measurement Invariance

Due to the longitudinal nature of the data, it was important to confirm that the factor structure was invariant across time. 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 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 reliability and inter-rater reliability was analyzed. Test-retest reliability was estimated from 16 participants who had a second BOSCC conducted within one-month of each other. We had hoped to collect more than 16 test-retest videos; however, this became impossible due to the COVID-19 pandemic and subsequent restrictions. Inter-rater reliability was estimated

from 54 videos which were double coded. Absolute agreement was assessed using two-way random effects models. Inter-rater and test-retest reliability results were described for each domain of the BOSCC-PSYF derived from the factor analyses, as well as the Core Total.

Change Analyses

Following procedures from the analyses of the minimally verbal version of the BOSCC (Grzadzinski et al., 2016; Kim et al., 2019), first, paired sample t-tests were used to compare BOSCC Core Total and Domain 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 other behavioral measures with sufficient data (i.e., VABS Communication, ADOS CSS and SRS Total Scores) as the dependent variable. This involved fitting a linear regression separately for each participant using participant's age at the time of assessment as the independent variable in order to generate an average rate of change per month. Rate of change was converted to represent expected change over 4.5 months, the average length of time between the intake and exit appointments in our sample. To be consistent with prior analyses, these rates of change were also converted to represent expected change over 6 months. These rates were then divided by the standard deviation of the measure at intake 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 and IQ at intake were related to or moderated change in BOSCC scores over time.

Lastly, again following the procedures of Grzadinski et al. (2016) and Kim et al. (2019), response status was determined separately based on change scores in each of the behavioral measures (i.e., VABS Communication, ADOS CSS and SRS Total Scores). Change of greater

than or equal to 8 points on VABS Communication standard scores and SRS Total scores (half a standard deviation) and greater than or equal to 1 point on ADOS CSS Scores were used to classify responders. 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

Figure 1 depicts the distribution of BOSCC-PSYF Core codes (averaged for Segment A and B) across the 14 out of the 17 items in the final version of the BOSCC-PSYF. 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 are 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 would best fit the data (see Table 1 for fit statistics). The best solution based on the fit statistics was the four-factor solution; the parameters are included in Table 2. 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; fit statistics across each model are included in Table 3. Parameter estimates for the three-factor model are included in Table 4. 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. Due to the clinical value that the "Play with Objects" item provides and the possibility of play skills to improve with interventions, substantive decisions were made to keep this item in the final algorithm but remove it from the measure's factor structure. Figure 2 depicts the items, domains, and Core Total.

The CFA model adequately fit the data (CFI = .937, TLI = .921 and RMSEA = .076). Item loadings across factors were high with the exception of some RRB items. The factor loadings of the items onto the Early Communication factor ranged from .63 to .95, loadings of the Social Reciprocity/Language factor ranged from .71 to .84 and the loadings of the RRB factor ranged from .25 to .58. Model parameters across items are include in Table 6.

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 7.

Test-Retest Reliability

Test-retest reliability was estimated from 16 videos. Adequate reliability for one item, "Engagement in Play with Others" was not able to be reached, likely do the subjective and variable nature of the construct it measures. Additionally, when coding this item, coders reported anecdotally that this item seemed to fluctuate based on the child's mood and disposition throughout 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 domain and fair for the Social Reciprocity/Language and RRB domains. The ICC value for the Early Communication domain was 0.82, 95% CI [0.49,0.95], 0.53, 95% CI [-0.05,0.84] for the Social Reciprocity/Language domain and 0.42, 95% CI [-0.19, 0.79] for the RRB domain.

Inter-Rater Reliability

Inter-rater reliability was estimated from 54 double-coded videos. Overall, inter-rater reliability was good for the Early Communication domain and Social Reciprocity/Language domain and was fair for the RRB domain. The ICC value for the Early Communication domain was 0.85, 95% CI [0.75,0.91], 0.87, 95% CI [0.78, 0.92] for the Social Reciprocity/Language domain and 0.60, 95% CI [0.38, 0.75] for the RRB domain.

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, SD = .62, $t(111)$, = 2.61, $p = 0.01$, Cohen's $d = -0.25$), the Social Reciprocity/Language domain (M = -1.32, SD = .62, $t(111)$, = 4.21, $p < 0.01$, Cohen's $d = -0.40$), the combined Social Communication domain (M = -2.03, SD = .95, $t(111)$, = 4.26, $p < 0.01$, Cohen's $d = -0.40$), and the Core Total (M = -2.02, SD = 1.05, $t(111)$, = 3.80, $p < 0.01$, Cohen's $d = -0.36$). 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, SD =.67, $t(36) = 1.48$, $p = 0.49$, Cohen's $d = -0.24$).

Results from the individual growth models indicated that the average rates of change over 4.5 months was small in the Early Communication domain (Cohen's $d = -0.28$, 95% CI [-0.65, 0.09], see Figure 3), and greater for the Social Reciprocity/Language domain (Cohen's $d = -0.45$, 95% CI [-0.71, -0.19], see Figure 4), the combined Social Communication domain (Cohen's $d = -0.41$, 95% CI [-0.66, -0.16], see Figure 5), and the Core Total (Cohen's $d = -0.41$, 95% CI [-0.67, -0.15], see Figure 6), with larger changes when comparisons were made for 6 months. The average rate of change over 4.5 months was smaller in the ADOS CSS (Cohen's $d = -0.32$, 95% CI [-0.80, 0.16]) and the VABS Communication Standard Score (Cohen's $d = -0.07$, 95% CI [-0.23, -0.08]). 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]).

Moderating Variables

The results of the linear mixed models found that children's chronological age ($t = -2.104$, $p = .039$) though not VIQ ($t = -1.786$, $p = .078$) nor NVIQ ($t = .0.161$, $p = .87$) was related to BOSCC Early Communication scores, where younger children had higher (more impaired) scores on the Early Communication domain. For the Social Reciprocity/Language domain, children's chronological age ($t = -2.43$, $p = .017$) and VIQ ($t = -3.49$, $p < .001$), though not NVIQ ($t = 1.42$, $p = .156$) were related to scores on this domain, where younger children and children with lower VIQ's had higher BOSCC Social Reciprocity/Language scores. BOSCC Combined Social Communication total scores were related to both children's age ($t = -2.42$, $p = .02$) and VIQ ($t = -2.88$, $p = .005$), though not NVIQ ($t = .90$, $p = .37$). VIQ ($t = -4.05$, $p < .001$), though

neither chronological age (t = -1.868, p = .06) nor NVIQ (t = .958, p = .34), was related to BOSCC Core Total scores. 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 = .03) and combined Social Communication domains than SRS "non-responders" (t(17) = 2.21, p = .04). There were no statistically significant differences on any BOSCC domain score for VABS Communication or ADOS CSS responders.

Discussion

Results from our analyses confirm prior literature that the BOSCC is a promising outcome measure of treatment 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 treatment 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. 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 being conducted.

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 effect sizes of the ADOS CSS and VABS Communication score over the same 4.5-month period were much 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 treatment at the time. These results suggest that the BOSCC may be more sensitive to changes over brief periods

of time than the ADOS CSS and VABS, 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 the 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., 2015; Lord et al., 2006). It may be that subtle changes in RRBs are difficult to capture in a brief observational measure, or that these behaviors do not vary as much over time (e.g., they are either present or not). Nonetheless, the BOSCC-PSYF Core Total scores (i.e., the RRB domain combined with the two Social Communication domains) demonstrated significant amounts of change indicating that RRBs are worth considering in conjunction with social communicative behaviors. It will be important to continue to collect parent report data on RRBs to be 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 has moderate test-retest reliability for each domain and total, except for RRBs. There are a few plausible reasons for the adequate test-retest reliability. First, due to collecting data during the COVID-19 pandemic, the number of test-retest cases we had available to analyze (n = 16) was lower than would be preferred. 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. 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 treatment effects.

Limitations

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 when comparing 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. 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

While the BOSCC has been validated in minimally verbal children, preliminary research demonstrated that the BOSCC-MV did not reliably detect changes in older or more verbal individuals. Thus, another version of the BOSCC, the BOSCC-PSYF, was developed. Now, 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. The flexible nature of the BOSCC allows for individuals of varying skill level to administer and code the assessment. 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 manner in which 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 reports, has the potential to fill an important gap in currently available outcome measures used in ASD intervention research.

Figure 1. Distribution of BOSCC-PSYF items.

Table 1

ESEM with Play

	X^2 (df)	RMSEA [90% CI]	CFI	TLI
1 Factor	257.02 (77)	.122 [.106, .139]	.818	.785
2 Factor	162.47 (64)	.099 [.080, .118]	.901	.859
3 Factor	94.82 (52)	.073 [.049, .096]	.957	.924
4 Factor	54.136 (41)	.045 [.000, .075]	.987	.971

Table 2

ESEM Factor Loadings with Play

	Standardized Loading (SE)			
	Factor 1	**Factor 2**	**Factor 3**	**Factor 4**
Eye Contact	0.004	0.816	-0.240	0.029
Facial Expressions	-0.047	0.671	0.264	0.155
Gesture	0.269	0.426	0.107	-0.170
Integration of Non-verbal Communication	0.062	0.896	0.053	-0.027
Quality of Social Overtures	0.603	0.150	0.137	0.073
Quality of Social Responses	0.731	0.079	-0.066	0.144
Verbal Exchanges	0.862	-0.030	0.082	0.007
Offering Information	0.855	0.067	-0.151	-0.052
Engagement in Play with Others	0.346	0.094	0.559	0.124
Play with Objects	-0.055	-0.011	0.677	-0.066
Stereotyped Speech	0.248	-0.024	-0.149	0.298
Sensory Behaviors	0.107	0.011	0.010	0.341
Mannerisms	-0.216	-0.026	0.033	0.385
Repetitive Behaviors	0.054	0.022	-0.032	0.712

Table 3

ESEM without Play

	X^2 (df)	RMSEA[(90% CI]	CFI	TLI
1 Factor	206.27 (65)	.118 [.100,.136]	.851	.821
2 Factor	114.45 (53)	.086 [.064,.108]	.935	.905
3 Factor	78.09 (42)	.074 [.048, .100]	.962	.929

Table 4

ESEM 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

Table 5

Model Fit CFA

	X^2 (df)	RMSEA[(90% CI]	CFI	TLI
3 Correlated Factors	186.846 (62)	.076 [.064, .089]	.937	.921

*Note: A 4-factor solution with the object play variable was fit but led to a negative variance estimate indicating a mis-specified model.

Table 6

Final CFA Parameter Estimates

	Standardized Loading (SE)		
	Factor 1	**Factor 2**	**Factor 3**
Eye Contact	.790		
Facial Expressions	.741		
Gesture	.630		
Integration of Non-verbal	.951		
Communication			
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

Table 7

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

Item	Domain		Total
Eye Contact	Early Communication	Social Communication	Core
Facial Expressions			
Gestures and Showing			
Integration of Vocal and Non-Vocal Communication			
Frequency and Quality of Social Overtures	Social Reciprocity/ Language		
Frequency and Quality of Social Responses			
Verbal Exchanges on a Topic			
Offering Information			
Stereotyped and Echoed Speech	Restricted and Repetitive Behaviors		
Unusual Sensory Interests			
Hand and Finger or Complex Body Mannerisms & Self Injurious Behaviors			
Unusually Repetitive Interests or Behaviors			
Play with Objects			
Social Engagement in Play Activities/Interaction	Other Abnormal Behaviors		
Activity Level			
Disruptive Behavior/Irritability			
Anxious Behaviors			

Figure 2. Visual depiction of BOSCC items, domains, and totals.

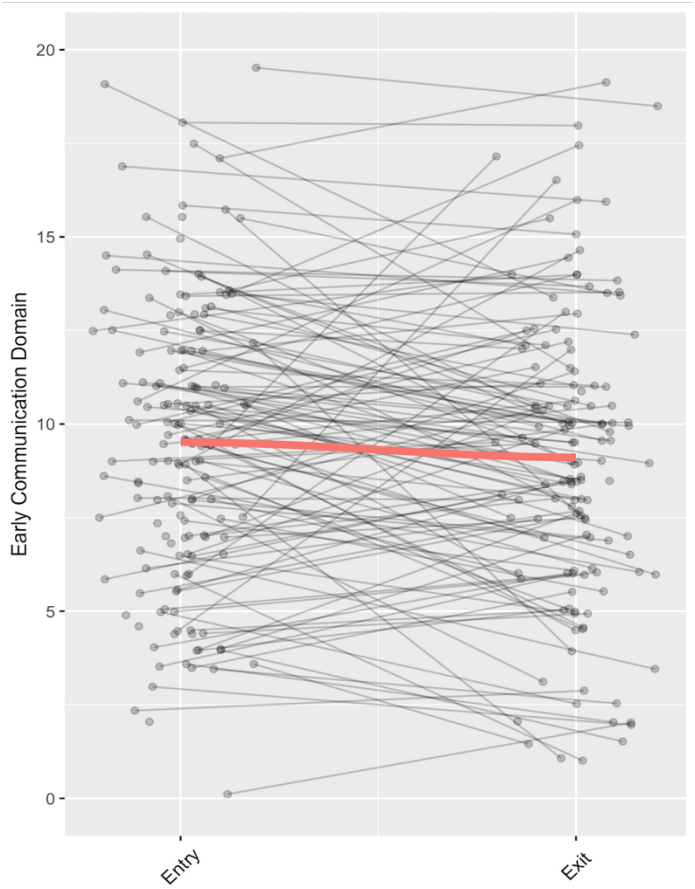

Figure 3. Decrease in Early Communication domain scores over 4.5-months.

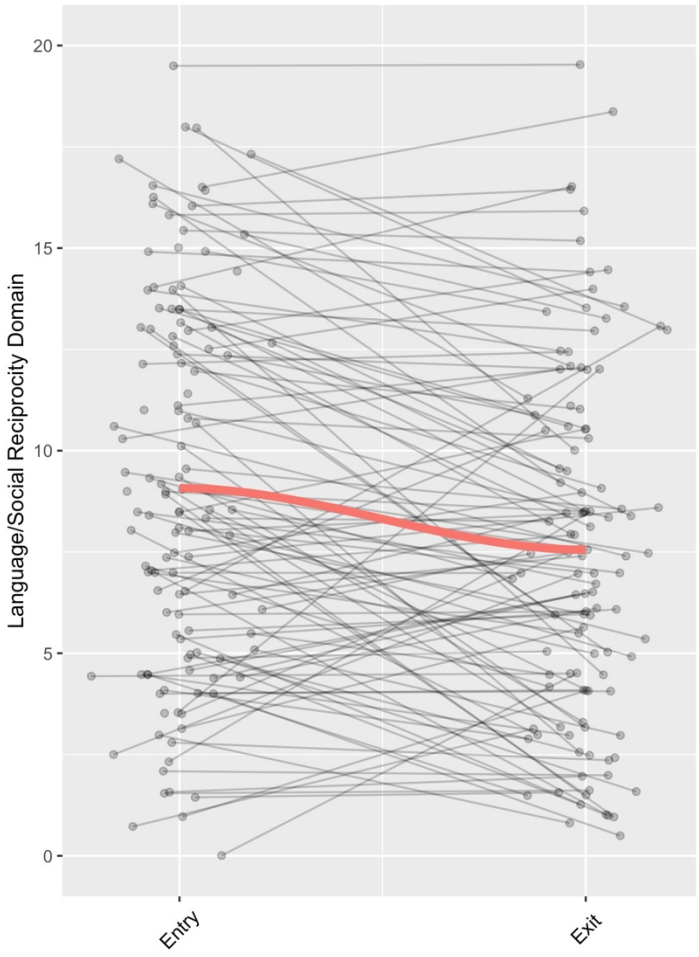

Figure 4. Decrease in Language/Social Reciprocity domain scores over 4.5-months.

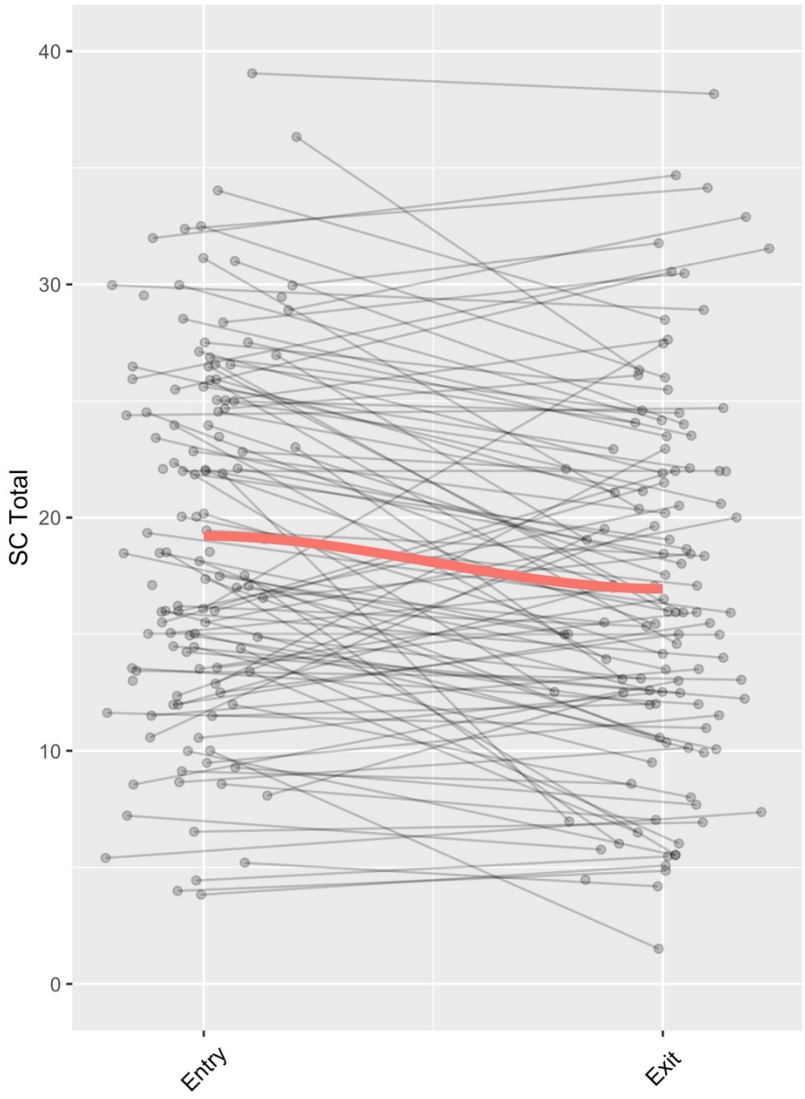

Figure 5. Decrease in combined Social Communication domain scores over 4.5-months.

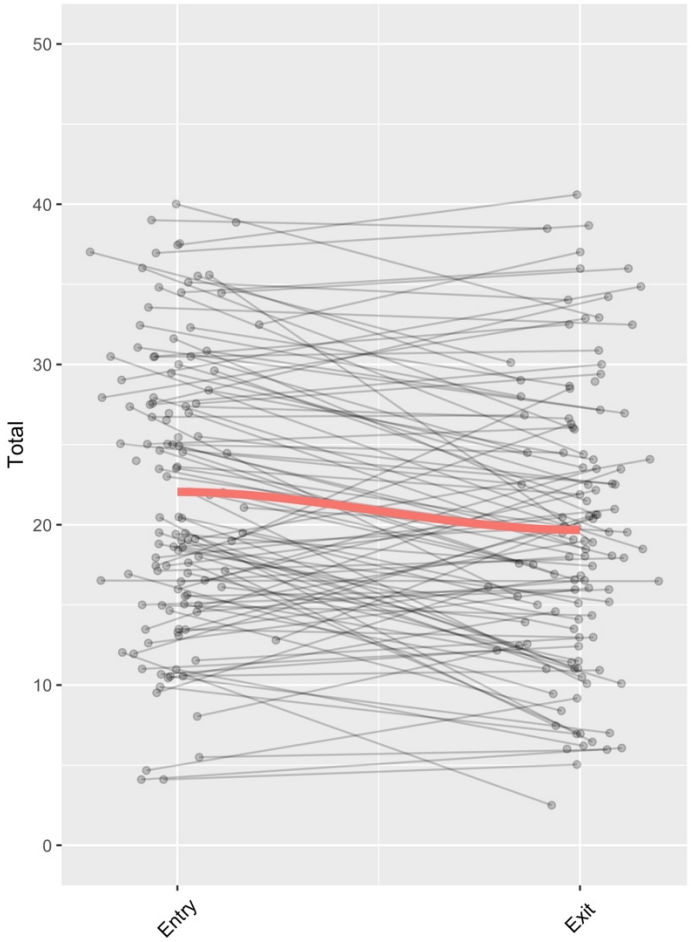

Figure 6. Decrease in combined Core Total scores over 4.5-months.

References

Aldred, C., Green, J., & Adams, C. (2004). A new social communication intervention for children with autism: Pilot randomised controlled treatment study suggesting effectiveness. *Journal of Child Psychology and Psychiatry, and Allied Disciplines*, *45*(8), 1420–1430. https://doi.org/10.1111/J.1469-7610.2004.00848.X

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

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. /pmc/articles/PMC2880930/

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*, *125*(1). https://doi.org/10.1542/peds.2009-0958

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 2015 45:9*, *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*, *144*(3). 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 2008 39:5*, *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*, *130*(5). https://doi.org/10.1542/peds.2011-3668

Green, J., Charman, T., McConachie, H., Aldred, C., Slonims, V., Howlin, P., Le Couteur, A., Leadbitter, K., Hudry, K., Byford, S., Barrett, B., Temple, K., Macdonald, W., & Pickles,

A. (2010). Parent-mediated Communication-Focused Treatment in children with autism (PACT): A randomised controlled trial. *The Lancet*, *375*(9732), 2152–2160. https://doi.org/10.1016/s0140-6736(10)60587-9

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

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: *Https://Doi.Org/10.1177/1362361318793253*, *23*(5), 1176–1185. https://doi.org/10.1177/1362361318793253

Kim, S. H., Grzadzinski, R., Martinez, K., & Lord, C. (2019). 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

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.

Magiati, I., Moss, J., Yates, R., Charman, T., & Howlin, P. (2011). Is the Autism Treatment Evaluation Checklist a useful tool for monitoring progress in children with autism spectrum disorders? *Journal of Intellectual Disability Research*, *55*(3), 302–312.

https://doi.org/10.1111/J.1365-2788.2010.01359.X

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

Mundy, P., Sigman, M., & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disorders*, *20*(1), 115–128. https://doi.org/10.1007/BF02206861

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*, *50*(8). 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

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.

Rogers, S. J., & Vismara, L. A. (2008). Evidence-Based Comprehensive Treatments for Early Autism. *Journal of Clinical Child and Adolescent Psychology : The Official Journal for the*

Society of Clinical Child and Adolescent Psychology, American Psychological Association, Division 53, *37*(1), 8. https://doi.org/10.1080/15374410701817808

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

Shumway, S., Farmer, C., Thurm, A., Joseph, L., Black, D., & Golden, C. (2012). The ADOS calibrated severity score: relationship to phenotypic variables and stability over time. *Autism Research : Official Journal of the International Society for Autism Research*, *5*(4), 267–276. https://doi.org/10.1002/aur.1238

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

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

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