

Measuring Change in Social Communication Behaviors:  
Reliability, Validity, and Application

Rebecca Grzadzinski

Submitted in partial fulfillment of the  
requirements for the degree of  
Doctor of Philosophy  
under the Executive Committee  
of the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2018

©2018  
Rebecca Grzadzinski  
All Rights Reserved

## ABSTRACT

Measuring Change in Social Communication Behaviors:

Reliability, Validity, and Application

Rebecca Grzadzinski

*Purpose:* The field of Autism Spectrum Disorder (ASD) intervention research is in need of treatment response measures that are sensitive to change and flexible enough to be used across studies. The Brief Observation of Social Communication Change (BOSCC) was developed to address this need. The purpose of this work is to examine the initial reliability and validity of the BOSCC in two samples of children with ASD.

*Method:* In a sample of 56 children participating in ongoing early intervention, the primary objectives of Study 1 were to 1) determine items for inclusion in the BOSCC coding scheme, 2) explore the relationships among items using factor analysis, 3) assess inter-rater and test-retest reliability, and 4) explore change over time. Using a sample of school-age, minimally-verbal children, the primary objectives of Study 2 were to extend the results of Study 1 to a new sample to 1) assess BOSCC changes over time, 2) compare changes in BOSCC to clinician determinations of improvement, 3) examine the relationship between change in BOSCC scores with changes in baseline cognitive skills, adaptive functioning, and ASD severity, and 4) compare changes in BOSCC scores in children who did and did not change on other standard measures.

*Results:* Study 1 revealed that the BOSCC has high to excellent inter-rater and test-retest reliability and shows convergent validity with measures of language and communication skills. The BOSCC Core total demonstrated statistically significant amounts of change over time while the ADOS Calibrated Severity Score over the same period did not. Results of Study 2 confirmed

excellent inter-rater reliability but the BOSCC did not change significantly over time. Most children were identified by clinicians as improving in response to treatment. However, only 15% of children changed significantly on the BOSCC over 16 weeks of intervention.

*Limitations:* Both studies had small samples of predominantly male, Caucasian children. When interpreting the results of these studies, it is important to consider the differences between samples, including the shorter time of treatment and more cognitively and language impaired children in Study 2.

*Conclusions:* These studies are a first step in the development of a novel outcome measure for social-communication behaviors with applications to clinical trials and longitudinal studies. Future work should continue to explore the benefits and limitations of the BOSCC in larger independent samples.

## **TABLE OF CONTENTS**

<u>LIST OF TABLES AND FIGURES</u>	<u>iv</u>
<u>ACKNOWLEDGMENTS</u>	<u>vi</u>
<u>DEDICATION</u>	<u>viii</u>
<u>INTRODUCTION</u>	<u>1</u>
WHAT IS AUTISM SPECTRUM DISORDER?	1
MEASURING PSYCHOLOGICAL CONSTRUCTS SUCH AS SOCIAL COMMUNICATION	3
THE COMPLEXITY OF MEASURING SOCIAL COMMUNICATION AND OTHER ASD SYMPTOMS	5
DO SOCIAL COMMUNICATION SYMPTOMS CHANGE OVER TIME?	7
THE CURRENT STATE OF ASSESSING CHANGE	8
COMMONLY USED MEASURES ASSESS CHANGE IN NON-ASD SYMPTOMS	10
COMMONLY USED MEASURES ASSESS CHANGE IN HIGHLY SPECIFIC BEHAVIORS	12
MEASURING CHANGE IN BROAD SOCIAL COMMUNICATION SKILLS	13
COMMONLY USED MEASURES RELY ON PARENT OR CLINICIAN REPORT	17
COMPLICATIONS OF ASSESSING SOCIAL COMMUNICATION CHANGE	18
THE BRIEF OBSERVATION OF SOCIAL COMMUNICATION CHANGE (BOSCC)	18
OBJECTIVES OF THIS WORK	19
<u>STUDY 1: INITIAL RELIABILITY AND VALIDITY METHOD</u>	<u>21</u>
PARTICIPANTS	21
PRIMARY MEASURE (BOSCC)	22
ADDITIONAL MEASURES	26

<b>STUDY 1 DATA ANALYSIS</b>	<b>29</b>
<b>STUDY 1 RESULTS</b>	<b>33</b>
INTER-RATER RELIABILITY	33
TEST-RETEST RELIABILITY	34
VALIDITY	34
POST-HOC ANALYSES	36
<b>STUDY 2: APPLICATION OF BOSCC TO AN INDEPENDENT SAMPLE METHOD</b>	<b>36</b>
PARTICIPANTS	36
PRIMARY MEASURE	38
ADDITIONAL MEASURES	40
<b>STUDY 2 DATA ANALYSIS</b>	<b>42</b>
<b>STUDY 2 RESULTS</b>	<b>43</b>
INTER-RATER RELIABILITY	43
CHANGE OVER TIME ACROSS MEASURES	43
BOSCC AND CGI RESPONDERS	44
BOSCC AND LEITER, PPVT-4, AND RBS-R RESPONDERS	45
ENTRY MEASURES AND CORRELATIONS WITH BOSCC	46
<b>DISCUSSION</b>	<b>47</b>
DOES THE BOSCC CAPTURE CHANGE OVER TIME?	47
RELIABILITY AND FACTOR STRUCTURE OF THE BOSCC	51
THE RELATIONSHIP BETWEEN THE BOSCC AND OTHER MEASURES AND A SINGLE TIME POINT	53
ADDITIONAL BENEFITS OF THE BOSCC	54

<b>ADDITIONAL CONSIDERATIONS</b>	<b>56</b>
<b>LIMITATIONS</b>	<b>57</b>
<b>FUTURE DIRECTIONS AND CONCLUSION</b>	<b>59</b>
<b>TABLES &amp; FIGURES</b>	<b>61</b>
<b>REFERENCES</b>	<b>84</b>
<b>APPENDIX</b>	<b>100</b>

## LIST OF TABLES AND FIGURES

### Tables

<u>Table 1.</u> <i>Study 1: Background and First Observation Information (n=56)</i> .....	61
<u>Table 2.</u> <i>Study 1: Information about Assessments Gather</i> .....	62
<u>Table 3.</u> <i>Study 1: Brief Observation of Social Communication Change (BOSCC)</i> <i>Exploratory Factor Analysis Model Comparison</i> .....	63
<u>Table 4.</u> <i>Study 1: 1, 2, and 3-Factor Model Loadings for Brief Observation of Social</i> <i>Communication Change Items</i> .....	64
<u>Table 5.</u> <i>Study 1: Inter-Rater Reliability for Domains (n=28)</i> .....	65
<u>Table 6.</u> <i>Study 1: Inter-Rater ICCs for Individual Items (Averaged across A and B) and</i> <i>Percent Agreement between Raters within 1 point</i> .....	66
<u>Table 7.</u> <i>Study 1: Test-Retest for Domains (n=20)</i> .....	67
<u>Table 8.</u> <i>Study 1: Test-Retest ICCs for Individual Items (Averaged across A and B)</i> .....	68
<u>Table 9.</u> <i>Study 2: Background Information (n=78)</i> .....	69
<u>Table 10.</u> <i>Study 2: Inter-Rater Reliability (n=22)</i> .....	70
<u>Table 11.</u> <i>Study 2: Inter-Rater ICCs for Individual Items (Averaged across A and B) and</i> <i>Percent Agreement between Raters within 1 point (n=22)</i> .....	71
<u>Table 12.</u> <i>Study 2: BOSCC, Leiter, and PPVT-4 Change Over Time</i> .....	72
<u>Table 13.</u> <i>Study 2: Frequency CGI and BOSCC Core Responders</i> .....	73
<u>Table 14.</u> <i>Study 2: T-Test CGI-I Groups and BOSCC Change from TP 1 to TP 3</i> .....	74
<u>Table 15.</u> <i>Study 2: Logistic Regression for Entry Measures Predicting CGI-I and BOSCC</i> <i>Core Responder Status</i> .....	75



<u>Table 16.</u> <i>Study 2: Leiter, PPVT-4, and RBS-R Change Groups and Change in BOSCC Domains from TP 1 to TP 3</i> .....	76
<u>Table 17.</u> <i>Study 2: BOSCC Scores and BOSCC Change Correlations with Entry Measures: ADOS, VABS, IQ and Age</i> .....	77
<b>Figures</b>	
<u>Figure 1.</u> <i>BOSCC Example Item</i> .....	78
<u>Figure 2.</u> <i>Study 1: Distributions for 12 Core BOSCC Items (Averaged across Segment A and B)</i> .....	79
<u>Figure 3.</u> <i>Study 1: Exploratory Factor Domains</i> .....	80
<u>Figure 4.</u> <i>Study 1: Responder Groups Defined by MSEL, VABS, or ADOS-2 in Early Intervention Studies</i> .....	81
<u>Figure 5.</u> <i>Study 1: Responder Groups defined by Clinician Global Impression (CGI) in Community-Based Intervention Study</i> .....	82
<u>Figure 6.</u> <i>Study 2: Item Distributions: Averages Across Segments A and B</i> .....	83

## ACKNOWLEDGMENTS

I would especially like to thank Dr. Catherine Lord, my amazing mentor. Cathy's expertise, continual guidance, professional and personal support provided me with the tools necessary to complete this dissertation. Cathy's dedication to me and my undergraduate and doctoral studies has given me a training model I hope to emulate. Her personal dedication to my professional growth and advocacy for me and all her trainees, has made me feel blessed to have learned from her. I look forward to continuing to learn from Cathy as she mentors me in my future endeavors.

I would also like to thank Drs. Bonanno, Jahromi, Rosa, and Tottenham, for their individual contributions to this project and participation on my committee. Their guidance through the process and feedback on this research has been invaluable. I would like to thank Dr. George Bonanno for his sustaining dedication to me throughout the course of my graduate studies and for being my connection with the clinical psychology department at Teachers College, Columbia University. In addition, I am eternally grateful for Dr. Bonanno's willingness to invest himself in my project as well as his invaluable insights into scientific methodology of this work. I would also like to thank Dr. Laudan Jahromi for her commitment to this project, insights into the ASD and special needs population, and for her supportive demeanor. It was truly a pleasure to learn from Dr. Jahromi, whose knowledge and positive attitude I hope to carry with me throughout my career. I also owe many thanks to Dr. Dinelia Rosa. I have learned so much from her clinical expertise and I thank her for challenging me to think beyond where I thought I was capable. I am also grateful for her kindness in guiding me through the clinical psychology program and her commitment to my professional development. Last, I would like to thank Dr. Nim Tottenham for

supporting me for the entirety of my graduate studies and from whose teaching style I hope to follow. Her discerning suggestions for this work have been truly invaluable.

Much gratitude is extended to Dr. Andrew Pickles for statistical consultation on these projects. I have truly enjoyed learning from Dr. Pickles and the knowledge he has shared will be carried with me indefinitely. In addition, I would like to acknowledge the many colleagues that have contributed to the development and coding of the Brief Observation of Social Communication Change (BOSCC) including Eliana Ajodan, Nurit Benrey-Grinberg, Juliana Boucher, Emily Campi, Themba Carr, Morgan Cohen, Costanza Colombi, Catherine Dick, Sarah Dufek, Kyle Frost, Gabrielle Gunin, Michelle Heyman, Natalie Hong, Rebecca Jones, So Hyun (Sophy) Kim, Cassandra Martinez, Allison Megale, Sophie Manevich, Kelly McGuire, Anna Marie Paolicelli, Lauren Pepa, Gabbie Ranger-Murdock, and Melanie Silverman. I would also like to thank Shanping Qui for her data support. Sincerest gratitude is extended to Susan Skrelja for her unending personal and professional support.

Portions of this proposal have been published in the *Journal of Autism and Developmental Disorders*. A Dennis Weatherstone pre-doctoral fellowship from Autism Speaks awarded to Rebecca Grzadzinski supported this work. Additional funding sources for data include grants awarded to Dr. Catherine Lord from NIMH (R01MH081757, 1RC1MH089721, R01RFAAMH14100, and R01MH078165), HRSA (UA3MC11055), and Autism Speaks (5766) as well as a grant awarded to Dr. Connie Kasari (R01HD073975).

## DEDICATION

I dedicate this work to my family. To my husband, Oliver, whose unending support made this work as well as all professional and personal pursuits possible. To my siblings, Allison, Sarah, Jacob, and Ian for their encouragement which has provided grounding to me during chaotic times. To my parents, Linda Galesewicz, Gerald Grzadzinski, Renee Clark, and Daniel Galesewicz, whose enduring reassurance pushed this work to fruition. I would also like to extend special gratitude to Leah and Logan Grzadzinski, who have been the inspiration for my commitment to improving the lives of those with autism spectrum disorder and their families.

Measuring Change in Social Communication Behaviors:  
Reliability, Validity, and Application

*“Measure what is measurable, and make measurable what is not.”*

*-Galileo Galilei*

*What is Autism Spectrum Disorder?*

Autism Spectrum Disorder (ASD) is one of the most commonly diagnosed childhood neurodevelopmental disorders with an estimated prevalence rate as high as 1 in 68 children (APA, 2013; Baio, 2014). ASD, like all psychopathology, is defined by the presence of atypical behavior and/or the absence of typical behavior (APA, 2013). Specifically, ASD is defined by both deficits in social communication skills as well as the presence of restricted, repetitive patterns of behavior or interests (RRB; APA, 2013). Recognition that these symptoms cluster together in some individuals, led to the diagnosis we now call ASD (Asperger, 1979; Kanner, 1967). The social communication and RRB constructs have been empirically validated in the ASD population (Guthrie, Swineford, Wetherby, & Lord, 2013; Mandy, Charman, & Skuse, 2012; Shuster, Perry, Bebko, & Toplak, 2014). Each domain is composed of specific behaviors. Social communication refers to behaviors associated with sending and receiving messages from one individual to another or differences in how one person interacts with another. Social communication deficits in ASD refers to specific behaviors such as impairments in non-verbal communication (e.g., use of eye contact, facial expressions, and gestures), social-emotional reciprocity (e.g., reduced sharing of emotions, limited reciprocal conversation), and development and maintenance of social relationships (e.g., challenges understanding social norms and making friends; APA, 2013). Within the RRB domain, specific behaviors include an interest in unusual

objects (e.g., an interest in street signs), an excessive interest in a particular topic (e.g., *Thomas the Tank Engine*), an interest in the sensory components of materials or people (e.g., smelling toys), hand/finger/other complex mannerisms (repetitive motor movements), or repetitive use of objects (e.g., lining up objects repetitively).

Researchers have explored the factor structure of these separate constructs. Studies of social communication behaviors in ASD and non-ASD populations have typically found that social communication behaviors cluster on one common factor (Frazier et al., 2012; Georgiades et al., 2007; Mandy, Charman & Skuse, 2012; Norris, Lecavalier & Edwards, 2012). However, one recent study found two separate social communication factors: “basic social communication,” predominantly composed of non-verbal social behaviors (e.g., use of eye gaze, gestures, and facial expressions), and “interaction quality,” composed of behaviors related to overall rapport and reciprocal aspects of an interaction (Bishop, Havdahl, Huerta, & Lord, 2016). Similarly, some research suggests that there is more than one factor within the RRB construct (Bishop et al., 2013; Cuccaro et al., 2003; Georgiades, Papageorgiou, & Anagnostou, 2010; Lam, Bodfish, & Piven, 2008). This research has demonstrated that, while the constructs of social communication and RRB are still being understood, they are composed of many specific behaviors that may all need to be measured discretely. Adequate operationalization of each discrete social communication and RRB behavior is necessary to ensure that the broader constructs are measured sufficiently.

### *Measuring Psychological Constructs such as Social Communication*

The ways in which the field of ASD research systematically measures social communication and RRB behaviors is consistent with the method of measurement used for other psychological phenomena, though ASD-specific considerations are necessary. In all psychological measurement, there are several qualities that must be considered when deciding the best method for quantifying behavior (AERA, APA, & NCME, 2014). How the information is gathered is key because behaviors can be influenced by the chosen methods (AERA, APA, & NCME, 2014). This may be particularly relevant for the measurement of social communication behaviors because these behaviors are transactional, or related to another person, as is inherent in social behavior.

Typical methods for measurement in psychology include direct observation, testing (e.g., standardized), questionnaires (e.g., self or other report), and interviews. All of these methods have been used in the measurement of social communication behaviors. For example, the Test of Pragmatic Language (Phelps-Terasaki & Phelps-Gunn, 2007) assesses social aspects of communication using a standardized testing procedure. Parent report questionnaires such as the Social Responsiveness Scale (SRS; Constantino, 2002), Social Communication Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003), and the Children's Communication Checklist (Bishop, 2006) all measure aspects of, and deficits in, social communication behavior. Direct observational methods of social communication include the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012a; Lord, Luyster, Gotham, & Guthrie, 2012), the Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner), and the Psychoeducational Profile (Schopler, Lansing, Reichler, & Marcus, 2005). Interviews with caregivers about social communication skills are also available, such as the Autism Diagnostic Interview, Revised (ADI-

R; Lord, Rutter, & Le Couteur, 1994) and the Diagnostic Interview for Social and Communication Disorders (DISCO; Wing, Leekam, Libby, Gould, & Larcombe 2002). Each method has advantages and disadvantages (see Lord, Corsello, & Grzadzinski, 2014). For example, while questionnaires may be a useful, relatively quick method by which to gather information about a child's social communication behaviors, they rely on the ability of the reporter to understand the questions posed and to accurately report their observations or impressions. In contrast, in-person testing conditions may produce results that minimize some aspects of measurement error through administration of standardized prompts and environmental arrangements, but the generalization of these results to more naturalistic contexts may be limited.

Perhaps the most objective method, direct observation, allows for the opportunity to assess social communication behaviors in naturalistic settings, adding to the ecological validity of this method of measurement (AERA, APA, & NCME, 2014). However, quantifying social communication behaviors in such unstructured settings can be challenging (Wetherby, 2008), given the limited standardization of the social context and the reliance on the observer to quantify behaviors accurately and reliably. Given these limitations, when considering the use of direct observation, it is essential that the measure demonstrate strong psychometric properties (reliability and validity; AERA, APA & NCME, 2014; Wetherby, 2008).

All psychological measurement tools must withstand rigorous psychometric tests, including reliability and validity assessments, to ensure that the tools align as closely as possible with the behaviors being measured (AERA, APA, & NCME, 2014). Reliability is the consistency of a measure, for example, over time (Test-retest) or between coders (Inter-rater). Reliability ensures that the tool is measuring the behavior with as little error as possible. Reliability assessments use statistical methods to determine the degree to which a measure is



consistent, usually by calculating reliability coefficients, or percent agreement, for scales (e.g., Intraclass Correlation Coefficients). Especially for direct observation, clear definitions of the behaviors being measured are essential to ensure that coders can reliably assess behaviors across individuals and time. Though the creation of these definitions may be tedious and time-consuming, they guarantee the quality of the measurement.

Another quality to consider is the validity of a measure. Validity refers to the extent to which the measure truly assesses the construct (e.g. social communication) that it intends to measure (construct validity). Evaluations of an instrument's quality examine whether it appears to quantify what it intends to (face validity), whether it aligns with other assessments of the same or similar variables (convergent validity), and whether the instrument accurately discriminates between what it intends to measure and other variables that may impact it (divergent validity). Measures developed to assess symptoms of ASD need to meet these standard criteria of reliability and validity.

### *The Complexity of Measuring Social Communication and Other ASD Symptoms*

Using various methods including direct observation, questionnaires, and structured interviews, the field of ASD has excelled in quantifying social communication and RRB behaviors to both define the phenotype of ASD and differentiate ASD from other forms of developmental psychopathology (Esler et al., 2015; Gotham, Pickles, & Lord, 2009; Kim & Lord, 2010; Lord et al., 2012a; Lord, Luyster, Gotham, & Guthrie, 2012; Shumway et al., 2012; Wetherby & Prizant, 2002). For example, most diagnostic instruments can accurately identify ASD in >80% of individuals with clinical best-estimate expert diagnoses of ASD across age and language levels and can distinguish ASD from other neurodevelopmental disorders at an equally

high frequency (Chlebowski, Green, Barton, & Fein, 2010; Gotham et al., 2008; Lord et al., 2012b, c; Matson et al., 2009). Over the last 20 years, significant advances have been made in the screening and identification of ASD (Daniels & Mandell, 2013; Daniels, Halladay, Shih, Elder, & Dawson, 2014; Dawson & Bernier, 2013; Guthrie, Swineford, Nottke, & Wetherby, 2013; Woolfenden, Sarkozy, Ridley, & Williams, 2012), leading to reductions in age of diagnosis and better understanding of the social communication and RRB constructs within an ASD diagnosis (Lord et al., 2012b; Oosterling et al., 2010; Robins et al., 2014; Wetherby, 2008).

Understanding the social communication and RRB constructs of ASD has been particularly important because individuals with ASD vary in type, quality, and frequency of social communication deficits as well as the type and intensity of RRBs. These ASD-specific symptoms differ based on age and are related to cognitive and language abilities (Boucher, 2012; Lord et al., 2006; Ingram, Takahashi, & Miles, 2008; Matson & Shoemaker, 2009; Tager-Flusberg, Paul, & Lord, 2001). Children with ASD vary greatly in both intellectual and language functioning (Boucher, 2012; Matson & Shoemaker, 2009; Tager-Flusberg, Paul, & Lord, 2001), ranging from profoundly intellectually impaired to intellectually gifted (Matson & Shoemaker, 2009) and from no verbal speech to verbal fluency (Boucher, 2012; Tager-Flusberg et al., 2001). For young, minimally-verbal children with ASD, symptom presentations include decreased frequency or lack of a response to one's name or other bids for social engagement, limited use of nonverbal communication (e.g., pointing, facial expressions, eye gaze), decreased initiations for social engagement, and diminished joint attention skills (Reznick, Baranek, Reavis, Watson, & Crais, 2007). For older, highly verbal children, ASD symptoms most often involve continued deficits in nonverbal aspects of communication (e.g., use of facial expressions and eye gaze) as well as the impairments in the reciprocal aspect of social interchanges (e.g., conversation, use of

comments and social chit-chat) and understanding of social nuances (APA, 2013; Lord et al., 2012c).

Verbal skills are often correlated with overall intellectual functioning and together, they relate to the presentation of ASD symptoms. For example, certain RRBs have been found to be more frequent in individuals with lower cognitive abilities (Bishop et al., 2013; Bishop, Richler, & Lord, 2006; Lam, Bodfish, & Piven, 2008). In addition, many children with ASD, especially as they age, may have co-occurring conditions (e.g., anxiety, depression, Attention Deficit/Hyperactivity Disorder) that may impact the presentation of ASD symptoms (Kaat, Gadow, & Lecavalier, 2013; Lever & Geurts, 2016; Salazar et al., 2015; Wallace, Budgett, Charlton, 2016). The boundaries between ASD and other disorders can be tricky to disentangle, particularly when distinguishing social communication impairments that could be the result of ASD, another disorder, or both (Matson & Cervantes, 2014; Ronald, Larsson, Anckarsatar, & Lichtenstein, 2014; Simonoff et al., 2008; Sprenger et al., 2014). The variety of ASD symptom presentations as well as the complexity resulting from the effects of age, language level, co-morbidities, and cognitive skills, all contribute to the heterogeneous nature of the ASD phenotype. This heterogeneity also highlights the need for measures of social communication deficits that account for these complexities associated with age, language, and cognitive level.

### *Do Social Communication Symptoms Change Over Time?*

While the field has advanced in quantifying symptoms of ASD for the purposes of diagnosis at a single moment in time, symptoms of ASD are not stagnant over time. Some recent studies have suggested that approximately 5-10% of young children diagnosed with ASD may “outgrow” the ASD diagnosis by adulthood (Helt et al., 2008; Fein et al., 2013, Anderson, Liang,

& Lord, 2014). Yet, for the vast majority of individuals, the continued presence of deficits in social communication as well as RRBs leads to the maintenance of a diagnosis of ASD throughout one's lifetime (Anderson, Liang, & Lord, 2014; Fein et al., 2013; Helt et al., 2008; Lord et al., 2006). Despite the continued presence of an ASD diagnosis, the quality, intensity, and frequency of ASD symptoms, including social communication behaviors, change over time (Gotham, Pickles, & Lord, 2012; Lord, Bishop, & Anderson, 2015). This contrasts with social communication behaviors in the general population, which are thought to be relatively consistent across time and development (Constantino et al., 2003; Hoekstra, Bartels, Verweij, & Boomsma, 2007; Robinson et al., 2011). Changes observed in social communication behaviors over time in individuals with ASD may be due to general maturation or because many ASD-specific interventions directly target social communication behaviors (Hanson, Blakely, Dolata, Raulston, & Machalicek, 2014; Reichow & Volkmar, 2010; Rogers & Vismara, 2008). For example, interventions such as Early Start Denver Model (ESDM), Joint Attention Symbolic Play Engagement and Regulation (JASPER), and Parent-Mediated Communication-Focused Treatment (PACT) all target aspects of social communication deficits seen in ASD. These targeted goals include increasing the frequency and quality of interpersonal exchanges, joint attention, reciprocal social engagement, interactive communication and positive affect (Dawson et al., 2010; Green et al., 2010; Kasari, Gulsrud, Paparella, Helleman, & Berry, 2015). However, quantifying these changes has been challenging for many reasons.

### *The Current State of Assessing Change*

Though quantifying broad deficits in social communication behaviors has been fruitful for diagnostic determination, measuring changes in social communication behaviors over time

has been less successful (Anagnostou et al., 2015; Fletcher-Watson & McConachie, 2015; McConachie et al., 2015). Currently, there are few outcome measures available to identify changes social communication behaviors over short periods (e.g., months), though most treatments for ASD focus on improvement in these skills (Anagnostou et al., 2015; Fletcher-Watson & McConachie, 2015; McConachie et al., 2015). Without appropriate measures of change in social communication, the field of ASD intervention research is limited in its ability to identify efficacious interventions (Bolte & Diehl, 2013; Danial & Wood, 2013; McConachie et al., 2015).

A recent review noted that in 195 behavioral intervention trials for ASD, over 200 different measurement tools were used to assess treatment response (Bolte & Diehl, 2013). Sixty percent of these tools were used in only a single study, with only three tools used in more than two percent of studies (Bolte & Diehl, 2013). In addition, another review noted that most commonly used tools have little validity as outcome measures (McConachie et al., 2015). After a series of consensus meetings held by Autism Speaks, a panel of ASD experts determined that only a handful of existing measures are appropriate for identifying treatment response in ASD (Anagnostou et al., 2015; Scahill et al., 2015). Among these recommended instruments are two of the most commonly used tools: the Aberrant Behavior Checklist (ABC; Aman, Singh, Stewart, & Field, 1985) and the Vineland Adaptive Behavior Scales (VABS; Sparrow, Cicchetti, & Balla, 2005). It should be noted that these tools were not recommended for use as outcome measures in a contradicting review from researchers in the United Kingdom (McConachie et al., 2015). A third commonly used measure, the Clinical Global Impressions (CGI; Guy, 1976), was not recommended for use by either group of researchers.

### *Commonly Used Measures Assess Change in Non-ASD Symptoms*

Though not specific to ASD symptoms or social communication, the ABC, CGI, and VABS are some of the most commonly used measures to assess treatment response in individuals with ASD. The ABC captures behavior across a range of psychiatric symptoms including subscales in the domains of irritability, lethargy, stereotypic behavior, hyperactivity and inappropriate speech (Aman, Singh, Stewart, & Field, 1985). The CGI is a clinician-rated measure with two 7-point scales to rate both the severity of symptoms (ranging from 1 “normal” to 7 “among the most extremely ill patients”) as well as amount of improvement (from 1 “very much improved” to 7 “very much worse”; Guy, 1976). The ABC and CGI are most commonly used in psychopharmacology trials (Bolte & Diehl, 2013). The VABS is a parent-completed interview that assesses a range of adaptive behaviors, or skills needed for daily functioning (Sparrow, Cicchetti, & Balla, 2005). VABS parent and teacher questionnaire rating forms are also available. The VABS measures behaviors in the domains of Socialization, Communication, Daily Living, and Motor Skills (Sparrow, Cicchetti, & Balla, 2005). The updated edition of the VABS, the VABS-3 (Third Edition; Sparrow, Cicchetti, & Saulnier, 2016), includes several new items and updated norms. The VABS has been most commonly used to assess treatment response in psychological studies of children with ASD. The ABC, CGI, and VABS capture a broad range of psychopathology and functional impairment. Research suggests at least a small relationship between scores on the ABC and VABS and impairments in social communication, perhaps because social communication behaviors span across most areas of life (Capone, Grados, Kaufmann, Bernad-Ripoll, & Jewell, 2005; Frost, Hong, & Lord, 2017).

The Behavior Assessment System for Children (BASC-2; Reynolds & Kamphaus, 2006), the Social Skills Improvement System (SSIS; Gresham & Elliot, 2008), the Communication and

Symbolic Behavior Scales (CSBS; Wetherby & Prizant, 2002), and the Early Social Communication Scales (ESCS; Mundy et al., 2007) were also identified as appropriate for use as treatment response measures for ASD (Anagnostou et al., 2015; Scahill et al., 2015). The BASC-2 is a parent or teacher questionnaire used to rate a child's behavior in the realms of adaptive skills, behavioral symptoms, externalizing problems, internalizing problems, and school problems (Reynolds & Kamphaus, 2006). Two subscales, social withdrawal and social skills, most consistently identify changes in children with ASD (Lopata et al., 2010; Sim, 2006; Solomon, Ono, Timmer, & Goodlin-Jones, 2008). The SSIS is another parent and/or teacher report measure that quantifies behaviors in the realms of social skills and problem behaviors (Gresham & Elliot, 2008). Developed to measure preverbal communication skills in children between the ages of 12 and 24 months of age, the CSBS consists of a parent questionnaire and an hour-long, parent-child play observation coded by a clinician (Wetherby & Prizant, 2002). The last measure deemed appropriate for identifying change in children with ASD is the ESCS. The ESCS is an observation-based measure that quantifies the frequency of behaviors related to non-verbal communication, specifically behaviors related to joint attention (Kaale, Smith, & Sponheim, 2012). Though the ESCS, CSBC, and BASC are recommended for use, they have limitations (Anagnostou et al., 2015; Scahill et al., 2015). Some are only appropriate for a small range of mental or chronological ages. For example, the ESCS is appropriate for children up to 2 ½ years mental age (Kaale, Smith, & Sponheim, 2012) and the CSBS is only appropriate for children up to 2 years chronological age.

Other commonly used outcome measures, though not recommended by the expert panel (Anagnostou et al., 2015; Scahill et al., 2015), include assessments of cognitive skills (Anan, Warner, McGillivray, Chong, & Hines, 2008; Eapan, Črnčec, & Walter, 2013; Dawson et al.,

2010; Grindle et al., 2012; Rogers et al., 2012; Siller, Hutman, & Sigman, 2013; Zachor & Itzchak, 2010), such as the Leiter International Performance Scale- Revised (Leiter; Roid & Miller, 1997) or the Mullen Scales of Early Learning (MSEL; Mullen, 1995). Measures of cognitive skills are often used to assess co-occurring verbal or non-verbal developmental delays. In an ESDM trial, researchers found statistically significant improvements in MSEL receptive language after 1 year of treatment as well as statistically significant expressive and receptive improvements after two years (Dawson et al., 2010). Similarly, in another study, children who entered treatment with MSEL expressive language levels below 12 months, gained significantly more language over the course of intervention compared to those with higher language levels at entry (Siller, Hutman, & Sigman, 2013). These studies indicate that measures of cognitive functioning may be useful in assessing changes that are not specific to ASD symptoms.

Overall, many measures are used to assess cognitive, language, or adaptive functioning outcomes instead of ASD-specific symptoms or social communication behaviors (Anderson et al., 2007; Matson, 2007; Spence & Thurm, 2010; Warren et al., 2011; Wolery & Gerfinkle, 2002). While it is important for researchers to measure improvements in skills that are not specific to a diagnosis of ASD, these skills are often not directly targeted in intervention (Matson, 2007; Spence & Thurm, 2010; Wolery & Garfinkle, 2002). Therefore, measuring improvements in cognitive and adaptive functioning may be a beneficial way to quantify generalization of skills, but this is less informative for understanding how a child's ASD symptoms, those that have been specifically targeted in treatment, have improved.

#### *Commonly Used Measures Assess Change in Highly Specific Behaviors*

Alternatively, researchers may use treatment response measures that are highly specific to certain aspects of social communication, such as assessments of the frequency of joint attention,



rather than quantifying the broad range of social communication deficits targeted in intervention (Green et al., 2010; Kaale, Smith, & Sponheim, 2012; Kasari, Gulsrud, Freeman, Paparella, & Hellemann, 2012; Rogers et al., 2012; Yoder, Woynaroski, Fey, & Warren, 2014). For example, researchers create a measure that captures the frequency of a specific operationalized behavior that is one of many behaviors targeted in treatment, such as joint attention or imitation (Kaale, Smith, & Sponheim, 2012; Rogers et al., 2012). Measuring these specific behaviors over time is useful in quantifying changes in behaviors targeted in intervention, but these measures do not address changes in broader symptoms of ASD that may aid in understanding the overall effectiveness of an intervention. Synthesizing results across studies can also be difficult because researchers often operationalize highly specific behaviors differently across studies even when using the same or similar terminology to identify behaviors (Wolery & Garfinkle, 2002). Limiting outcome measures to behaviors that are highly specific to a certain treatment may also lead to accentuated interpretation of treatment effects that have not generalized beyond very specific behaviors (Yoder, Bottema-Beutel, Woynaroski, Chandrasekhar, & Sandbank, 2013).

#### *Measuring Change in Broad Social Communication Skills*

Though measures that focus on highly specific behaviors have limitations, research that has attempted to quantify changes in broader social communication skills has usually been unsuccessful (Brian, Smith, Zwaigenbaum, Roberts, Bryson, 2015; Dawson et al., 2010; Estes et al., 2015; Shumway et al., 2012; Thurm, Manwaring, Swineford, & Farmer, 2015). This may be explained by the common use of measures intended for diagnosing ASD rather than quantifying change in symptoms. Because an ASD diagnosis is usually stable over time (Lord et al., 2006; Woolfenden et al., 2012), current diagnostic and screening measures, which are intended to provide information about the presence or absence of the disorder, are typically not sensitive

enough to changes in ASD symptoms that occur over shorter periods of time (months as opposed to years) or in response to treatment (Anagnostou et al., 2015). This is in part because changes in social communication behaviors are often subtle, making it difficult to find measures that are sensitive enough to capture smaller, though potentially meaningful, changes (Anagnostou et al., 2015; Cunningham, 2012; Matson, 2007; McConachie et al., 2015; Yoder, Bottema-Beutel, Woynaroski, Chandrasekhar, & Sandbank, 2013).

The Autism Diagnostic Observation Schedule (ADOS-2; Lord, Luyster, Gotham, & Guthrie, 2012; Lord et al., 2012a), a measure intended for diagnostic purposes, has been used by researchers in an attempt to capture improvements in broad social communication skills. Using raw scores from the ADOS-2 has usually been unsuccessful in assessing changes, and, when changes have been identified, the clinical utility of these changes is difficult to interpret because ADOS-2 raw scores are not intended for use as interval data or for measuring change (Green et al., 2010; Gutstein, Burgess, & Montfort, 2007; Owley et al., 2001; Pickles et al., 2015).

Recently, researchers standardized the ADOS-2 scores creating the ADOS-2 Calibrated Severity Score (CSS; Esler et al., 2015; Gotham, Pickles, & Lord, 2009). The ADOS-2 CSS is a metric based on ADOS-2 raw totals, developed to quantify the severity of ASD symptoms on a 1 (least severe symptoms) to 10 (most severe symptoms) scale. The ADOS-2 CSS aims to provide a better way to assess the severity of ASD symptoms and quantify change over time (Gotham, Pickles, & Lord, 2009). Research indicates that the ADOS-2 CSS has been successful in identifying changes over the course of years (Gotham, Pickles, & Lord, 2012; Lord, Luyster, Guthrie, & Pickles, 2012). For example, when assessing change in ASD severity over the course of 13 years (from ages 2 to 15), one study found four trajectories: two trajectories that represented stability in severity over time (stable high severity, stable moderate severity) as well

as two trajectories that represented change (increasing severity and decreasing severity groups; Gotham, Pickles, & Lord, 2012). Similar patterns but with different proportions of participants in different trajectory groups were found for toddlers over the course of 1 ½ years (18 to 36 months of age; Brian, Smith, Zwaigenbaum, Roberts, & Bryson, 2015; Lord, Luyster, Guthrie, & Pickles, 2012). Despite these promising results, the ADOS-2 CSS has been less successful in identifying changes in ASD symptoms over shorter periods of time (Brian, Smith, Zwaigenbaum, Roberts, & Bryson, 2015; Dawson et al., 2010; Estes et al., 2015; Shumway et al., 2012; Thurm, Manwaring, Swineford, & Farmer, 2015). For example, in an ESDM intervention trial and in a longitudinal study of ASD severity in toddlers, the ADOS-2 CSS did not change over the course of one year (Dawson et al., 2010; Thurm, Manwaring, Swineford, & Farmer, 2015).

Another diagnostic measure researchers have used to assess change is the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994). The ADI-R is a semi-structured, standardized parent interview that gathers information about diagnostic symptoms of ASD and provides information about the presence of symptoms in the past (ever or between the ages of 4 and 5) as well as currently. Analyses of the ADI-R (Lord, Rutter, & Le Couteur, 1994) have been useful in identifying trajectories of change over the course of years (Lord, Bishop, & Anderson, 2015). For example, research has demonstrated that the ADI-R in combination with information about intellectual ability, can yield different trajectories of change in social skills and repetitive behaviors from ages 2 to 19 years old (Lord, Bishop, & Anderson, 2015). The ADI-R has also shown change from middle childhood through adulthood (from age 10 to age 53) in another sample (Seltzer et al., 2003). However, assessing change over shorter periods of time is less clear because studies exploring this question have often not used control groups or

changes are seen in both treatment and control groups, suggesting that developmental trajectories of change are not specific to treatment (Gutstein, Burgess, & Montfort, 2007; Sallows & Graupner, 2005).

When considering use of the ADOS-2 and ADI-R, the significant training required to be able to administer and score reliably is a hindrance. In addition, the ADOS-2 typically requires between 45-60 minutes while the ADI-R requires 90-150 minutes to administer and score. This is a substantial time commitment from both patients and the trained clinician. As a result, use of the ADOS-2 and ADI-R in large-scale, multi-site studies is often not feasible. Given the high time and resource burden as well as the limited evidence of sensitivity of change over short periods of time, the ADOS-2 and ADI-R are no longer recommended for use as treatment response measures (Anagnostou et al., 2015) though use of these tools was previously encouraged for this purpose (Cunningham, 2012; Matson, 2007). Despite limitations in measuring change over short periods of time, the ADOS-2 and ADI-R are some of the most well-validated, psychometrically-sound tools that quantify symptoms of ASD. Therefore, the benefits of the ADOS-2 and ADI-R for standardized diagnostic purposes is evident (De Bilt, 2004; Lord et al., 2000).

An additional challenge when attempting to use measures that capture a broad range of social-communication behaviors is that they are often confounded by co-occurring intellectual deficits and behavior and/or language problems (Hus, Bishop, Gotham, Huerta, & Lord, 2013). Though the ADOS-2 CSS has been found to be relatively independent of intellectual ability (Esler et al., 2015; Gotham, Pickles, & Lord, 2009), other measures of broad social communication skills, such as the Social Responsiveness Scale (Constnatio, 2002) are highly related to age, general behavior problems, language impairment, and cognitive skills. The

influence of these confounds may make it difficult to disentangle meaningful changes in ASD-specific social-communication behaviors from other non-ASD-specific behaviors. While it seems unrealistic to expect a measure of ASD symptoms to be completely uncorrelated with language and IQ, a measure that takes into account these aspects of an individual's presentation may lessen these confounds (McConachie et al., 2015).

#### *Commonly Used Measures Rely on Parent or Clinician Report*

Other measures commonly used to assess treatment response rely on caregiver or clinician report, such as the CGI (Busner & Targum, 2007). Use of these measures is limited because placebo effects are particularly strong for caregiver or clinician report measures (Anagnostou et al., 2015; Bolte & Diehl, 2013; Jones, Carberry, Hamo, & Lord, in press). For example, in a recent study tracking children over the course of eight weeks, parents reported significant decreases in problem behaviors as well as ASD symptoms even though no treatment was provided (Jones, Carberry, Hamo, & Lord, in press). These “placebo-like” effects may even outweigh more subtle changes that occur over time or in response to interventions (Guastella et al., 2015; Lord, Luyster, Guthrie, & Pickles, 2012; Owley et al., 2001). In a recent paper, caregiver-report measures of response to treatment were more related to the caregiver's belief that the child was receiving the experimental treatment than to the treatment itself (Guastella et al., 2015). This may be related to the use of measures that rely on the report of someone who is aware, and not intended to be unaware, of the child's treatment status, which leads to reporting bias (Wolery & Garfinkle, 2002). At other times the person rating improvement becomes “unblinded,” i.e., *becomes* aware of whether the child is receiving treatment and/or which treatment the child is receiving. For psychopharmacology trials, caregivers and clinicians are often aware if the child is experiencing significant side effects. Using parent or clinician report

measures provides valuable information about the parent or clinician's opinion of whether the child is improving, but they may be influenced by whether the parent or clinician is aware of the child's treatment status.

#### *Complications of Assessing Social Communication Change*

A further complexity relates to the heterogeneous ASD phenotype which makes measuring changes over time difficult because improvements across individuals may differ dramatically. Also, one would not necessarily expect every child to improve from every treatment, further complicating the interpretation of intervention trials (Georgiades et al., 2013; Grzadzinski, Huerta, & Lord, 2013; Hus, Pickles, Cook, Risi, & Lord, 2007; Ingram, Takahashi, Miles, 2008; Spiker, Lotspeich, Dimiceli, Myers, & Risch, 2002). Nevertheless, most treatments for children with ASD focus on decreasing core symptoms, especially social communication impairments (Rogers & Vismara, 2008), making the quantification of subtle social communication improvements particularly important for the measurement of outcomes associated with efficacious treatments, despite the inherent challenges in this endeavor (Anagnostou et al., 2015).

#### *The Brief Observation of Social Communication Change (BOSCC)*

The limitations of currently used measures make it difficult for clinicians and researchers to measure changes in broad social communication skills and determine the effectiveness of interventions. This may be related to why few ASD interventions have met standard criteria for efficacy (Chambless & Hollon, 1998; Danial & Wood, 2013; Levy, Mandell, & Schultz, 2009; Rogers & Vismara, 2008; Spreckley & Boyd, 2009; Wong et al., 2015). Given this critical need, researchers have begun to focus efforts on developing measures that are sensitive to change

(Fletcher-Watson & McConachie, 2015; McConachie et al., 2015). The Brief Observation of Social Communication Change (BOSCC) is an initial attempt to address the limitations of commonly used measures.

The BOSCC is a new measure consisting of specific items that were developed to identify changes in social communication behaviors over relatively short periods of time by quantifying subtleties in both the frequency and the quality of specific behaviors. The BOSCC is a coding scheme that was developed by modifying and expanding codes from the ADOS-2 (Lord, Luyster, Gotham, & Guthrie, 2012) to capture more subtle variations in behaviors. The goal of the BOSCC is to provide researchers and clinicians with an outcome measure that is flexible, easy to code, and minimally-biased by caregiver or clinician report. The BOSCC is flexible enough to be used across a variety of settings (e.g., across multi-site studies, in clinics or at home) and to be carried out and coded by clinicians or researchers who are blind to the child's treatment status and are new to ASD research (e.g., research assistants). There are plans to expand the BOSCC to a range of ages and developmental levels, however, the version of the BOSCC described in this work is applicable to minimally-verbal children.

#### *Objectives of This Work.*

The broad purpose of this work is to assess the reliability and validity of the BOSCC in two studies of children with ASD. The first study provides an opportunity to explore the initial psychometric properties and construct validity of the BOSCC in a sample of toddlers and preschoolers with ASD who were participating in ~6 months of early intervention. The second study provides another opportunity to assess the reliability and construct validity of the BOSCC while extending this exploration to a sample of school-age children who are also minimally

verbal (used fewer than 20 spontaneous words) over a shorter intervention period. This exploration uses a unique sample of very impaired children who are frequently neglected in treatment studies because they are often considered non-responders or slower to respond to treatment (Koegel, Shiratova, & Koegel, 2009; Sheinkopf & Siegel, 1998; Tager-Flusberg & Kasari, 2013). The second study also includes clinician ratings of improvement (CGI). Because clinicians base their determination of whether a child is improving (CGI-I) on a child's behavior during intervention sessions while, the BOSCC provides a blinded observation of behavior during play with a parent, these explorations also provide an opportunity to assess a child's ability to generalize skills outside of the intervention context (Yoder et al., 2013). The combination of these studies contributes to our understanding of the robustness and limitations of the BOSCC.

The specific aims of the first study (Study 1: Initial Reliability and Validity) are to 1) determine items for inclusion in the final BOSCC coding scheme through exploration of item correlations, 2) explore the relationship among items using factor analysis, 3) examine the BOSCC's psychometric properties, including inter-rater and test-retest reliability, and 4) provide evidence for validity of the BOSCC through explorations of change in BOSCC scores over time compared with changes in scores from other well-established standard measures.

The specific aims of the second study (Study 2: Application of BOSCC to an Independent Sample) are to 1) explore whether the BOSCC or other measures change significantly over time, 2) compare changes in BOSCC to clinician judgments of improvement (CGI), 3) compare changes in BOSCC scores in children who are and are not showing change on standardized measures of receptive language, cognitive ability, and repetitive behavior, and 4) examine the



relationship between changes in BOSCC scores with baseline cognitive skills, adaptive functioning, and ASD symptom severity.

## STUDY 1: INITIAL RELIABILITY AND VALIDITY

### Method

*Participants.* Fifty-six children (44 males) with a Best Estimate Clinical Diagnosis (BEC; Anderson, Liang, & Lord, 2014) of an Autism Spectrum Disorder (ASD) were included in this study. Diagnoses of ASD were determined based on thorough diagnostic evaluations, including administration of the Autism Diagnostic Interview-Revised (ADI-R; Lord, Rutter, & Le Couteur, 1994) and the Autism Diagnostic Observation Schedule (ADOS-2; Lord et al., 2012a; Lord, Luyster, Gotham, & Guthrie, 2012). All participants had elected to join various treatment studies (Kasari, Gulsrud, Freeman, Paparella, & Helleman, 2012; Rogers et al., 2012; Wetherby et al., 2014) depending on which studies were available at the time and parents provided written informed consent for their child to participate in the intervention trial. Children were then randomized into a treatment condition at the University of Michigan Autism and Communication Disorders Center (UMACC; n=49) or the Center for Autism and the Developing Brain (CADB; n=6), except for one participant. Data from this one child was extracted from an existing database of children whose parents had provided written informed consent for their child's clinical information/assessments to be included in an Institutional Review Board (IRB)-approved database.

All children included in this study except for one child were receiving some form of intervention while participating, either through the treatment condition in the clinical trials or elsewhere, though the interventions varied in frequency and type (see Kasari, Gulsrud, Freeman, Paparella, & Helleman, 2012; Rogers et al., 2012; Wetherby et al., 2014 for details regarding

intervention trials). For the one child who was not receiving any form of intervention, only one BOSCC observation was available; accordingly, this child was not included in analyses of change over time. For the purposes of this study, which focuses on the validity and reliability of the BOSCC, the effects of specific treatment conditions are not explored; future work will address this question. All children included in the study were between 1 and 5 years of age, walking independently, and using minimal spontaneous language (phrase speech or less, equivalent to ADOS-2 Module Toddler, 1, or 2), as is appropriate for the current BOSCC coding scheme (described below). See Table 1 for demographic and initial observation information.

*Primary measure (BOSCC).*

*BOSCC Development.* Over several versions of the BOSCC coding scheme, numerous codes and coding structures were generated and tested. A uniform (flat) distribution over the coding range for items was desirable because the aim of the BOSCC is to capture change within items over time. For example, if normal distributions were desired, most children would obtain items scores in the middle of the distribution (e.g. code of 3) across time. By creating a scale with more uniform distributions, the hope is that children will move from one score to another (change) as intended. Item codes were re-written over several versions to better achieve this distribution. As changes were made to the BOSCC while it was under development, videos were re-coded to reflect these changes. Other publications used a preliminary version of the BOSCC (from February 2014; Fletcher-Watson et al., 2015; Kitzerow, Teufel, Wilker, & Freitag, 2015; Pijl et al., 2016). Over the course of development items with a poor distribution across codes or with high inter-item correlations ( $r \geq 0.7$ ) were modified or eliminated. All data in this study used an updated version of the BOSCC coding scheme (September 2015 version). A final BOSCC coding scheme consisting of 15 items was chosen and applied to play samples described above.

The 15 BOSCC items are coded on a 6-point scale from 0 (abnormality is not present) to 5 (abnormality is present and significantly impairs functioning). Eight items relate to social communication behaviors. One item related to play and three items related to restricted, repetitive behaviors/interests (RRBs) seen in ASD. Because many children with ASD do not show all the coding ranges for RRBs (Kim & Lord, 2010), we did not expect normal or uniform distributions for the three items related to these behaviors, namely *Sensory Interests*, *Hand/Finger Mannerisms*, and *Restricted/Repetitive Behaviors/Interests*. Three items quantify Other Abnormal Behaviors often seen in ASD (*Activity Level*, *Disruptive Behavior/Irritability*, *Anxious Behaviors*). However, these behaviors were rarely observed in this sample of children playing with their caregiver; thus, these items were not included in analyses. However, these items were retained in the BOSCC coding scheme because they may yield valuable information when determining whether the BOSCC observation is a valid representation of the child's behavior (e.g., high scores on these items may suggest that the BOSCC observation was not representative of the child's typical behavior).

Each BOSCC item is coded using a novel decision tree originally derived from the well-validated ADOS-2 definitions of ASD behaviors. These decision trees capture detailed information about specific behaviors, including, for example, information about a behavior's frequency and quality (see Figure 1 for example item). At each branch of the decision tree, the coder answers a question about the child's behavior before proceeding on to the next question or arriving at a code. For example, the *Directed Vocalizations* item first asks whether the child directs vocalizations to another person (branch 1), then asks whether this ever occurs beyond echoed or highly routinized speech (branch 2), how often these more flexible directed vocalizations occur (branch 3), in what pragmatic contexts these occur (branches 4 and 5), and in

how many activities (branch 6). The BOSCC is coded in two 5-minute segments of a 10-minute video (Segment A and Segment B), with codes averaged across the two segments. The initial coding process relied on viewing each video segment once and then coding. Over the course of development, this process was modified such that each video segment was watched and coded twice, with the second codes deemed final and used for analyses in this study. Observing and coding each segment twice resulted in greater accuracy in capturing behaviors, higher reliability amongst coders on individual items, and greater confidence in coding decisions. Coding a BOSCC video takes a trained coder about 30 minutes to complete.

*BOSCC Application.* For the purposes of assessing the initial psychometric properties and validity, the BOSCC coding scheme was applied to 10-minute videos of free-play interactions between a caregiver and a child, gathered over the course of the child's participation in an intervention trial. A parent was the play partner for most of the BOSCC observations (97%,  $n=171$ ), with 94% ( $n=160$ ) of these conducted with mothers. For the remaining videos, the interaction was gathered with the child and another caregiver (e.g., grandparent). Most observations were gathered in the clinic setting ( $n=147$ , 83%) while the remaining observations were gathered at the child's home. These play samples were determined to be adequate for applying the BOSCC coding scheme because they fulfilled sufficient criteria for recommended BOSCC observations including minimal structure, a variety of toys (such as cause and effect and pretend objects), and duplicates of toys (to promote interactive play). Caregivers were given minimal instruction and simply told to play "how you typically would" with the child.

Between one and eight videos were available per child. Two or more videos were available for a subset of children ( $n=50$ ) with an average of 5.9 months ( $SD=3.1$ ) from first to last video observation. Children were between the ages of 12 and 56 months at their first

observation ( $M=29$ ,  $SD=11$ ) and between the ages of 18 and 62 months at their last observation ( $M=35$ ,  $SD=11$ ).

Coders of data presented here were one psychologist, one psychiatrist, one clinical psychology graduate student, and several research assistants. All coders were blind to the child's treatment status as well as the treatment time point. Before coding independently, coders obtained inter-rater agreement standards deemed adequate. These standards consisted of no more than four items with more than one point disagreement *and* within four points across summed totals for all items. These criteria had to be met for segments A and B for three consecutive videos. Training involved review of the BOSCC coding scheme, practice watching and coding video observations, and participation in coding discussions with reliable coders. How quickly trainees reached these inter-rater agreement standards varied though most met standards after practice coding approximately 10 to 12 videos. Codes from coders that were "in training" (had not yet met the above inter-rater agreement standards) were never included in datasets. Most coders of the BOSCC used in this study (September 2015 version) had been involved in coding that used previous versions of the BOSCC coding scheme while it was under development; as such, these coders, though many were bachelor-level assistants with limited previous ASD experience, had exposure to the BOSCC measure over several months. In addition, coders began training on the BOSCC at different points in the study, each participating in coding and consensus discussions of videos. Codes were only used in this study from coders who had attained the inter-rater agreement standards described above.

To ensure that no coder was coding any item significantly differently than other coders, ongoing reliability checks of individual coders were conducted; overall, there were no significant coding discrepancies between coders except for one coder who consistently under-scored

behaviors in the RRB domain. As such, *Sensory Interests*, *Hand/Finger Mannerisms*, and *Restricted/Repetitive Behaviors/Interests* that were coded by this coder were re-coded by other coders. Using the final version of the BOSCC described in this paper (September 2015 version), Figure 2 depicts the averaged (across segment A and B) item distributions for the 12 BOSCC items (BOSCC Core).

A random sub-sample of videos (approximately every 6<sup>th</sup> video) was chosen for coding by multiple coders (ranging from 2 coders to 7 coders) to assess inter-rater reliability and to ensure that inter-rater agreement standards were retained over time (see below). During consensus meetings for these multiply-coded videos, coders determined consensus codes; validity data presented here uses the consensus codes (16%, 28 videos) when applicable (but these codes were not used for inter-rater reliability, see below).

*Additional Measures.* As part of participation in the intervention trials, children completed several assessments, including tests of cognitive functioning, adaptive functioning, and diagnostic instruments. These additional measures provided an opportunity to explore the convergent validity of the BOSCC. See Table 2 for a summary of measures included.

*Adaptive Functioning.* The Vineland Adaptive Behavior Scales (VABS; Sparrow, Cicchetti, & Balla, 2005) was completed with the caregiver(s) of a subset of children (n=31) at two or more time points. The VABS is a caregiver interview of adaptive functioning that provides standard scores in the domains of socialization, communication, daily living, and motor skills as well as an overall adaptive behavior composite standard score (ABC). See Table 1 for information about VABS Domain scores at the initial observation.

*Cognitive Functioning.* Children were administered either the Mullen Scales of Early Learning (MSEL; Mullen, 1995) or the Differential Abilities Scales (DAS; Elliot, 2007),

depending on the child's ability level. The MSEL (collected from 36 children at two or more time points) provides standard scores in the domains of expressive language, receptive language, visual reception (non-verbal problem-solving), and fine motor skills. The DAS provides standard scores in the domains of verbal and nonverbal cognition. However, ratio IQs were calculated in some children for both scales due to the inability to calculate norm-referenced standard scores in some children because the child's age exceeded standard cut-offs and/or their developmental levels were too low to be calculated using standard metrics (Bishop, Guthrie, Coffing, & Lord, 2011). Ratio IQs are commonly used as an estimate of verbal and nonverbal IQ in samples of children with ASD and have been found to correspond with standardized metrics of VIQ and NVIQ (Bishop, Guthrie, Coffing, & Lord, 2011; Richler, Bishop, Kleinke, & Lord, 2007). Consistent with the methodology used by other researchers (Bishop, Guthrie, Coffing, & Lord, 2011; Richler, Bishop, Kleinke, & Lord, 2007), ratio NVIQ was calculated by averaging the age equivalents from the Visual Reception and Fine Motor domains, then dividing by the child's chronological age and multiplied by 100. Similarly, for ratio VIQ, age equivalents from the Receptive Language and Expressive Language domains were averaged, then divided by the child's chronological age and multiplied by 100. None of the children received the DAS at more than one time point. As a result, only the participants with multiple MSEL scores were included in analyses addressing change in cognitive scores. See Table 1 for information about cognitive functioning at the first observation.

*ASD Symptoms.* The Autism Diagnostic Observation Schedule, 2<sup>nd</sup> Edition (ADOS-2; Lord et al., 2012a; Lord, Luyster, Gotham, & Guthrie, 2012) was administered to 55 children at one time point. The ADOS-2 is a standardized observation-based measure of ASD symptoms that is divided into five modules, one of which is chosen depending on the child's age and

language level. The Toddler module of the ADOS-2 is appropriate for use with children up to 31 months of age who are using no words to those using simple phrase speech. Module 1 of the ADOS-2 is appropriate for use with children 31 months and older who are using no words to simple phrase speech. Module 2 of the ADOS-2 is appropriate for children who are using flexible phrase speech. Modules 3 and 4 of the ADOS-2 are appropriate for children, adolescents and adults with fluent language abilities (e.g., multi-clausal utterances, references to events or objects not present). Most children received ADOS-2 Module 1 or the Toddler Module (85%, n=47), while the remaining children (n=8) received ADOS-2 Module 2. A subset of children (n=41) received an ADOS-2 at two or more time points, allowing for exploration of change over time.

The ADOS-2 obtains information to contribute to a diagnosis of ASD through direct observation by a clinician. All clinicians involved in administering the ADOS-2 established research reliability on the measure prior to administration. None of the clinicians involved in administering/scoring of the ADOS were involved in administering/coding of the BOSCC. The ADOS-2 yields a Calibrated Severity Score (CSS) for the algorithm total (CSS Overall) and domain severity scores in the areas of Social Affect (CSS SA) and Restricted and Repetitive Behavior (CSS RRB; Esler et al., 2015; Gotham, Pickles & Lord, 2009). These scores provide a cross-module comparison that accounts for language level and age. See Table 1 for information about the ADOS-2 CSS at the first observation.

*Clinical Global Impression-Improvement (CGI-I)*. The CGI-I is a measure used by clinicians to evaluate whether an individual is responding to treatment (Busner & Targum, 2007). Clinicians rate the participant's level of improvement on a 7-point scale ranging from "very much improved" (1) to "very much worse" (7). In Study 1, a CGI-I was collected on six



children who participated in an intervention trial at CADB, for whom we also had BOSCCs at multiple time points (see below for more about CGI in Study 2). None of the clinicians who rated the CGI-I also coded the BOSCC.

#### Data Analysis.

*Exploratory Factor Analyses (EFA) of BOSCC Items.* To determine domain scores, Exploratory Factor Analyses (EFA) were conducted for the 12 Core BOSCC items (*Other Abnormal Behavior Items: Activity Level, Disruptive Behavior/Irritability, and Anxious Behaviors* were not included due to minimal endorsement). For the EFA, scores for the RRB items that had skewed distributions (*Sensory Interests, Hand/Finger Mannerisms, and Restricted/Repetitive Behaviors/Interests*) were collapsed into 3 or 4 categories based on the item distribution and treated as ordinal scores. EFA was conducted using all available codings, which includes multiple codings from different coders of the same video (308 total available codings). Analyses were undertaken in Mplus (Muthen & Muthen, 1998-2012) using a promax oblique rotation, taking into account the multiple codings (from multiple raters) and observations (of children over time; between 1 and 8 videos per child) by using complex survey adjustment with the child as the cluster-level unit. The complex survey adjustment method takes into account the non-independence of observations from the same child in the calculation of inferential statistics by using sampling weights in the estimation of the parameters. Subsequently calculated sums of raw scores formed over items were all appropriately normal and suitable for analysis by means of normal theory models. EFA of the 12 Core items, of which the last three items were treated categorically, gave eigenvalues of 5.48, 1.58, and 1.05 and RMSEA values of 0.107, 0.067, and 0.037 for the one, two, and three factor solutions, respectively (see Tables 3 and 4).

For subsequent analyses, the two-factor solution was chosen as a plausible parsimonious fit for the data because eigenvalues were substantially greater than 1, the RMSEA value was under 0.07 (Browne & Cudeck, 1993), and the model shared theoretical overlaps with other two-factor solutions found in ASD literature (Guthrie, Swineford, Wetherby, & Lord, 2013; Mandy, Charman, & Skuse, 2012; Shuster, Perry, Bebko, & Toplak, 2014). Factor 1, the Social Communication domain, consisted of items 1-8 (SC domain). Although some studies suggest that play is a separate factor (Boomsma et al., 2008; van Lang et al., 2006), the BOSCC *Play* item cross-loaded both on factor 1 and 2 and was placed in the RRB domain (items 9-12) for subsequent analyses due to item content that most closely related to play with materials instead of social aspects of play. The two domains, Social Communication (SC; items 1-8) and Restricted, Repetitive Behaviors (RRB; items 9-12) will be referred to in subsequent analyses as well as the Core total (items 1-12; see Figure 3). As described above, the three items related to Other Abnormal Behaviors were not included due to the rare presentation of these behaviors in this sample of children.

*Inter-Rater Reliability.* Sums for items in the factors (domains) defined by the EFA results (see above) were calculated as well as a sum for Core items (1-12, Core total). As described above, approximately every 6<sup>th</sup> video (28 videos) was coded by multiple coders. These double codings were used to obtain estimates for inter-rater reliability by randomly selecting two coders when more than two coders (up to 7 coders) coded a video. Consensus codes (mutually agreed upon codes for multiply-coded videos) for these 28 (16%) videos were not used for inter-rater reliability. Rather, original codes from two randomly selected coders were used for this purpose. Intraclass Correlation Coefficients (ICCs) for inter-rater reliability on domains (SC and RRB defined from the EFA) and Core total (items 1-12) were calculated. For the three skewed

items (*Sensory Interests, Hand/Finger Mannerisms, and Restricted/Repetitive Behaviors/Interests*) these results should be interpreted cautiously. ICCs for individual items (averaged from segment A and B) were also calculated.

*Test-Retest Reliability.* For estimation of test-retest reliability, a test-retest sub-sample of 20 individuals that had BOSCC videos available on two occasions less than one-month apart were randomly assigned to available coders (40 videos in total). ICCs on domains (defined from the EFA) and the Core total were estimated and ICCs were also estimated for individual items (summed from segment A and B).

*Validity.* To assess the validity of the BOSCC as a measure of relevant change, first, paired t-tests with  $\alpha = .05$  and effect sizes of changes (Cohen's D; the mean difference between first and last observation divided by the pooled standard deviation) were used to examine whether significant amounts of change in BOSCC and ADOS-2 CSS scores were present from the first to last observation.

Second, individual change models were fitted to all the available data on each child for the BOSCC Core total (items 1-12) and the ADOS-2 CSS (treated as a 10-point ordinal scale). Individuals in this data varied in terms of the time between observations, the number of observations available, and the time points at which observations were gathered, making the results sensitive to model specification errors. To limit this, these factors were taken into account by applying growth curve models. Specifically, the participant was used as the primary unit with repeated recordings treated as repeated measures. For each participant in turn, a linear regression was fitted to the data and the coefficient associated with the age at each assessment was used as the average rate of change score for that participant. This method allowed for the inclusion of a varied number of observations on the same individual at various time points (see Table 2) and

limiting model specification errors. To assist comparison for each observation we standardized the expected change over 6 months by the standard deviation at baseline. This can be thought of as the effect size (Cohen's D) that would have been obtained using each measurement had all of the children been followed for 6 months from baseline and compared to a randomized control group. A comparison of the effect sizes for the ADOS-2 CSS and BOSCC was constructed using the mean difference between the experimental and control group divided by the standard deviation of the control group. These quantities were obtained from a multivariate regression. These analyses were also conducted on the BOSCC SC (Items 1-8) domain separately.

Third, correlations of cross-sectional and change scores across these measures were conducted to assess convergent validity. Fourth, discriminant validity and coding contamination from maternal education and family income was tested by examining their association with BOSCC scores when included as fixed predictors within a mixed effects model for the repeated BOSCCs.

*Post-Hoc Analyses.* Given the phenotypic heterogeneity of ASD, it was expected that not all children would respond to treatment (Rogers & Vismara, 2008; Spence & Thurm, 2010). Therefore, responders and non-responders were identified based on changes from first to last observation based on other standardized measures of skills that were used as outcome measures in previous studies (MSEL, VABS, ADOS-2; Dawson et al., 2010; Wetherby et al., 2014). First, responders were defined based on changes in MSEL Receptive Language and, second, based on VABS Communication Standard Scores, consistent with changes observed in the two most recent major early intervention trials (Dawson et al., 2010; Wetherby et al., 2014). Specifically, children who demonstrated an increase in MSEL Receptive Language standard score of  $>5$  points ( $1/2$  standard deviation) were defined as responders ( $n=15$ ) while the remaining children

were defined as non-responders (n=21). Using the VABS standard Communication score, children were defined as responders if they demonstrated an increase of  $\geq 8$  points (1/2 standard deviation; n=16), while the remaining children were defined as non-responders (n=15). Third, children were defined as responders if they demonstrated an ADOS-2 CSS score decrease of  $\geq 1$  point (1 standard deviation; n=16), while the remaining children were defined as non-responders (n=25). Convergent validity was assessed using t-tests comparing the amount of change in BOSCC SC and RRB domains and Core Totals between responder and non-responder groups as defined by the above definitions on these measures.

Finally, to explore whether decreases in the BOSCC domain scores align with clinician's impressions of improvement, BOSCC scores for six other children participating in an early intervention trial at CADB (DeGeorge, Dufek, & Lord, in prep) were separated into responders and non-responders based on CGI scores. Four children received CGI scores of "much improved" (responders) while two children received CGI scores of "no change" (non-responders). No statistical analyses were conducted on these six children due to small sample size.

## Results

*Inter-Rater Reliability.* The estimated inter-rater reliability from the 28 videos that were coded by multiple coders (two randomly selected coders) was excellent for SC and RRB domains, as well as for the Core Total, with ICCs ranging from 0.94, 95% CI [0.87-0.97], to 0.96, 95% CI [0.93-0.99] (See Table 5). ICCs of individual items (averaged across segment A and B) ranged from 0.71, 95% CI [0.37-0.86] for the *Play* item to 0.93, 95% CI [0.86-0.97] for the *Unusual Sensory Interests* item (see Table 6). Given that coders had to be within 1 point on 80% of items for three consecutive videos prior to independent coding, the percent agreement

within 1 point for each item is also displayed in Table 6. The percent agreement within 1 point across items ranged from 68% for Segment B *Engagement* to 96% for Segment A *Vocalizations* and Segment B *Eye Contact*.

*Test-Retest Reliability.* Using a sub-set of children ( $n=20$ ) with two video observations separated by less than one month (40 videos total), the estimated test-retest reliabilities (ICCs) were high: 0.87, 95% CI [0.68, 0.95], for the Social-Communication domain, 0.78, 95% CI [0.47, 0.91], for the RRB domain, 0.90, 95% CI [0.75, 0.96], for the Core Total, and 0.91, 95% CI [0.78, 0.96], for the Overall Total (See Table 7). ICCs of individual items (averaged across segment A and B) ranged from 0.36, 95% CI [-0.56, 0.74] for the *Repetitive/Stereotyped Interests/Behaviors* item to 0.89, 95% CI [0.73, 0.96] for the *Vocalizations* item (see Table 8).

*Validity.* First, results of paired t-tests indicated that from first to last BOSCC observation ( $n=50$ ), statistically significant changes were found in the Core Total ( $M=-2.53$ ,  $SD=8.01$ ), [ $t(49)=2.23$ ,  $p<0.05$ ], corresponding to an effect size of 0.26. Changes in the separate SC and RRB domains were not statistically significant. Twenty-three (46%) of children showed decreases of four ( $1/2$  SD) or more points over time. Paired t-tests from first to last ADOS-2 observation ( $n=41$ ) indicated that there were no statistically significant changes in ADOS-2 CSS ( $M=-0.29$ ,  $SD=1.75$ ,  $d=0.15$ ), ADOS-2 SA CSS ( $M=-0.42$ ,  $SD=1.91$ ,  $d=0.21$ ), or ADOS-2 RRB CSS ( $M=0.42$ ,  $SD=1.84$ ,  $d=0.20$ ). See Table 1 for time between first and last ADOS and BOSCC observations.

Second, results from individual growth curve models indicated that the average rate of change in the ADOS-2 CSS score over an estimated 6 months was 0.33, which corresponded to an effect size of -0.15, 95% CI [-0.44, 0.15]. The average rate of change in the BOSCC Core Total over an estimated 6 months was -4.2, corresponding to an effect size of -0.37, 95% CI [-

0.73, -0.01]. Corresponding values for the BOSCC SC domain score were -3.4 with an effect size of -0.38, 95% CI [-0.81, 0.05]. Though the effect sizes were larger for the BOSCC, a comparison of the difference in effect sizes between changes in BOSCC Core Total and changes in ADOS-2 CSS indicated no statistically significant difference (multivariable regression;  $p=0.35$ ). However, the effect size of the BOSCC Core Total was statistically different from a no change alternative ( $p<0.05$ ) while the effect sizes of the ADOS-2 CSS and BOSCC SC domain were not statistically different from a no change alternative ( $p=0.33$  and  $p=0.08$ , respectively).

Third, in cross-sectional correlations, the BOSCC Core Total and the ADOS-2 CSS score were strongly associated (Pearson correlation of 0.48,  $p<0.001$ ). When correlating change scores to assess convergent validity, the MSEL Receptive Language and VABS Communication Standard scores showed highly correlated change scores ( $r=0.69$ ,  $p<0.001$ ). For the ADOS-2 CSS change score, evidence for convergent validity with the MSEL Receptive Language and the VABS Communication Standard score was neither significant nor consistent, while for the BOSCC Core Total, correlations were in the expected direction and, in the case of the MSEL Receptive Language, approached significance ( $r=-0.35$ ,  $p=0.05$ ). The correlation of ADOS-2 CSS to change in ADOS-2 CSS was 0.28 ( $p=0.08$ ) and of the BOSCC Core Total to change in BOSCC Core Total was -0.37 ( $p=0.08$ ).

Fourth, results of discriminant validity analyses indicated no associations of maternal education or family income with the BOSCC Social Communication domain ( $\chi^2(2)=1.94$ ,  $p=0.38$ ), RRB domain ( $\chi^2(2)=1.75$ ,  $p=0.42$ ) or the BOSCC Core Total ( $\chi^2(2)=1.53$ ,  $p=0.47$ ). There was also no association of maternal education and family income with the ADOS-2 CSS ( $\chi^2(2)=3.40$ ,  $p=0.18$ ).

*Post-Hoc Analyses.* T-tests comparing the amount of change in BOSCC scores between groups indicated that the MSEL responder group demonstrated significantly more change in the BOSCC SC domain ( $t(34)=3.04, p<0.01$ ) and the BOSCC Core Total ( $t(34)=3.58, p<0.01$ ) than the MSEL non-responders group (See Figure 4). Results of t-tests also indicated that the VABS responder group demonstrated significantly more change in the BOSCC RRB domain ( $t(29)=2.51, p<0.05$ ) and the BOSCC Core Total ( $t(29)=2.40, p<0.05$ ) than the VABS non-responder group. In contrast, BOSCC domains and the BOSCC Core Total did not statistically differ in the ADOS-2 CSS responder and non-responder groups (non-significant results).

As shown in Figure 4, except for the BOSCC RRB domain, from first to last time point, BOSCC scores for the CGI-I responders consistently decreased more than the CGI-I non-responders. Figure 5 shows the amounts of change in BOSCC domains for CGI-I responders and non-responders. Figure 5 is provided for illustrative purposes only because no statistical analyses were conducted on these groups given the small sample size.

## STUDY 2: APPLICATION OF BOSCC TO AN INDEPENDENT SAMPLE

### Method

*Participants.* Seventy-eight children (63 males) with a BEC (Anderson, Liang, & Lord, 2014) of ASD, confirmed through administration of the ADOS-2 (Lord et al., 2012a; Lord, Luyster, Gotham, & Guthrie, 2012), were included in this study. Children were between 4 and 8 years old and were participating in an on-going multi-site, school-based sequential multiple assignment randomized trial (SMART; 1R01HD073975, lead PI=Dr. Connie Kasari, University of California at Los Angeles: UCLA; Almirall, Compton, Rynn, Walkup, & Murphy, 2012; Murphy, 2005) aimed at increasing spontaneous speech in minimally-verbal children with ASD. For inclusion in the study, all children used fewer than 20 spontaneous words in a 20-minute



language sample gathered at screening. Children were from four sites: UCLA (n=11), CADB (n=29), University of Rochester (n=10), and Vanderbilt University (n=28). There was not a statistically significant difference between sites with regards to gender [ $\chi^2(3, N = 78) = 1.90, p = 0.59$ ], race [ $\chi^2(9, N = 72) = 10.04, p = 0.35$ ], ethnicity [ $\chi^2(3, N = 72) = 7.70, p = 0.05$ ], age at entry [ $F(3, 74) = 2.63, p = 0.06$ ] and IQ at entry [Leiter;  $F(3, 63) = 1.12, p = 0.35$ ]. See Table 9.

Children in this intervention trial were randomly assigned at entry to receive either Discrete Trial Training (DTT) or Joint Attention Spontaneous Play Enhanced Milieu Training (JASP-EMT). DTT is an Applied Behavior Analysis (ABA) approach that teaches children skills using small units of instruction (5-10 seconds) within a one-on-one, distraction-free environment (Smith, 2001). DTT is structured such that the child is provided a cue, the instructor prompts for a response (either modelling the response to the child or aiding the child in producing the correct response), the teacher then reinforces the child's correct response with a reward (e.g., praise, favorite toy, activity, or food), and then the instructor pauses a few seconds before cueing the next trial. JASP-EMT is an integration of two treatments aimed at increasing communication skills in children with ASD: Joint Attention Spontaneous Play Engagement and Regulation (JASPER) and Enhanced Milieu Teaching (EMT; Kasari et al., 2014). JASPER and EMT are interventions that embed ABA teaching strategies into naturalistic interactions. JASPER focuses on promoting the use of prelinguistic communication (e.g., joint attention) and the development of play skills during an interaction with an adult. While focusing on both behavioral strategies and social interaction, EMT uses environmental arrangement, responsive interaction techniques, and milieu teaching procedures such as time delays (e.g., a prompting strategy in which the child is not provided with a reward or desired object until he/she has provided the targeted behavior) and incidental teaching (e.g., unintentional teaching that occurs during natural activities), to

encourage use of language during social interactions (Kaiser, Hancock, Nietfeld, 2000).

Integrating JASPER and EMT (JASP-EMT), instructors implement a flexible, naturalistic, play-based intervention aimed at increasing spontaneous communication.

Utilizing the SMART design protocol, each child is randomly assigned again at midpoint to continue in the current treatment condition (DTT or JASP-EMT), receive the current treatment condition as well as parent training, or receive a blend of DTT and JASP-EMT, depending on the child's progress which is determined by the treating clinician in collaboration with study PIs. It should be noted that the focus of this work is on the validity of the BOSCC; therefore, analyses will focus on the ability of the BOSCC to assess change in this sample rather than overall effectiveness of different interventions. This study was approved by the Institutional Review Board (IRB) at each site (see Appendix 1 for IRB approval documentation from CADB and Teachers College).

*Primary Measure.* The 15-item BOSCC coding scheme (see Study 1) was applied to Caregiver-Child-Interaction (CCX) videos gathered at entry (pre-treatment), midpoint (after 6 weeks of treatment), and exit (after 16 weeks of treatment). Children were between the ages of 4.5 and 8.6 years ( $M=6.2$  years,  $SD=1.3$  years) at entry, 4.7 and 8.8 years ( $M=6.3$  years,  $SD=1.3$  years) at midpoint, and 4.9 years and 8.9 years ( $M=6.6$  years,  $SD=1.3$  years) at exit. CCX videos are 15-minute parent-child play interactions composed of three five-minute segments (Free Play, Hidden Objects, and Drawing/Coloring) during which the parents are instructed to play however they typically would with their child (e.g., showing the child a toy, labeling the toy, etc.). During the Free-play segment, parents are provided with a variety of materials to play with (e.g., pegs and pegboard, play-doh). For the Hidden Objects segment, parents are provided with a bag filled with four different toys (a book, a vehicle and track/road, puppets, and various blocks). For the

Drawing segment, parents were provided with coloring sheets (e.g. pictures of Thomas the Tank Engine), markers, crayons, and textured rubbing plates. For the purposes of this work, the BOSCC coding scheme was applied to two of the three five-minute segments: Free-play and Hidden Objects. These segments were chosen as they most closely align with the recommended elements of BOSCC observation contexts, including minimal structure and a variety of toys.

Two children from UCLA were missing CCXs at midpoint. Accordingly, a total of 232 CCX videos were coded using the BOSCC coding scheme. For 41 (18%) of the CCX videos, Free Play and/or Hidden Objects was less than 5 minutes long. In these instances, a portion of the video (mean= 34 seconds, SD=24) from the Drawing segment was used to create a complete 5-minute video. The child was usually interacting with his/her mother (86%, n=200) and most videos were gathered in the child's home (59%, n=138) or the clinic (36%, n=84).

One clinical psychology graduate student and two research assistants completed BOSCC coding. All coders were blind to the child's treatment status and treatment time point. All coders obtained inter-rater agreement standards described above in Study 1. Coding procedures were conducted as described above in Study 1. A random sub-sample of videos (n=22; approximately every 10<sup>th</sup> video) was coded by two randomly determined coders to ensure that inter-rater agreement standards were retained over time and to assess inter-rater agreement. Over the course of coding, five videos were coded by all coders and a final set of consensus codes was agreed upon for these videos; if available, consensus codes were used except when assessing inter-rater reliability. BOSCC domain scores, as described in Study 1 (Social Communication, RRB, Core, and Overall), were calculated and used for analyses. Children were defined as BOSCC responders (decrease of  $\geq 1$  SD of BOSCC Core change from entry to exit; n=12) or non-responders (decrease of  $< 1$  SD of BOSCC Core change from entry to exit; n=66).

*Additional Measures.* As part of participation in the intervention trial, children completed several assessments, including measures of cognitive functioning, adaptive functioning, and ASD symptoms.

*Adaptive Functioning.* The VABS (Sparrow, Cicchetti, & Balla, 2005) was completed with the caregiver (n=70) at entry to the study. As described in Study 1, the VABS provides standard scores in the domains of socialization, communication, and daily living. The motor skills domain is only administered for children under age 7. Given the age range for this study (4-8 years old), the motor skills domain was not explored because it was only administered to a subset of children (n=41). A standard score could not be calculated for the daily living domain for one child due to the caregiver reporting “I don’t know” for more than 7 items.

*Cognitive Functioning.* Estimates of non-verbal cognitive ability were gathered through completion of the Leiter International Performance Scale (Leiter; Roid & Miller, 1997). Four subtests of the Leiter were administered to obtain age-normed Brief IQ scores at entry and exit. The Leiter is a standard measure of non-verbal cognitive abilities that is administered completely non-verbally and does not require the child to speak to be able to complete the tasks. Standard scores on the Leiter have a mean of 100 and standard deviation of 15. Four children were not administered the Leiter at the last time point (exit).

The Peabody Picture Vocabulary Test (PPVT-4; Dunn & Dunn, 2007) was gathered at entry, midpoint, and exit to provide an estimate of the child’s single-word receptive language abilities. The PPVT-4 is administered by asking the child to identify a word out of a four-picture set (e.g., “point to dog”). The PPVT-4 provides age-normed standard scores ( $M=100$ ,  $SD=15$ ). Two children were not administered the PPVT-4 at entry and midpoint.

*ASD Symptoms.* At entry to the study, the ADOS-2 (Lord et al., 2012a) was collected as part of the standard diagnostic battery and to measure ASD symptom severity (see Study 1). All children in this study received Module 1 of the ADOS-2, which is appropriate for children who are older than 31 months (chronologically) and are pre-verbal or using single words. See Table 9 for a summary of ADOS-2 domain totals and CSS at entry.

In addition, parents completed the Repetitive Behavior Scale-Revised (RBS-R; Lam & Aman, 2007), a 43-item questionnaire of RRB symptoms at entry and exit. Each item on the RBS-R rates RRB symptoms on a 4-point Likert scale ranging from 0, behavior does not occur, to 3, behavior occurs and is a severe problem. The RBS-R is comprised of six subscales that do not have content overlap. The six subscales measure stereotyped behavior, self-injurious behavior, compulsive behavior, routinized behavior, insistence on sameness, and restricted behavior. The RBS-R does not provide normed scores. Items from these subscales are summed to create an RBS-R total (Esbensen, Mailick Seltzer, Lam, & Bodfish, 2009). The RBS-R has shown high internal consistency and adequate psychometric properties (Bodfish, Symons, Parker, & Lewis, 2000; Lam & Aman, 2007).

*Clinical Global Impression, Severity and Improvement (CGI-S and CGI-I).* The CGI is a measure used by clinicians to evaluate whether an individual is responding to treatment (Busner & Targum, 2007). Using standard procedures developed for the intervention trial (e.g., to capture increases in spontaneous communication), interventionists and a review committee from across sites rated the child's severity of social communication impairment on a 7-point scale (CGI-S) at entry, midpoint, and exit. In addition, interventionists and the committee also rated the child's level of improvement (CGI-I), on a 7-point scale ranging from "very much improved" (1) to "very much worse" (7), at midpoint and exit. Children were defined as responders ("CGI-I

responders”) to treatment if they received a CGI-I score of 1 (“very much improved”) or 2 (“much improved”; n=29). Analyses were also conducted when defining children as CGI-I responders if they received a CGI-I score of 1 or “very much improved” (n=22). Children were also defined as CGI-S responders (“CGI-S responders”) to treatment if they decreased 1 point or more on the CGI-S from entry to exit. Five children were missing CGI-I and CGI-S information at exit.

### Data Analysis

Preliminary data explorations were conducted to explore item distributions (see Figure 6) and inter-rater reliability (by calculating ICCs; See Study 1) for BOSCC domains and items. Data analyses of BOSCC change scores over time focused on total change from entry to exit (over the whole course of treatment). To address our first aim, paired t-tests were conducted and effect sizes (Cohen’s D; the mean difference between first and last observation divided by the pooled standard deviation; see Study 1) were calculated to explore whether significant changes in BOSCC scores or other measures occurred from entry to exit. To address our second aim,  $\chi^2$  analyses were conducted to compare the number of BOSCC Core “responders” to the number of children defined as CGI-I and CGI-S responders. A t-test was conducted to compare change in BOSCC scores between CGI-I Responders and Non-Responders. In addition, a series of logistic regressions were conducted to explore whether VABS domains, Leiter IQ, or ADOS-2 CSS at entry predicted status as either a CGI-I or BOSCC Core responder. To address our third aim, consistent with post-hoc analyses in Study 1, children were defined as responders or non-responders based on other standard measures (Leiter, PPVT-4, and RBS-R). Specifically, children were defined as Responders on the Leiter, PPVT-4, or the RBS-R if they changed  $\frac{1}{2}$  SD from entry to exit on each respective measure. The remaining children were defined as non-

responders on these measures. Then, t-tests were conducted to compare changes in BOSCC domain scores in responders to non-responders based on these measures. To address our final aim, correlations were conducted to assess whether nonverbal IQ (Leiter), adaptive functioning (VABS), or ASD symptom severity (ADOS-2 CSS) at entry was significantly correlated with BOSCC domain scores or the change in BOSCC domain scores from entry to exit. To account for multiple comparisons (Bonferroni method), a significance threshold was set at  $\alpha = 0.01$  across analyses.

## Results

*Inter-Rater Reliability.* To assess the inter-rater reliability of the BOSCC, ICCs were calculated from the 22 videos coded by multiple coders (two randomly selected coders). ICCs were high across domains, ranging from 0.70, 95% CI [0.41,0.86] for the RRB domain to 0.87, 95% CI [0.70,0.94] for the SC domain (See Table 10). ICCs of individual items (sums across segment A and B) ranged from 0.46, 95% CI [0.06, 0.74] for the *Social Responses* item to 0.89, 95% CI [0.75,0.95] for *Vocalizations* item (See Table 11). Given that coders had to be within 1 point on 80% of items for three consecutive videos prior to independent coding, the percent agreement within 1 point for each item is also displayed in Table 11. The percent agreement within 1 point across items ranged from 64% for Segment A *Restricted/Repetitive Interests/Behaviors* to 100% Segment A *Facial Expressions* and Segment B *Play*.

*Change Over Time Across Measures.* To assess the construct and convergent validity of the BOSCC, change in BOSCC and other measures from entry to exit was examined. Table 12 presents mean values across time points and results from paired t-tests examining the amount of change in BOSCC domains, the Leiter, PPVT-4, and RBS-R from entry to exit. Results indicate that no measure demonstrated significant amounts of change from entry to exit, with

correspondingly small effect sizes (Cohen's *d*) ranging from 0.003 for the PPVT-4 to 0.17 for the RBS-R. The direction of change, though not significant, was in the expected direction for the RBS-R, though not for the other measures. On the BOSCC, 12 children (15%) demonstrated decreases of 6 or more points (1 SD) in the Core Domain from entry to exit (BOSCC Core Responders); the remaining 66 children were considered BOSCC Core Non-Responders. While most of the 12 BOSCC Core Responders (n=10; 83%) showed some decrease from entry to midpoint (ranging from 1.5 points to 12 points), only four of these children demonstrated a decrease of 6 or more points (1 SD) in the BOSCC Core Domain from entry to midpoint.

*BOSCC & CGI Responders.* Given that we would not necessarily expect all children to demonstrate improvements over time and to assess convergent validity between the BOSCC and a clinician's impression of change, we explored the proportion of children who were defined as responders on the CGI-I and CGI-S compared to those who showed significant amounts of change on the BOSCC. Three of twelve children defined as BOSCC Core Responders were missing CGI-I data at exit. Of the nine remaining children, six (67%) were defined as CGI-I Responders ("much improved" or "very much improved") and seven (one additional child; 78%) were defined as CGI-S Responders (decrease of at least one point). Results of a  $\chi^2$  analyses indicated that there was not a statistically significant difference between the proportion of children defined as CGI-I or CGI-S responders/non-responders and BOSCC Core responders/non-responders,  $\chi^2 (1, N = 73) = 0.05, p = 0.82$  and  $\chi^2 (1, N = 73) = 0.31, p = 0.58$ , respectively. See Table 13.

Overall, clinicians thought far more children responded than those who demonstrated change on the BOSCC. In addition, there were a few children (n=3) who demonstrated significant amounts of change on the BOSCC who were not identified as responders by the CGI-



I. A t-test indicated that there was no statistically significant difference between CGI-I Responders and Non-Responders in change in any BOSCC domain from entry to exit (See Table 14). While there was no statistical difference in BOSCC change scores between CGI-I Responders and CGI-I Non-Responders, the BOSCC change scores for CGI-I Non-Responders appeared to be increasing slightly over time (ranging from increase of 0.52 to 1.89) while BOSCC change scores for CGI-I responders are stable over time (ranging from decrease of 0.01 to an increase of 0.07). In addition, there was not a statistically significant difference between CGI-I responders and CGI-I non-responders in Leiter, PPVT-4, and RBS-R change scores (see Table 14).

Last, a series of logistic regressions were conducted to assess whether VABS domains, Leiter IQ, and ADOS-2 CSS at entry predicted CGI-I and BOSCC Core responder status. Results indicated that the VABS domains, Leiter IQ, and ADOS-CSS at entry did not predict CGI-I or BOSCC Core responder status (See Table 15). Results were consistent when defining children as CGI-I responders if they received a score of 1 (“very much improved”; n=22).

*BOSCC and Leiter, PPVT-4, and RBS-R Responders.* As stated above, since we would not necessarily expect all children to demonstrate improvements over time and to assess convergent validity between the BOSCC and responder status based on other standard measures, we identified children that were showing change on the Leiter, PPVT-4 and RBS-R. Twelve (22%) children were identified as “responders” on the Leiter (showing increases of at least  $\frac{1}{2}$  SD or 8 points). Fourteen (22%) children were identified as “responders” on the PPVT-4 (showing increases of at least  $\frac{1}{2}$  SD or 8 points). Twenty-three (31%) children were identified as “responders” on the RBS-R (showing decreases of at least  $\frac{1}{2}$  SD or 8 points). T-tests comparing the amount of change in BOSCC scores between the groups identified as responders and non-

responders on the Leiter, PPVT-4 and RBS-R indicated that there were no statistically significant differences between the groups in the amount of change in BOSCC domains from entry to exit.

See Table 16.

*Entry Measures and Correlations with BOSCC.* To assess whether symptom presentations at entry were related to BOSCC scores at either time point or amount of change in BOSCC, we conducted correlations between the BOSCC and other measures. Results indicated several statistically significant correlations between entry measures and entry BOSCC totals. Specifically, ADOS-2 CSS at entry and BOSCC SC domain total at entry were significantly correlated ( $r=0.30$ ), such that children with more severe ASD symptoms (higher ADOS CSS) showed more impaired social communication skills on the BOSCC (higher BOSCC SC). These results indicate convergent validity between the BOSCC and ADOS-2 supporting the utility of the BOSCC as a possible metric of social communication severity. A significant correlation was also found between VABS Daily Living Skills and BOSCC SC, BOSCC RRB, BOSCC Core, and BOSCC Overall at entry ( $r=-0.33$ ,  $r=-0.35$ ,  $r=-0.43$ , and  $r=-0.47$ , respectively), such that children with more impaired daily living skills (lower VABS Daily Living Skills) showed more impairment across BOSCC domains (higher BOSCC SC, RRB, Core, and Overall totals) at entry. The VABS Communication was also significantly correlated with BOSCC Core and Overall Totals at Entry ( $r=-0.31$  and  $r=-0.37$ , respectively), such that children with more impaired communication abilities (lower VABS Communication) showed more impairment during the BOSCC (higher BOSCC Core and Overall totals) at entry. These results indicate convergent validity between the BOSCC scores and other aspects of current functioning. The Leiter Brief IQ was significantly correlated with BOSCC RRB total at entry ( $r=-0.36$ ), such that children with lower Leiter Brief IQ showed more repetitive behaviors during the BOSCC.

No statistically significant correlations were found between entry measures and exit BOSCC totals or between entry measures and change in BOSCC totals from entry to exit, which is not surprising given the minimal amounts of change observed over time in this sample. In addition, age at entry was not significantly correlated with BOSCC domain scores at entry, exit, or change in BOSCC from entry to exit; this suggests that the BOSCC scores are relatively independent of age. See Table 17.

## DISCUSSION

These two studies provide details on the preliminary utility and limitations of the BOSCC. Overall, the results of these analyses suggest that the BOSCC is a promising outcome measure that may be sensitive to subtle changes in social communication behaviors over time. To our knowledge, the BOSCC is the first brief, observation-based measure of dimensional treatment response that quantifies a broad range of social communication behaviors and can be administered and coded by an inexperienced researcher who is blind to the child's treatment status and time point. The over-arching aim of these two studies was to assess the ability of the BOSCC to quantify changes over time and to assess the psychometric properties of the BOSCC including inter-rater reliability, test-retest reliability, and the factor structure of the measure.

### *Does the BOSCC Capture Change Over Time?*

To assess change in the BOSCC over time, we explored pre- to post-treatment changes in BOSCC scores and compared changes in the BOSCC scores to changes in other standard measures, such as the ADOS-2 CSS, MSEL, VABS, PPVT, Leiter, and RBS-R. Study 1 indicated that the BOSCC scores demonstrated small to medium effect sizes of change in toddlers with ASD after an average of 6 months of early intervention. In contrast to Study 1, Study 2 revealed that after four months of an intervention aimed at increasing spontaneous

communication and engagement in an older more intellectually and language impaired sample, the BOSCC scores did not change significantly over time. The mixed results of these studies indicate that the BOSCC's ability to quantify change has limitations.

When comparing change in the BOSCC scores to change in the ADOS-2 CSS, results from Study 1 showed that effect sizes of change in BOSCC scores were not statistically different than the effect sizes of change on the ADOS-2 CSS. Nevertheless, the effect size of change in BOSCC scores, considering the small sample size, is promising. In addition, the BOSCC scores demonstrated statistically significant changes over time while the ADOS-2 CSS scores did not (when compared to a no change alternative). This suggests that the BOSCC may be more sensitive to changes in social communication behavior than the ADOS-2 CSS, and hence more successful in identifying changes in response to treatments over shorter periods of time (Brian, Smith, Zwaigenbaum, Roberts, & Bryson, 2015; Dawson et al., 2010; Shumway et al., 2012; Thurm, Manwaring, Swineford, & Farmer, 2015).

When comparing changes in the BOSCC scores to changes in other standard measures, post-hoc analyses from Study 1 indicated that when children were defined as either responders or non-responders based on the VABS or MSEL, BOSCC scores decreased significantly more in responders than non-responders, suggesting convergence of these multiple assessments of change. Similarly, Study 1 revealed preliminary evidence of convergent validity with the clinician's impression of improvement (CGI-I) in a very small sample. Overall, results of Study 1 are the first indication that the BOSCC has convergent validity with social communication changes seen in other measures, including a caregiver report measure (VABS), a standardized cognitive measure (MSEL), and clinician impression (CGI-I). Despite these results, Study 1 found that there was not a significant correlation between change in the BOSCC scores and

change in the ADOS-2 CSS scores, possibly due to the minimal amounts of change or limited range of scores observed in the ADOS-2 CSS, consistent with other studies (Dawson et al., 2010).

In contrast, results of Study 2 indicated that none of the measures (BOSCC, PPVT-4, Leiter, and RBS-R) changed significantly over time. When children in Study 2 were defined as either responders or non-responders based on the PPVT-4, Leiter, or RBS-R, similar to the post-hoc analyses of Study 1, change in BOSCC scores did not differ between groups. Study 2 also provided an opportunity to further explore the relationship between the CGI-I and the BOSCC; results of Study 2 indicated that the CGI-I identified most children as responders to treatment, whereas the BOSCC did not. There was no difference between children who were defined by the clinicians as responders to treatment (CGI-I Responders) and those who were defined by the clinicians as non-responders to treatment (CGI-I Non-responders) in amount of change across all other measures (BOSCC, PPVT-4, Leiter, and RBS-R).

Overall, the results of Study 1 were encouraging. Study 2 yielded many interesting results but it did not replicate the results of Study 1 as was anticipated. There are several possible explanations for the minimal changes seen in the BOSCC scores in Study 2. None of the standard measures (Leiter, PPVT-4, and RBS-R) changed over time, suggesting that the lack of change in BOSCC scores seen in Study 2 may reflect the reality that these children are not changing significantly over the course of treatment. The only measure that identified children as improving in Study 2 was the treating clinician's impression of improvement (CGI). Clinicians identified 70% of children as "much improved" or "very much improved" after four months of treatment. In fact, clinicians identified 96% of children as at least "minimally improved." It is important to note that the CGI was based on a narrowly defined study goal: an increase in

spontaneous communication and/or engagement. In contrast, the BOSCC quantifies more general changes in social communication, beyond the speech and engagement goals of this intervention. Study 2 showed that there was no difference between CGI-I Responders and CGI-I Non-Responders in BOSCC, Leiter, PPVT-4, or RBS-R, indicating that the clinician's impression of change did not align with BOSCC or any other measure. Given the different behaviors that the CGI-I and the BOSCC aim to quantify, it is important to consider what changes the BOSCC may not capture. For example, while using the CGI-I clinicians may identify children whose ability to attend to play materials increases or whose disruptive behavior decreases over the course of treatment, the BOSCC may not quantify these other important behavioral changes. This may explain some of the difference between the CGI-I and BOSCC in Study 2.

It is worth noting that such a high proportion of responders as measured by the CGI seems inconsistent with the heterogeneous nature and variability of treatment response seen in ASD (Rogers & Vismara, 2008; Spence & Thurm, 2010). Nevertheless, if accurate, the high proportion of CGI responders suggests that children are demonstrating change within the therapeutic context at school, where they are seen by clinicians, but may not be generalizing these skills to the home context with parents, the context to which the BOSCC was applied. The CGI-I captures changes that are occurring with the treating clinician within the treatment session at school. In contrast, the BOSCC coding scheme was applied to parent-child play sessions usually at home. It may be that children are showing change within the treatment session that is not generalizing to the broader home or parental context. Future work should explore the ability of children to generalize treatment gains to different contexts as well as the ability of the BOSCC to capture improvements across various contexts.

This sample's unique characteristics may be another contributing factor to the lack of change seen in the BOSCC in Study 2. This sample is composed of children who were about four years older and who were more language and intellectually impaired than the children in Study 1. These children are often considered slower to respond and less likely to respond to treatment (Koegel, Shiratova, & Koegel, 2009; Tager-Flusberg & Kasari, 2013). Changes in the BOSCC have been seen in other samples of young children or toddlers, like in Study 1 (Kitzerow, Teufel, Wilker, & Freitag, 2015; Pijl et al., 2016), but changes were not found in a slightly older sample of children (Fletcher-Watson et al., 2015), indicating that perhaps the BOSCC is most effective at quantifying change in younger children. It is also possible that the lack of change in the BOSCC in Study 2 is a result of the shorter period over which these children were followed (four months in Study 2 as compared to 6 months in Study 1). Other studies that have utilized the BOSCC have found change over the course of six months to one year (Kitzerow, Teufel, Wilker, & Freitag, 2015; Pijl et al., 2016) but the BOSC has been less successful at identifying change over the course of eight to ten weeks (Fletcher-Watson et al., 2015; Nordahl-Hansen, Fletcher-Watson, McConachie, Kaale, 2016). Future work should explore the utility of the BOSCC to identify meaningful improvements in this population of children and over short periods of time in order to better understand the limitations of the BOSCC.

#### *Reliability and Factor Structure of the BOSCC*

These studies also explored the inter-rater and test-retest reliability as well as the factor structure of the BOSCC. Analyses of the psychometric properties indicate that the BOSCC has excellent inter-rater reliability and high test-retest reliability, meeting recommended standards (Cunningham, 2012) and consistent with other work using an earlier version of the measure

(Kitzerow, Teufel, Wilker, & Freitag, 2015). Using the BOSCC domain totals is recommended, given the lower ICCs when using individual items.

Results of this work indicated that the items on the BOSCC comprise two-factors, supporting a Social Communication domain separate from RRBs, consistent with other models of ASD symptoms (Guthrie, Swineford, Wetherby, & Lord, 2013; Mandy, Charman, & Skuse, 2012; Shuster, Perry, Bebko, & Toplak, 2014). The separation of the two domains allows future researchers to explore changes in social communication skills in children with social-communication impairments who do not necessarily have RRBs or meet criteria for ASD (e.g. Social Communication Disorder, Social Anxiety Disorder).

When considering the importance of the two BOSCC domains, improvements in the BOSCC Core Total (items 1-12, combining Social Communication and RRB domains) most consistently converged with improvement in other standard measures (VABS, MSEL), while change in the separate BOSCC SC and RRB domains was less consistent. Although the separate SC and RRB domains may prove useful in non-ASD populations or when assessing change specific to one domain, this work suggests that the BOSCC Core Total may be the most appropriate domain to identify improvement in young, minimally verbal children with ASD (Study 1). This needs to be confirmed in future work with larger, independent samples.

Of note, only three items on the BOSCC attempt to capture RRB behaviors across a continuum. Item distributions, though better in Study 2 than Study 1, indicated that obtaining a continuum for these behaviors was challenging. It may be that these behaviors are either clearly present or not present at all (with little variation in between) or that subtle variations in these behaviors are difficult to capture within five minutes. Though still adequate, the RRB domain score demonstrated lower test-retest and inter-rater reliability than the SC domain, consistent



with earlier iterations of the BOSCC (Kitzerow, Teufel, Wilker, & Freitag, 2015) and the ADOS-2, from which initial drafts of these items were developed. As mentioned, the BOSCC Core Total (combining SC and RRB domains) was most successful in identifying change, indicating the importance of these behaviors in combination with the SC behaviors, at least in these ASD samples. Perhaps this is a result of the strong relationship between these domains in the ASD population (Richler, Huerta, Bishop, & Lord, 2010). The BOSCC RRB domain may not prove to be a useful domain in which to measure change on its own but additional studies are needed to thoroughly assess this. In the meantime, it may be helpful to use other measures of RRB behaviors in combination with the BOSCC, such as the RBS-R (Lam & Aman, 2007), which demonstrated more change than other measures in Study 2. Future studies should continue to examine the convergence between changes in the BOSCC RRB domain and other measures of RRBs to aid in understanding the utility of the BOSCC in measuring these symptoms over time.

#### *The Relationship Between the BOSCC and other Measures at a Single Time Point*

A secondary aim of these studies was to explore the relationship between BOSCC scores and scores on other measures at a single time point. In both studies, the BOSCC Core Total and the ADOS-2 CSS score were highly correlated with each other at one time point. These results suggest that the BOSCC may be a metric of ASD severity. Because the BOSCC is not normed or standardized for use as a population-based metric of ASD symptom severity and because coders are only reliable within a site, it is important to consider the BOSCC as a potential metric of severity that is specific to only the sample the BOSCC was applied to (study specific). Future work should continue to explore the utility of the BOSCC as a metric of within-sample severity, especially since this was not the intended goal of development.

Study 2 also revealed significant correlations between entry BOSCC scores and entry adaptive behaviors and intellectual ability (in the expected direction). This provides further evidence for the BOSCC as a broad measure of within-sample social communication severity that may be related to other aspects of functioning. For example, the BOSCC RRB total was negatively correlated with IQ in Study 2, indicating that children with more RRBs had lower IQs. This is consistent with other work suggesting that lower intellectual ability may be related to more or specific types of RRB behaviors (Bishop, Richler, & Lord, 2006; Cuccaro et al., 2003; Cannon et al., 2010; Silverman et al., 2008).

In addition, despite the significant correlation between the BOSCC and ADOS-2 CSS score in both studies, the BOSCC is not intended to be a measure of diagnostic classification. Rather the BOSCC was developed to capture nuanced social communication behaviors that may change over relatively brief periods of time. This distinction is important to prevent misuse of this new measure.

#### *Additional Benefits of the BOSCC*

The utility of the BOSCC is enhanced by its ability to be administered and scored by someone with minimal experience and who is unaware of the child's treatment status and time point. In our studies, post-baccalaureate research assistants reliably coded the BOSCC. The BOSCC does not require a highly experienced or credentialed coder, unlike other commonly used measures (Bolte & Diehl, 2013). Because the BOSCC assesses change in a single individual over time and these blind coders could be randomly assigned a video from any time point, high levels of agreement amongst coders in a coding team is particularly crucial. In contrast, agreement in BOSCC scores is less crucial across sites, unlike reliability training for the ADOS-2. The high inter-rater agreements in our group indicate that it is feasible to attain strong

concordance across a range of experience levels. However, one highly experienced coder in our group, a child psychiatry fellow, consistently under-scored behaviors in the RRB domain. This may suggest that the BOSCC is initially more challenging for someone who has more advanced training or experience, particularly in a specific framework, though this remains to be thoroughly explored.

Another goal when developing the BOSCC was to create a measure that did not rely on parent or clinician report. The BOSCC was developed to be coded by researchers who are unaware of the child's treatment status and time point, minimizing measurement bias associated with parent or clinician report (Anagnostou et al., 2015; Bolte & Diehl, 2013; Guastella et al., 2015). This is particularly important as evidence is growing that parent report measures may produce placebo-like results (Jones, Carberry, Hamo, & Lord, in press) that may even outweigh true treatment effects (Guastella et al., 2015).

The BOSCC's minimally structured, naturalistic context places little demand on administration and contributes to the measure's ecological validity. In addition, the use of video-based data has many benefits (Jewitt, 2012). Video data captures the richness and totality of behavior and eliminates the pressure of quick assessments as in live coding situations. In this way, coders can thoroughly consider and dissect behaviors. Video also captures aspects of the surrounding context, allowing coders to place behaviors in relation to the environment and sequence of events. Video data also provides opportunities for data reuse. For example, videos could be submitted securely online, allowing for a centralized cross-study coding process and for the application of automated analyses of visual information (e.g., physical orientation of the interactors' heads; distance between interactors; Dawson et al., 2004), acoustic information (e.g., features of conversational turn-taking or vocal prosody; Bone et al., 2012; Narayanan &

Georgiou, 2013), or other aspects of behavior. Despite the advantage of video coding, our group aims to also explore the utility of the BOSCC in live coding situations, as this method would not require video cameras or rely on audio/visual recording quality.

*Additional Considerations.*

It is also important to consider the limited endorsement in our samples of any of the Other Abnormal Behaviors (*Activity Level, Disruptive Behavior/Irritability, Anxious Behaviors*) within parent-child play-based contexts. Despite the infrequent display of these behaviors, we retained these items in the coding scheme because other researchers may want to consider them in future analyses. It is possible that these behaviors may impact social communication and RRB behaviors captured in other codes or may be more common in other contexts. These behaviors may also be useful in determining whether the BOSCC observation is a valid representation of the child's behavior. For example, if a child is particularly irritable, the observation may not represent the child's typical behavior. Ensuring that a BOSCC observation is an adequate representation of the child's behavior is an FDA requirement when deciding on the validity of an observation-based instrument (Chan, 2014).

An additional consideration when implementing the BOSCC is the context used to gather a valid sample of the child's behavior. Previous work has emphasized the importance of the context in which changes are assessed (Yoder et al., 2013), therefore whichever social and environmental context is chosen for the BOSCC observation, the context should be as consistent as possible (e.g., same play partner, same materials, same location) to ensure the validity of the observations gathered. At the same time, measures that go beyond a single context are clearly necessary to ensure generalization of skills gained in treatment. Future work should assess the

influence of different BOSCC contexts on the child's behavior and BOSCC scores to determine whether some contexts are more appropriate than others.

*Limitations.*

Although the initial results of the BOSCC are promising, especially when applied to toddlers across six months of treatment, they should be interpreted in light of several limitations of this project, including the relatively small sample sizes. These studies focus on samples of 56 and 78 children with ASD, with even smaller samples of children with multiple observations of other measures (e.g., CGI, VABS, MSEL, ADOS-2, Leiter, PPVT-4, and RBS-R). Our small samples did not allow for analyses of differences by sex, race, or ethnicity or to control for baseline observations. In addition, these studies did not explore the effects of specific treatment or control conditions. We hope to expand this work to larger samples comparing different treatment conditions, employing the BOSCC as an independent measure of treatment response.

While all children in Study 2 were using fewer than 20 spontaneous words (equivalent to Module 1 of the ADOS-2) and most children in Study 1 used simple phrase speech or less (equivalent to the Toddler Module or Module 1 of the ADOS-2), a subsample of eight children in Study 1 completed Module 2 of the ADOS-2, equivalent to flexible phrase speech. Given the small sample of children using flexible phrase speech, these studies did not explore whether the BOSCC coding scheme is as effective at capturing change in this more verbal group as it is in children with less speech. Similarly, given that the BOSCC was not as effective at capturing change in Study 2, as highlighted above, the BOSCC could be less effective at capturing change in an older, more intellectually impaired sample of children. Future work should address whether modifications to the BOSCC coding scheme are necessary to adequately capture changes in children using phrase speech and children who are older yet minimally-verbal.

When interpreting the results of these studies, changes in the BOSCC, as a representation of the child's behavior, must be considered in light of changes in the caregiver's behavior. Parent-child interactions are often described as bi-directional--the child's behaviors impact the parent and vice versa (Ginn, Clionsky, Eyberg, Warner-Metzger, & Abner, 2015; Rutgers, Bakermans-Kranenburg, van Ijzendoorn, & van Berckelaer-Onnes, 2004; Siller & Sigman, 2008; Slaughter & Ong, 2014; Zhou & Yi, 2014). A recent parent-focused intervention study found that changes in ASD symptoms, as measured by the ADOS-2 CSS, were mediated by parental synchrony (Pickles et al., 2015) and another study found a high correlation between the quality of the parent-child interaction and the child's ASD severity (using the ADOS-2 CSS; Hobson, Tarver, Beurkens, & Hobson, 2016). Research has also shown that children's language development may be influenced by a parent's responsiveness during play interactions (Siller & Sigman, 2008). Similarly, research has shown that a clinician's amount and prosody of speech is influenced by the child's speech during ADOS-2 administrations (Bone et al., 2014). Our studies did not assess whether the caregiver's behavior significantly impacted the child's BOSCC scores or if the child's severity of ASD or other behaviors impacted the caregiver's behavior. Given this potential confound, some researchers may choose to have an examiner who is blind to the child's treatment status interact with the child during the BOSCC. If the caregiver is chosen as a BOSCC play partner, researchers should consider collecting additional measures of caregiver behaviors that may contribute to observed changes in the child's behavior (Pickles et al., 2015). For example, measuring the parent's behavior using the Emotional Availability Scales (Biringen, Robinson, & Emde, 2000) or analyzing the amount and quality of the parent's speech production (Bone et al., 2014) may yield useful information related to the child's behavior. When using the parent as a play partner in BOSCC administrations, disentangling changes in the child's behavior

from changes in the parent's behavior will be essential to understanding the effectiveness of an intervention.

*Future Directions and Conclusions.*

Given the limited results of Study 2, our group is expanding this work by applying the BOSCC coding scheme to clinician-child treatment sessions gathered at the same time points as the parent-child play sessions used in Study 2. This would allow us to explore whether the child is improving within the context that intervention is delivered but perhaps not generalizing to the parent-child play context. Our group is also working on several lines of research related to the development of the BOSCC, including expanding the BOSCC to individuals with verbal fluency. We are also modifying the BOSCC for application to segments of ADOS-2 videos, consistent with the work of others (Kitzerow, Teufel, Wilker, & Freitag 2015). Given that researchers are familiar with the ADOS-2 and many have been collecting ADOS-2 videos at multiple treatment time points, we aim to confirm the validity of this method, allowing researchers to apply the BOSCC to pre- and post-treatment ADOS-2 videos from previously collected data.

In conclusion, our ongoing work and the work of other researchers (Fletcher-Watson et al., 2015; Kitzerow, Teufel, Wilker, & Freitag, 2015; Nordahl-Hansen, Fletcher-Watson, McConachie, & Kaale, 2016; Pijl et al., 2016) will continue to provide larger samples across multiple sites to contribute to our continued understanding of the value and limitations of the BOSCC. Though there is a critical need for outcome measures for FDA-recommended Randomized Controlled Trials (RCTs), we encourage the BOSCC be used in combination with other measures of change because the BOSCC is new and additional testing of its ability to capture meaningful change needs to be completed. Because the FDA requires researchers to specify a primary outcome measure prior to data collection, this can be a tricky decision but this

recommendation is consistent with suggestions from other researchers endorsing multiple means of assessing treatment outcome (Cunningham, 2012). Also, the BOSCC may be useful in clarifying potential placebo effects often found in caregiver reports, allowing for more effective use of parent report measures. As the field focuses efforts on finding appropriate outcome measures for longitudinal studies and RCTs, we look forward to the continued validation of measures such as the BOSCC that may provide unique, objective observational data to aid in assessing the efficacy and course of treatments aimed at improving social communication skills.



Table 1. Study 1: Background and First Observation Information (n=56).

	Mean (SD)	Range
Age (months)	28.9 (10.5)	12-56
VABS (Standard Score) (n=55)		
Communication	78.7 (17.5)	29-121
Socialization	79.0 (12.1)	32-110
Daily Living	84.0 (13.3)	36-117
Motor Skills	89.1 (13.8)	34-113
Adaptive Behavior Composite	79.6 (12.9)	32-113
MSEL (Ratio) (n=54)		
VIQ	62.9 (21.9)	29-123
NVIQ	78.5 (23.7)	30-145
ADOS-2 (n=55)		
CSS	7.6 (2.0)	3-10
SA CSS	7.7 (2.1)	3-10
RRB CSS	7.0 (2.1)	1-10
	n (%)	
ADOS-2 Module (Toddler or 1)	47 (85)	
Sex (Males)	44 (79)	
Race <sup>a</sup>		
Caucasian	34 (61)	
African American	5 (9)	
Other	5 (9)	
Ethnicity <sup>b</sup> (Hispanic)	6 (11)	
Maternal Education <sup>c</sup> (4+ years College)	30 (57)	

Note: ADOS-2= Autism Diagnostic Observation Schedule, 2nd Edition; CSS= Calibrated Severity Score; MSEL= Mullen Scales of Early Learning; RRB CSS= Restricted, Repetitive Behavior Calibrated Severity Scores; SA CSS= Social Affect Calibrated Severity Scores;SD= Standard Deviation; VABS= Vineland Adaptive Behavior Scales. <sup>a</sup>Twelve participants (21%) did not provide race information. <sup>b</sup>Four (7%) participants did not provide ethnicity information. <sup>c</sup>Three participants (5%) did not provide information about maternal education.

Table 2. Study 1: Information about Assessments Gathered.

Assessment	N with $\geq 2$ Observations	# of Observations (Mean)	# of Observations (Range)	# Months Between First and Last Observation (Mean)
BOSCC	50	3.4	1-8	5.9
ADOS-2	41	2.5	1-5	5.9
MSEL	36	2.0	1-3	9.2
VABS	31	2.1	1-3	9.5

Note: ADOS-2= Autism Diagnostic Observation Schedule, 2nd Edition; BOSCC= Brief Observation of Social Communication Change; MSEL= Mullen Scales of Early Learning; VABS= Vineland Adaptive Behavior Scales

Table 3. Study 1: Brief Observation of Social Communication Change (BOSCC)  
Exploratory Factor Analysis Model Comparison.

Model ( <i>df</i> )	$\chi^2$ Test of Model Fit	df	p	Eigenvalue	RMSEA
1-Factor (54)	221.29	54	<0.001	5.48	0.107
2-Factor (43)	101.85	43	<0.001	1.58	0.067
3-Factor (33)	46.70	33	0.057	1.05	0.037

Note: df=degrees of freedom; RMSEA= Root mean square error of approximation

Table 4. Study 1: 1, 2, and 3-Factor Model Factor Loadings for Brief Observation of Social Communication Change (BOSCC) Items.

Item Name (abbreviated)	1-Factor Model	2-Factor Model (promax)		3-Factor Model (promax)		
	Factor 1	Factor 1	Factor 2	Factor 1	Factor 2	Factor 3
Eye Contact	<b>0.66</b>	<b>0.78</b>	-0.06	<b>1.00</b>	-0.16	0.11
Facial Expressions	<b>0.51</b>	<b>0.62</b>	-0.09	<b>0.59</b>	0.05	-0.02
Gestures	<b>0.50</b>	<b>0.73</b>	-0.21	<b>0.42</b>	0.36	-0.27
Vocalizations	<b>0.77</b>	<b>0.63</b>	0.24	0.10	<b>0.80</b>	-0.05
Integration of Vocal and Non-Vocal	<b>0.84</b>	<b>0.87</b>	0.07	<b>0.65</b>	0.32	0.05
Social Overtures	<b>0.79</b>	<b>0.71</b>	0.16	<b>0.51</b>	0.34	0.09
Social Responses	<b>0.76</b>	<b>0.56</b>	0.31	0.14	<b>0.67</b>	0.05
Engagement	<b>0.62</b>	<b>0.40</b>	0.32	-0.06	<b>0.72</b>	0.03
Play	<b>0.50</b>	0.27	0.32	-0.15	<b>0.69</b>	0.01
Unusual Sensory Interests	<b>0.57</b>	-0.07	<b>0.85</b>	0.09	-0.05	<b>0.95</b>
Hand/Finger/Body Mannerisms	<b>0.40</b>	-0.12	<b>0.67</b>	-0.03	0.09	<b>0.55</b>
Repetitive Interests/Behaviors	<b>0.58</b>	0.17	<b>0.55</b>	0.06	0.32	0.38

Note: All factor loadings  $\geq 0.4$  shown in bold. Analyses included all available codings, which includes multiple codings from different coders of the same video (308 total codings).

Table 5. Study 1: Inter-Rater Reliability for Domains (28 videos, 2 randomly selected coders).

	Coder 1		Coder 2		Intra-Class Correlation (ICC)		ANOVA		
	Mean	SD	SE	Mean	SD	SE	Absolute Agreement	F	p
Overall Total (Items 1-15)	28.63	11.63	2.20	29.34	11.59	2.19	0.96	1.30	0.26
Core Subtotal (Items 1-12)	27.34	11.21	2.13	28.18	10.93	2.06	0.96	2.02	0.17
SC Subtotal (Items 1-8)	20.88	7.88	1.49	21.32	7.65	1.45	0.94	0.75	0.39
RRB Subtotal (Items 9-12)	6.45	4.31	0.81	6.82	4.26	0.81	0.94	1.73	0.20

Table 6. Study 1: Inter-Rater ICCs for Individual Items (Averaged across A and B) and Percent Agreement between Raters within 1 point.

	<i>Item Averaged Across Segment A and B</i>		<i>Percent Agreement within 1 point</i>	
	ICC (Absolute Agreement)	95% CI (Absolute Agreement)	Segment A	Segment B
Eye Contact	0.80	0.61, 0.90	93	96
Facial Expressions	0.79	0.55, 0.90	86	71
Gestures	0.88	0.76, 0.94	93	82
Vocalizations	0.91	0.81, 0.96	96	82
Integration of Vocal and Non-Vocal	0.86	0.73, 0.93	89	93
Social Overtures	0.87	0.72, 0.94	75	79
Social Responses	0.79	0.54, 0.90	82	82
Engagement	0.75	0.46, 0.88	86	68
Play	0.71	0.37, 0.86	93	82
Unusual Sensory Interests	0.93	0.86, 0.97	93	93
Hand/Finger/Body Mannerisms	0.83	0.66, 0.92	89	82
Repetitive/Stereotyped Interests/Behaviors	0.91	0.82, 0.96	89	89

Note: 28 videos total with two randomly selected coders.

Table 7. Study 1: Test-Retest for Domains (20 children, two videos within 1 month; 40 videos total).

	Intra-Class Correlation (ICC)										ANOVA	
	Video 1					Video 2					F	p
	Mean	SD	SE	Mean	SD	SE	Absolute Agreement					
Overall Total (Items 1-15)	27.80	12.23	2.73	26.35	10.24	2.29	0.91	1.01	0.33			
Core Subtotal (Items 1-12)	26.85	11.98	2.68	25.15	9.58	2.14	0.90	1.34	0.26			
SC Subtotal (Items 1-8)	21.45	8.96	2.00	20.75	7.63	1.71	0.87	0.31	0.59			
RRB Subtotal (Items 9-12)	5.43	3.93	0.88	4.50	2.61	0.58	0.78	2.26	0.15			

Table 8. Study 1: Test-Retest ICCs for Individual Items (Averaged across A and B).

	<i>Item Averaged Across Segment A and B</i>	
	ICC (Absolute Agreement)	95% CI (Absolute Agreement)
Eye Contact	0.53	0.22, 0.82
Facial Expressions	0.81	0.53, 0.93
Gestures	0.69	0.20, 0.88
Vocalizations	0.89	0.73, 0.96
Integration of Vocal and Non-Vocal	0.81	0.51, 0.93
Social Overtures	0.63	0.04, 0.86
Social Responses	0.72	0.28, 0.898
Engagement	0.57	0.11, 0.83
Play	0.59	0.03, 0.84
Unusual Sensory Interests	0.72	0.28, 0.89
Hand/Finger/Body Mannerisms	0.74	0.34, 0.90
<u>Repetitive/Stereotyped Interests/Behaviors</u>	0.36	-0.56, 0.74

Note: 20 children with 2 videos within 1 month; 40 videos total.



Table 9. Study 2: Background Information (n=78).

	n (%)
Gender (Male)	63 (81)
Race <sup>a</sup>	
White	51 (71)
African American	6 (8)
Asian	5 (7)
More than one race	10 (14)
Hispanic <sup>b</sup>	12 (15)
Site	
UCLA	11 (14)
CADB	29 (37)
Rochester	10 (13)
Vanderbilt	28 (36)
	Mean (SD)
Age (TP1)	6.19 (1.29)
IQ (Leiter) <sup>c</sup>	61.04 (17.33)
ADOS-2 (Module 1)	
SA Total	14.51 (2.65)
RRB Total	5.28 (1.77)
CSS	7.22 (1.30)

Note: <sup>a</sup>8% (n=6) did not report race; <sup>b</sup>1% (n=1) did not report; <sup>c</sup>14% (n=11) did not complete Leiter; ADOS= Autism Diagnostic Observation Schedule, 2nd Edition, IQ=Intelligence Quotient, SA=Social Affect, RRB=Restricted, Repetitive Behavior, CSS= Calibrated Severity Score, TP=Time Point, UCLA = University of California at Los Angeles

Table 10. Study 2: Inter-Rater Reliability (22 videos, 2 randomly selected coders)

	Coder 1			Coder 2			Intra-Class Correlation (ICC)	ANOVA	
	Mean	SD	SE	Mean	SD	SE	Absolute Agreement	F	<i>p</i>
Overall	37.86	6.82	1.46	38.61	6.31	1.35	0.83	0.82	0.38
Core	35.64	6.54	1.39	36.27	6.03	1.29	0.85	0.74	0.40
SC	28.50	5.80	1.24	28.59	4.62	0.98	0.87	0.02	0.88
RRB	7.14	2.88	0.61	7.68	2.93	0.62	0.70	1.32	0.26

Table 11. Study 2: Inter-Rater ICCs for Individual Items (Averaged across A and B) and Percent Agreement between Raters within 1 point.

	<i>Item Averaged Across Segment A and B</i>		<i>Percent Agreement within 1 point</i>	
	ICC (Absolute Agreement)	95% CI (Absolute Agreement)	Segment A	Segment B
Eye Contact	0.75	0.49, 0.89	86	95
Facial Expressions	0.79	0.57, 0.91	100	95
Gestures	0.80	0.57, 0.91	77	95
Vocalizations	0.89	0.75, 0.95	91	91
Integration of Vocal and Non-Vocal	0.83	0.63, 0.93	91	91
Social Overtures	0.75	0.48, 0.89	91	91
Social Responses	0.46	0.06, 0.74	86	82
Engagement	0.52	0.13, 0.77	91	77
Play	0.55	0.12, 0.82	95	100
Unusual Sensory Interests	0.72	0.44, 0.88	82	73
Hand/Finger/Body Mannerisms	0.81	0.60, 0.92	86	95
Repetitive/Stereotyped Interests/Behaviors	0.50	0.11, 0.75	64	73

Note: 22 videos coded by two randomly selected coders.

Table 12. Study 2: BOSCC, Leiter, and PPVT-4 Change Over Time.

	Entry	Early Response <sup>a</sup>	Exit	$\Delta$ Entry to Exit	Paired t-test (Entry to Exit)		
	Mean (SD)	Mean (SD)	Mean (SD)	Mean (SD)	t	df	p
BOSCC <sup>b</sup>							
Overall	37.84 (7.80)	37.57 (7.87)	38.02 (8.04)	0.18 (6.62)	-0.24	77	0.81
Core	35.97 (7.13)	35.55 (7.26)	36.03 (7.40)	0.05 (5.80)	-0.08	77	0.94
SC	27.75 (5.68)	27.80 (5.92)	27.67 (5.96)	-0.08 (4.71)	0.16	77	0.88
RRB	8.22 (3.40)	7.75 (2.85)	8.35 (3.60)	0.13 (3.17)	-0.38	77	0.71
Leiter Brief IQ <sup>c</sup>	61.89 (18.15)	-	60.85 (20.28)	-1.04 (11.54)	0.66	54	0.51
PPVT-4 (SS) <sup>d</sup>	44.60 (20.14)	43.08 (20.05)	44.55 (19.84)	-0.05 (10.01)	0.04	64	0.97
RBS-R Total <sup>b, e</sup>	30.08 (17.81)	-	27.20 (15.50)	-2.88 (11.11)	2.25	74	0.03

Note: <sup>a</sup>2 children missing midpoint data; <sup>b</sup>Decreases over time indicate improvements; <sup>c</sup>55 children have Leiter available at entry and exit; <sup>d</sup>65 children have PPVT available at entry and exit; <sup>e</sup>75 children have RBSR at entry and exit; BOSCC= Brief Observation of Social Communication Change; PPVT= Peabody Picture Vocabulary Test; RBSR= Repetitive Behavior Scale Revised; RRB= Restricted, Repetitive Behavior Total; SC= Social Communication Total; SS = Standard Score.

Table 13. Study 2: Frequency CGI and BOSCC Core Responders.

	CGI-I Responder	CGI-I Non- Responder	$\chi^2$	df	<i>p</i>
BOSCC Core Responder	6	3	0.05	1	0.83
BOSCC Core Non- Responder	45	19			

Note: 5 children are missing CGI information at exit.

Table 14. Study 2: T-Test CGI-I Groups and BOSCC Change from TP 1 to TP 3.

	CGI-I Responder	CGI-I Non-Responder	t	df	p
	<i>CGI-I (1 or 2), n=51</i>	<i>CGI-I (3+), n=22</i>			
	Mean (SD)	Mean (SD)			
<b>BOSCC Change</b>					
Overall	0.01 (6.38)	1.89 (6.96)	1.12	71	0.27
Core	0.06 (5.80)	1.30 (5.65)	0.84	71	0.40
SC	0.07 (4.77)	0.52 (3.95)	0.39	71	0.70
RRB	-0.01 (2.95)	0.77 (2.97)	1.04	71	0.30
Leiter Brief IQ Change	-0.70 (11.25)	-3.07 (9.81)	0.69	49	0.49
PPVT-4 Change	-0.26 (10.09)	-0.53 (10.47)	-0.09	59	0.93
RBS-R Change	-1.92 (10.65)	-3.10 (11.34)	-0.42	68	0.68

Note: Discrepancies in df are related to missing data for Leiter, PPVT, and RBS-R. Negative numbers indicate lower scores over time (corresponding to improvement on BOSCC and worsening on Leiter, PPVT, and RBS-R). Results are consistent with TP1 to TP 2.

Table 15. Study 2: Logistic Regression for Entry Measures Predicting CGI-I and BOSCC Core Responder Status

Predictor	CGI-I Responder	CGI-I Non-Responder	Logistic Regression			
	<i>CGI-I (1 or 2), n=51</i>	<i>CGI-I (3+), n=22</i>	$\beta$	<i>SE</i>	<i>e<sup>B</sup></i>	<i>p</i>
ADOS-2 CSS	7.20 (1.28)	7.32 (1.36)	-0.04	0.20	0.96	0.82
Leiter Brief IQ	59.63 (19.48)	59.67 (20.63)	0.02	0.02	1.02	0.16
VABS Socialization	62.49 (9.13)	59.16 (6.93)	0.04	0.03	1.04	0.22
VABS Communication	60.26 (11.11)	60.42 (12.23)	0.01	0.02	1.01	0.69
VABS Daily Living	64.65 (11.11)	63.37 (11.46)	0.03	0.02	1.03	0.19

Predictor	BOSCC Core Responder	BOSCC Core Non-Responder	$\beta$	<i>SE</i>	<i>e<sup>B</sup></i>	<i>p</i>
	<i>n=12</i>	<i>n=66</i>				
ADOS-2 CSS	7.58 (1.17)	7.15 (1.29)	0.27	0.25	1.31	0.28
Leiter Brief IQ	62.00 (20.83)	59.90 (20.24)	0.01	0.02	1.01	0.50
VABS Socialization	62.55 (14.51)	61.27 (7.17)	0.02	0.04	1.02	0.65
VABS Communication	58.91 (8.48)	60.90 (12.23)	-0.02	0.03	0.99	0.60
VABS Daily Living	65.36 (14.89)	63.93 (10.84)	0.11	0.03	1.01	0.70

Table 16. Study 2: Leiter, PPVT-4, & RBS-R Change Groups and Change in BOSCC Domains from TP 1 to TP 3.

	Leiter Responder (n=12)	Leiter Non-Responder (n=43)	t	df	p-value
	Mean (SD)	Mean (SD)			
<b>BOSCC Change</b>					
Overall	2.38 (6.74)	-0.31 (6.28)	-1.29	53	0.20
Core	1.71 (6.05)	-0.13 (5.54)	-1.00	53	0.32
SC	0.42 (5.53)	-0.55 (4.46)	-0.63	53	0.53
RRB	1.29 (2.45)	0.42 (3.09)	-0.90	53	0.37
	PPVT-4 Responder (n=14)	PPVT-4 Non-Responder (n=51)			
<b>BOSCC Change</b>					
Overall	-0.04 (6.02)	0.38 (6.62)	0.21	63	0.83
Core	-0.50 (5.13)	0.26 (5.83)	0.44	63	0.66
SC	-0.39 (4.83)	0.09 (4.45)	0.35	63	0.73
RRB	-0.11 (2.18)	0.17 (3.34)	0.29	63	0.77
	RBS-R Responder (n=23)	RBS-R Non-Responder (n=52)			
<b>BOSCC Change</b>					
Overall	0.24 (8.01)	0.22 (6.04)	0.01	73	0.99
Core	-0.15 (6.59)	0.13 (5.53)	0.20	73	0.85
SC	0.30 (5.01)	-0.34 (4.68)	-0.54	73	0.59
RRB	-0.46 (3.95)	0.48 (2.70)	1.20	73	0.24

Note: BOSCC= Brief Observation of Social Communication Change; PPVT= Peabody Picture Vocabulary Test; RBSR= Repetitive Behavior Scale Revised; RRB= Restricted, Repetitive Behavior Total; SC= Social Communication Total.



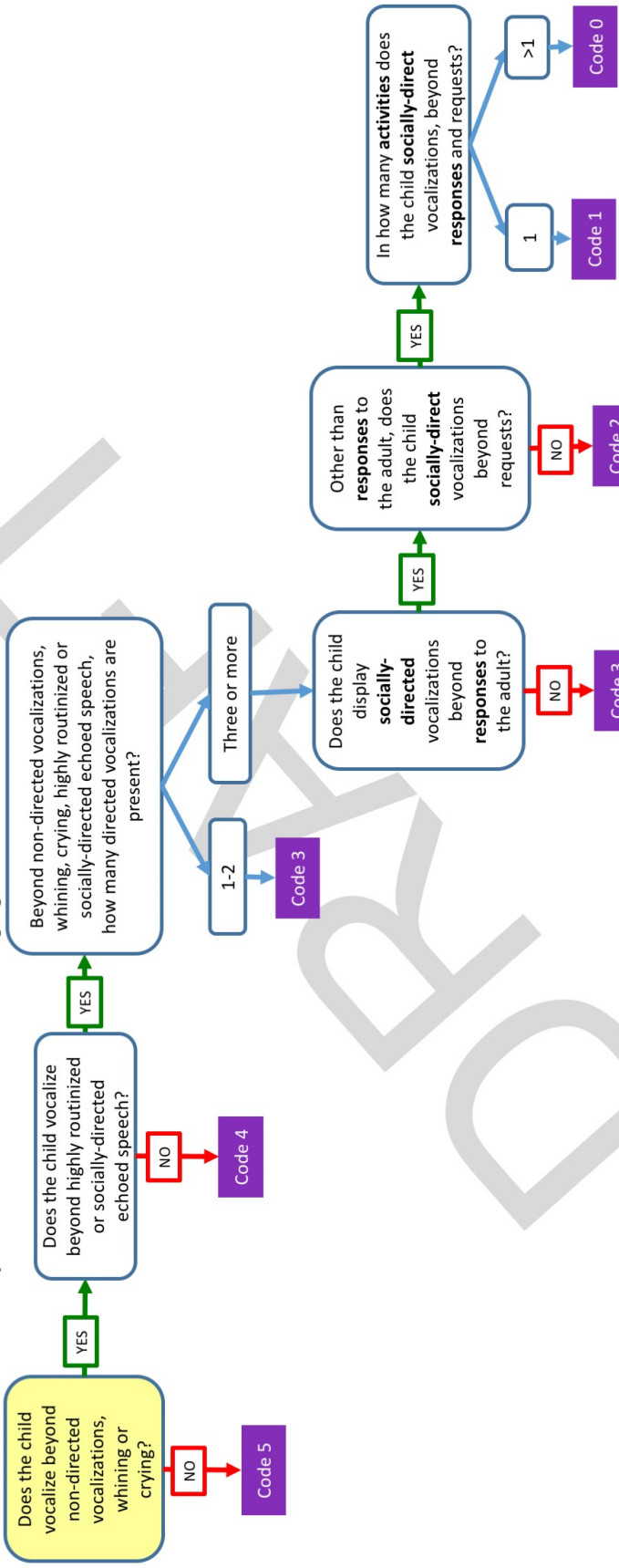
Table 17. Study 2: BOSCC Scores and BOSCC Change Correlations with Entry Measures: ADOS-2, VABS, IQ, and Age.

	BOSCC Domain Totals at Entry				BOSCC Domain Totals at Exit				Δ BOSCC Entry To Exit			
	SC	RRB	Core	Overall	SC	RRB	Core	Overall	SC	RRB	Core	Overall
ADOS-2 CSS	0.30*	0.02	0.25	0.27	0.15	0.02	0.14	0.14	-0.17	0.01	-0.13	-0.15
Leiter Brief IQ (Entry)	-0.13	-0.36*	-0.26	-0.30	-0.18	-0.22	-0.24	-0.26	-0.06	0.12	0.02	0.04
VABS Socialization	-0.16	-0.25	-0.25	-0.29	-0.14	-0.17	-0.19	-0.19	0.03	0.08	0.07	0.13
VABS Communication	-0.22	-0.28	-0.31*	-0.37*	-0.13	-0.11	-0.16	-0.17	0.11	0.17	0.19	0.24
VABS Daily Living	-0.33*	-0.35*	-0.43*	-0.47*	-0.24	-0.22	-0.30	-0.30	0.11	0.12	0.15	0.21
Age (at entry)	-0.22	0.18	-0.09	-0.09	-0.14	0.06	-0.08	-0.13	0.09	-0.12	0.01	-0.04

Note: \*<0.01

Figure 1. BOSCC Example Item

**4. Vocalizations Directed to Others.** Coding for this item considers 1) the frequency of vocalizations, 2) the social-directedness of vocalizations, 3) the type of vocalizations, and 4) the number of activities in which directed vocalizations occur. Socially-directed non-speech sounds should be considered here. Non-socially-directed vocalizations such singing or other odd non-social vocalizations can be counted here in a code of 5.



\*\*Example of prepublication/draft material for the BOSCC (formerly known as “ADOS-Change”) copyright © 2012-2015 by Western Psychological Services. Reprinted by the author with permission of the publisher, for the sole purpose of scholarly reference. All rights reserved (rights@wpspublish.com).\*\*

A. Item 4 Code=

Figure 2. Study 1: Distributions for 12 Core BOSCC Items (Averaged across Segments A and B).



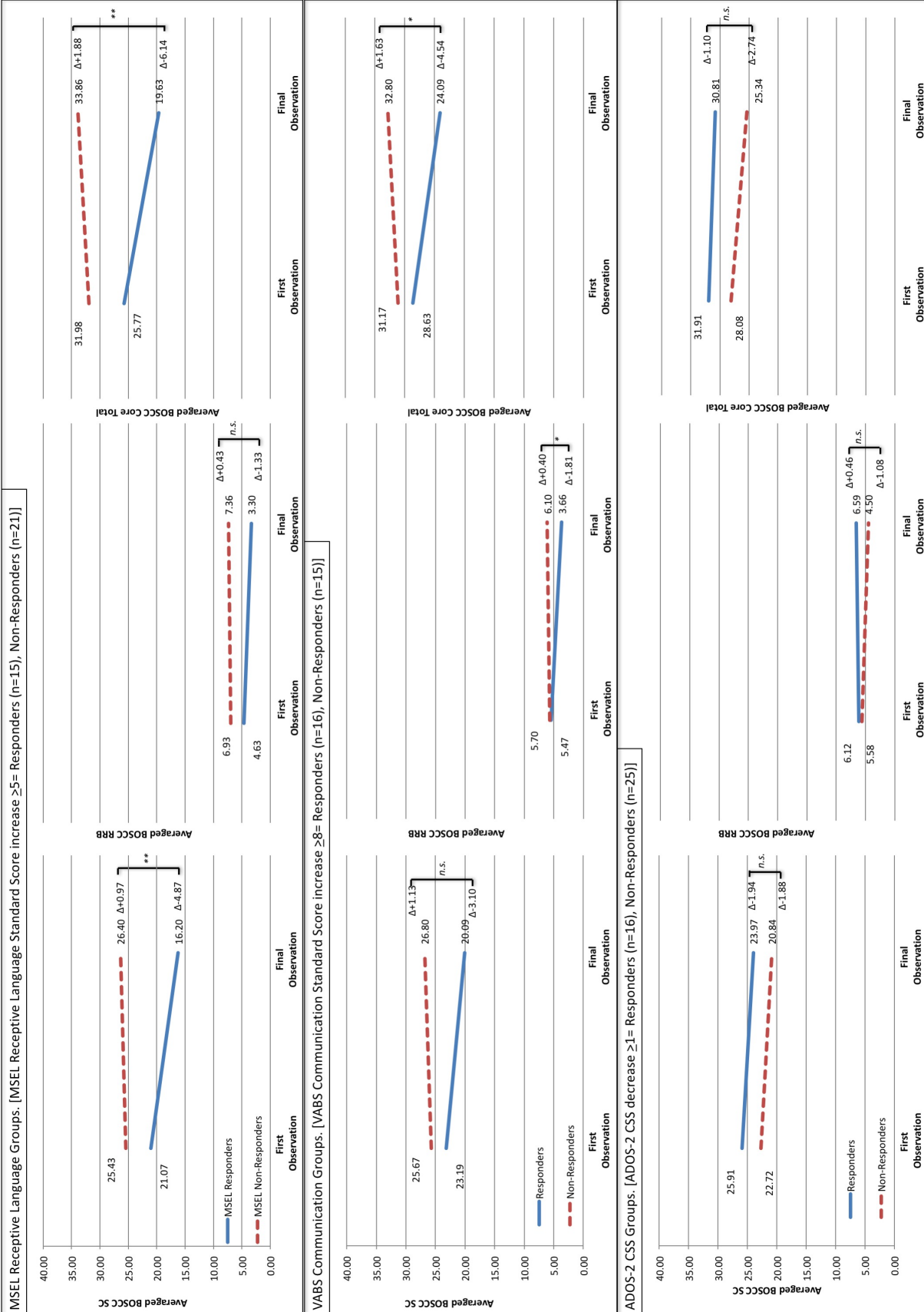
Note: Solid red represents items in the Social Communication domain; Stripped blue represents items in the Restricted, Repetitive Behaviors Domain.

Figure 3. Study 1: Exploratory Factor Domains.

	Item	Domain	Total
1	Eye Contact	Social-Communication	Core
2	Facial Expressions		
3	Gestures		
4	Vocalizations		
5	Integration of Vocal and Non-Vocal		
6	Social Overtures		
7	Social Responses		
8	Engagement		
9	Play	RRB	
10	Unusual Sensory Interests		
11	Hand/Finger/Body Mannerisms		
12	Repetitive/Stereotyped Interests/Behaviors		
13	Activity Level	Other Abnormal Behaviors	
14	Disruptive Behavior/Irritability		
15	Anxious Behaviors		

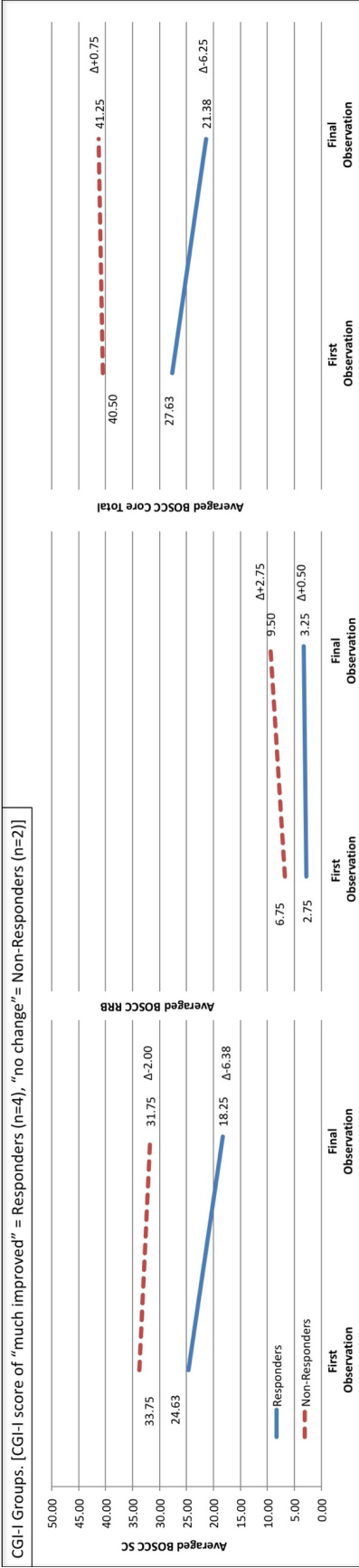
Note: RRB= Restricted/Repetitive Behavior/Interest

**Figure 4. Study 1: Responder Groups Defined by MSEL, VABS, or ADOS-2 in Early Intervention Studies.**



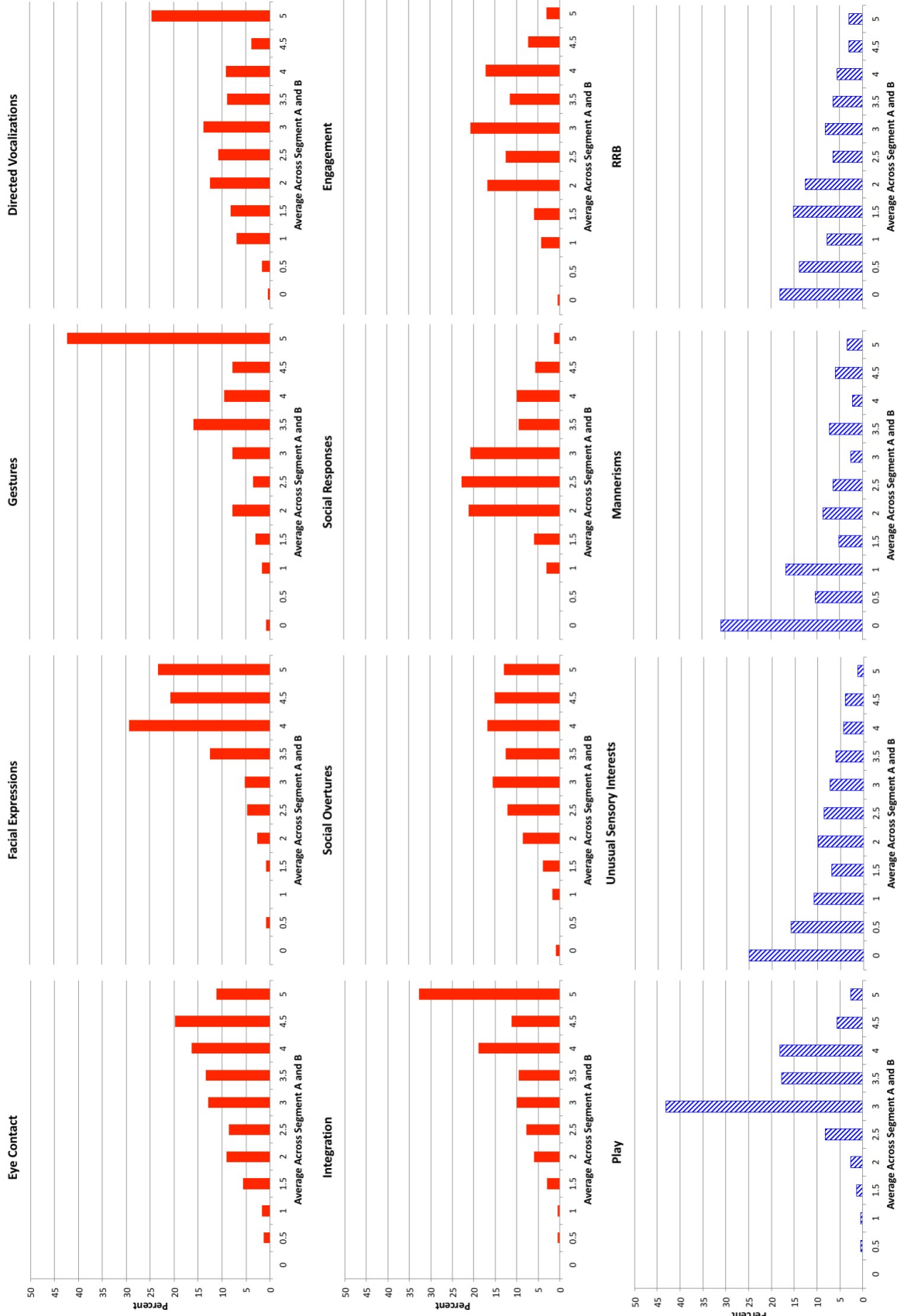
Note: \* $p < 0.05$ , \*\* $p < 0.01$ , \*\*\* $p < 0.001$ ; ADOS-2= Autism Diagnostic Observation Schedule, 2<sup>nd</sup> Edition; BOSCC= Brief Observation of Social Communication Change; CSS= ADOS Calibrated Severity Score; MSEL= Mullen Scales of Early Learning; n.s.= not significant; RRB= Restricted, Repetitive Behaviors BOSCC Domain; VABS= Vineland Adaptive Behavior Scales

**Figure 5. Study 1: Responder Groups defined by Clinician Global Impression (CGI) in Community-Based Intervention Study.**



Note: BOSCC= Brief Observation of Social Communication Change; CGI= Clinician's Global Impression-Improvement; RRB= Restricted, Repetitive Behaviors BOSCC Domain; SC= Social Communication BOSCC Domain

Figure 6. Study 2: Distributions for 12 Core BOSCC Items (Averages Across segments A and B).



Note: Solid red represents items in the Social Communication domain; striped blue represents items in the Restricted, Repetitive Behaviors domain.

## REFERENCES

- Almirall, D., Compton, S. N., Rynn, M. A., Walkup, J. T., & Murphy, S. A. (2012). SMARTer discontinuation trial designs for developing an adaptive treatment strategy. *Journal of child and adolescent psychopharmacology*, 22(5), 364-374.
- Aman, M.G., Singh, N.N., Stewart, A.W., & Field, C.J. (1985). The Aberrant Behavior Checklist: A behavior rating scale for the assessment of treatment effects. *Am J Ment De*, 89, 485-491.
- American Educational Research Association (AERA), American Psychological Association (APA), & National Council on Measurement in Education (NCME). (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.
- American Psychiatric Association (APA). (2013). *Diagnostic and statistical manual of mental disorders*. 5th ed. Washington, DC: APA.
- Anagnostou, E., Jones, N., Huerta, M., Halladay, A. K., Wang, P., Scahill, L., . . . Dawson, G. (2015). Measuring social communication behaviors as a treatment endpoint in individuals with autism spectrum disorder. *Autism*, 19(5), 622-636. doi:10.1177/1362361314542955
- Anan, R. M., Warner, L. J., McGillivray, J. E., Chong, I. M., & Hines, S. J. (2008). Group Intensive Family Training (GIFT) for preschoolers with autism spectrum disorders. *Behavioral Interventions*, 23(3), 165-180.
- Anderson, D. K., Liang, J. W., & Lord, C. (2014). Predicting young adult outcome among more and less cognitively able individuals with autism spectrum disorders. *J Child Psychol Psychiatry*, 55(5), 485-494. doi:10.1111/jcpp.12178
- Anderson, D. K., Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., ... & Pickles, A. (2007). Patterns of growth in verbal abilities among children with autism spectrum disorder. *Journal of consulting and clinical psychology*, 75(4), 594.
- Anderson, D. K., Oti, R. S., Lord, C., & Welch, K. (2009). Patterns of growth in adaptive social abilities among children with autism spectrum disorders. *Journal of abnormal child psychology*, 37(7), 1019-1034.
- Asperger, H. (1979). Problems of infantile autism. *Communication*, 13, 45-52.
- Baio J. (2014). Prevalence of Autism Spectrum Disorder Among Children Aged 8 Years — Autism and Developmental Disabilities Monitoring Network, 11 Sites, United States, 2010. CDC. *MMWR Surveillance Summaries*, 63(2).



- Biringen, Z., Robinson, J. L., & Emde, R. N. (2000). Appendix B: The emotional availability scales (; an abridged infancy/early childhood version). *Attachment & human development, 2*(2), 256-270.
- Bishop, D. V. M. (2006). *The Children's Communication Checklist, Second Edition*, U.S. Edition. New York, NY: The Psychological Corporation.
- Bishop, S. L., Guthrie, W., Coffing, M., & Lord, C. (2011). Convergent validity of the mullen scales of early learning and the differential ability scales in children with Autism Spectrum Disorders. *American Association on Intellectual and Developmental Disabilities, 116*(5), 331-343.
- Bishop, S. L., Havdahl, K. A., Huerta, M., & Lord, C. (2016). Subdimensions of social-communication impairment in autism spectrum disorder. *Journal of Child Psychology and Psychiatry.*
- Bishop, S. L., Hus, V., Duncan, A., Huerta, M., Gotham, K., Pickles, A., ... & Lord, C. (2013). Subcategories of restricted and repetitive behaviors in children with autism spectrum disorders. *Journal of autism and developmental disorders, 43*(6), 1287-1297.
- Bishop, S. L., Richler, J., & Lord, C. (2006). Association between restricted and repetitive behaviors and nonverbal IQ in children with autism spectrum disorders. *Child neuropsychology, 12*(4-5), 247-267.
- Bone, D., Black, M. P., Lee, C. C., Williams, M. E., Levitt, P., Lee, S., & Narayanan, S. (2012, September). Spontaneous-Speech Acoustic-Prosodic Features of Children with Autism and the Interacting Psychologist. In *InterSpeech* (pp. 1043-1046).
- Bone, D., Lee, C. C., Black, M. P., Williams, M. E., Lee, S., Levitt, P., & Narayanan, S. (2014). The psychologist as an interlocutor in autism spectrum disorder assessment: Insights from a study of spontaneous prosody. *Journal of Speech, Language, and Hearing Research, 57*(4), 1162-1177.
- Brian, J., Smith, I., Zwaigenbaum, L., Roberts, W., & Bryson, S. (2015). The Social ABCs Caregiver-Mediated Intervention for Toddlers With Autism Spectrum Disorder: Feasibility, Acceptability, and Evidence of Promise From a Multisite Study. *Autism Research*, Epub ahead of print. doi: 10.1002/aur.1582.
- Bodfish, J. W., Symons, F. J., Parker, D. E., & Lewis, M. H. (2000). Varieties of repetitive behavior in autism: Comparisons to mental retardation. *Journal of autism and developmental disorders, 30*(3), 237-243.
- Bolte, E. E., & Diehl, J. J. (2013). Measurement tools and target symptoms/skills used to assess treatment response for individuals with autism spectrum disorder. *J Autism Dev Disord, 43*(11), 2491-2501. doi:10.1007/s10803-013-1798-7

- Boomsma, A., Van Lang, N. D., De Jonge, M. V., De Bildt, A. A., Van Engeland, H., & Minderaa, R. B. (2008). A new symptom model for autism cross-validated in an independent sample. *J Child Psychol Psychiatry*, *49*(8), 809-816.
- Boucher, J. (2012). Research review: structural language in autistic spectrum disorder – characteristics and causes. *J Child Psychol Psychiatry*, *53*.
- Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K. A. Bollen & J. S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.
- Busner, J., & Targum, S. D. (2007). The Clinical Global Impression Scale. *Psychiatry*, *4*(7), 28-37.
- Cannon, D. S., Miller, J. S., Robison, R. J., Villalobos, M. E., Wahmhoff, N. K., Allen-Brady, K., ... & Coon, H. (2010). Genome-wide linkage analyses of two repetitive behavior phenotypes in Utah pedigrees with autism spectrum disorders. *Molecular autism*, *1*(1), 3.
- Capone, G. T., Grados, M. A., Kaufmann, W. E., Bernad-Ripoll, S., & Jewell, A. (2005). Down syndrome and comorbid autism-spectrum disorder: Characterization using the aberrant behavior checklist. *American journal of medical genetics Part A*, *134*(4), 373-380.
- Chambless, D. L., & Hollon, S. D. (1998). Defining empirically supported therapies. *J Consult Clin Psychol*, *66*(1), 7-18.
- Chan, E. K. (2014). Standards and guidelines for validation practices: development and evaluation of measurement instruments. In *Validity and validation in social, behavioral, and health sciences* (pp. 9-24). Springer International Publishing.
- Constantino, J. N. (2002). *The Social Responsiveness Scale*. Los Angeles: Western Psychological Services.
- Constantino, J. N., Davis, S. A., Todd, R. D., Schindler, M. K., Gross, M. M., Brophy, S. L., ... & Reich, W. (2003). Validation of a brief quantitative measure of autistic traits: comparison of the social responsiveness scale with the autism diagnostic interview-revised. *Journal of autism and developmental disorders*, *33*(4), 427-433.
- Cuccaro, M. L., Shao, Y., Grubber, J., Slifer, M., Wolpert, C. M., Donnelly, S. L., ... & Pericak-Vance, M. A. (2003). Factor analysis of restricted and repetitive behaviors in autism using the Autism Diagnostic Interview-R. *Child psychiatry and human development*, *34*(1), 3-17.
- Cunningham, A. (2012). Measuring change in social interaction skills of young children with Autism. *Journal of Autism and Developmental Disorders*, *42*, 593-605.

- Daniel, J. T., & Wood, J. J. (2013). Cognitive behavioral therapy for children with autism: review and considerations for future research. *J Dev Behav Pediatr*, *34*(9), 702-715. doi:10.1097/DBP.0b013e31829f676c
- Daniels, A.M., Halladay, A.K., Shih, A., Elder, L.M., & Dawson, G. (2014). Approaches to enhancing the early detection of autism spectrum disorders: a systematic review of the literature. *J Am Acad Child Adolesc Psychiatry*, *53*(2), 141-152.
- Daniels, A. M., & Mandell, D. S. (2013). Explaining differences in age at autism spectrum disorder diagnosis: A critical review. *Autism*, 1362361313480277.
- Dawson, G. & Bernier, R. (2013). A quarter century of progress on the early detection and treatment of autism spectrum disorder. *Development and Psychopathology*, *25*(4), 1455-1472.
- Dawson, G., Toth, K., Abbott, R., Osterling, J., Munson, J., Estes, A., & Liaw, J. (2004). Early social attention impairments in autism: social orienting, joint attention, and attention to distress. *Developmental psychology*, *40*(2), 271.
- Dawson, G., Rogers, S., Munson, J., Smith, M., Winter, J., Greenon, J., . . . Varley, J. (2010). Randomized, controlled trial of an intervention for toddlers with autism: the Early Start Denver Model. *Pediatrics*, *125*(1), e17-23. doi:10.1542/peds.2009-0958
- De Bildt, A. (2004). Interrelationship between Autism Diagnostic Observation Schedule—Generic (ADOS-G), Autism Diagnostic Interview—Revised (ADI-R), and the Diagnostic and Statistical Manual of Mental Disorders (DSM-IV-TR) classification in children and adolescents with mental retardation. *Journal of Autism and Developmental Disorders* *34*: 129–137.
- DeGeorge, A., Dufek, S., & Lord, C. (in prep). Effects of parent mediated coaching vs. psychoeducation on social communication development of underserved young children with autism: A pilot study.
- Dunn, L. M., & Dunn, D. M. (2007). Peabody Picture Vocabulary Test Fourth Edition. Minneapolis, MN: NCS Person. Inc. *Measure used with Cohort*, 3.
- Eapen, V., Črnčec, R., & Walter, A. (2013). Clinical outcomes of an early intervention program for preschool children with autism spectrum disorder in a community group setting. *BMC pediatrics*, *13*(1), 3.
- Esbensen, A. J., Seltzer, M. M., Lam, K. S., & Bodfish, J. W. (2009). Age-related differences in restricted repetitive behaviors in autism spectrum disorders. *Journal of autism and developmental disorders*, *39*(1), 57-66.
- Elliot, C. D. (2007). *Differential Ability Scales - Second Edition*. New York: Harcourt Brace Jovanovich.

- Esler, A. N., Bal, V. H., Guthrie, W., Wetherby, A., Ellis Weismer, S., & Lord, C. (2015). The Autism Diagnostic Observation Schedule, Toddler Module: Standardized Severity Scores. *J Autism Dev Disord*, *45*(9), 2704-2720. doi:10.1007/s10803-015-2432-7
- Estes, A., Munson, J., Rogers, S., Greenson, J., Winter, J., & Dawson, G. (2015). Long-Term Outcomes of Early Intervention in 6-Year-Old Children With Autism Spectrum Disorder. *J Am Acad Child Adolesc Psychiatry*, *54*(7), 580-587.
- Fein, D., Barton, M., Eigsti, I.M., Kelley, E., Naigles, L.R., Schultz, R.T., Tyson, K. (2013). Optimal outcome in individuals with a history of autism. *J. Child Psychol. Psychiatry*, *54* (2). <http://dx.doi.org/10.1111/jcpp.12037>.
- Fletcher-Watson, S. & McConachie, H. (2015). The search for an early intervention outcome measurement tool in autism. *Focus on Autism and Other Developmental Disabilities*, 1-10.
- Fletcher-Watson, S., Petrou, A., Scott-Barrett, J., Dicks, P., Graham, C., & O'Hare, A., et al. (2015). A trial of an iPad intervention targeting social communication skills in children with autism. *Autism*, (Epub ahead of print).
- Frazier, T., Youngstrom, E., Speer, L., Embacher, R., Law, P., Constantino, J. (2012). Validation of proposed DSM-5 criteria for autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry*, *51*, 28-40. 10.1016/j.jaac.2011.09.021.
- Frost, K. M., Hong, N., & Lord, C. (2017). Correlates of Adaptive Functioning in Minimally Verbal Children With Autism Spectrum Disorder. *American Journal on Intellectual and Developmental Disabilities*, *122*(1), 1-10.
- Georgiades, S., Papageorgiou, V., & Anagnostou, E. (2010). Brief report: Repetitive behaviours in Greek individuals with autism spectrum disorder. *Journal of autism and developmental disorders*, *40*(7), 903-906.
- Georgiades, S., Szatmari, P., Boyle, M., Hanna, S., Duku, & Zwaigenbaum, L. (2013). Investigating phenotypic heterogeneity in children with autism spectrum disorder: a factor mixture modeling approach. *J Child Psychol Psychiatry*, *54*, 206-215.
- Georgiades, S., Szatmari, P., Zwaigenbaum, L., Duku, E., Bryson, S., Roberts, W., ... & Mahoney, W. (2007). Structure of the autism symptom phenotype: A proposed multidimensional model. *Journal of the American Academy of Child & Adolescent Psychiatry*, *46*(2), 188-196.
- Ginn, N. C., Clionsky, L. N., Eyberg, S. M., Warner-Metzger, C., & Abner, J. P. (2015). Child-Directed Interaction Training for Young Children With Autism Spectrum Disorders: Parent and Child Outcomes. *J Clin Child Adolesc Psychol*, 1-9.

- Gotham, K., Pickles, A., & Lord, C. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *J Autism Dev Disord*, *39*(5), 693-705. doi: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), e1278-1284. doi:10.1542/peds.2011-3668
- Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., ... & Sigman, M. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) revised algorithms. *Journal of the American Academy of Child & Adolescent Psychiatry*, *47*(6), 642-651.
- Gotham, K., Risi, S., Pickles, A., & Lord, C. (2007). The Autism Diagnostic Observation Schedule: revised algorithms for improved diagnostic validity. *Journal of autism and developmental disorders*, *37*(4), 613-627.
- Gresham, F.M. & Elliott, S.N. (2008). *Social Skills Improvement System Rating Scales*. Minneapolis, MN: NCS Pearson.
- Green, J., Charman, T., McConachie, H., Aldred, C., Slonims, V., Howlin, P., . . . Consortium, P. (2010). Parent-mediated communication-focused treatment in children with autism (PACT): a randomised controlled trial. *Lancet*, *375*(9732), 2152-2160. doi:10.1016/S0140-6736(10)60587-9
- Grindle, C. F., Hastings, R. P., Saville, M., Hughes, J. C., Huxley, K., Kovshoff, H., ... & Remington, B. (2012). Outcomes of a behavioral education model for children with autism in a mainstream school setting. *Behavior modification*, *36*(3), 298-319.
- Grzadzinski, R., Huerta, M., & Lord, C. (2013). DSM-5 and autism spectrum disorders (ASDs): an opportunity for identifying ASD subtypes. *Mol Autism*, *4*(1).
- Guastella, A. J., Gray, K. M., Rinehart, N. J., Alvares, G. A., Tonge, B. J., Hickie, I. B., . . . 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. *J Child Psychol Psychiatry*, *56*(4), 444-452. doi:10.1111/jcpp.12305
- Guthrie, W., Swineford, L.B., Nottke, C., & Wetherby, A.M. (2013). Early diagnosis of autism spectrum disorder: Stability and change in clinical diagnosis and symptom presentation. *Journal of Child Psychology and Psychiatry*, *54*(5), 582-590.
- Guthrie, W., Swineford, L. B., Wetherby, A. M., & Lord, C. (2013). Comparison of DSM-IV and DSM-5 factor structure models for toddlers with autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry*, *52*(8), 797-805 e792. doi:10.1016/j.jaac.2013.05.004
- Gutstein, S., Burgess, A., & Montfort, K. (2007). Evaluation of the relationship development intervention program. *Autism*, *11*(5), 397-411.

- Guy, W. (1976). Clinical global impression scale. *The ECDEU Assessment Manual for Psychopharmacology-Revised Volume DHEW Publ No ADM, 76(338)*, 218-222.
- Hansen, S. G., Blakely, A. W., Dolata, J. K., Raulston, T., & Machalicek, W. (2014). Children with autism in the inclusive preschool classroom: A systematic review of single-subject design interventions on social communication skills. *Review Journal of Autism and Developmental Disorders, 1(3)*, 192-206.
- Helt, M., Kelley, E., Kinsbourne, M., Pandey, J., Boorstein, H., Herbert, M., Fein, D. (2008). Can children with autism recover? If so, how? *Neuropsychol. Rev.*, 18 (4). <http://dx.doi.org/10.1007/s11065-008-9075-9>.
- Hobson, J. A., Tarver, L., Beurkens, N., & Peter Hobson, R. (2016). The Relation between Severity of Autism and Caregiver-Child Interaction: A Study in the Context of Relationship Development Intervention. *J Abnorm Child Psychol, 44(4)*, 745-755.
- Hoekstra, R. A., Bartels, M., Verweij, C. J., & Boomsma, D. I. (2007). Heritability of autistic traits in the general population. *Archives of Pediatrics & Adolescent Medicine, 161(4)*, 372-377.
- Hus, V., Bishop, S., Gotham, K., Huerta, M., & Lord, C. (2013). Factors influencing scores on the social responsiveness scale. *J Child Psychol Psychiatry, 54(2)*, 216-224. doi:10.1111/j.1469-7610.2012.02589.x
- Hus, V., Pickles, A., Cook, E., Risi, S., & Lord, C. (2007). Using the Autism Diagnostic Interview-Revised to increase phenotypic homogeneity in genetic studies of autism. *Biol Psychiatry, 61*, 438-448.
- Ingram, D., Takahashi, N., & Miles, J. (2008). Defining autism subgroups: a taxometric solution. *J Autism Dev Disord, 38*, 950-960.
- Jewitt, C. (2012). An introduction to using video for research.
- Jones, R.M., Carberry, C., Hamo, A. & Lord, C. (in press). Placebo-like response in absence of treatment in children with autism. *Autism Research*.
- Kaale, A., Smith, L., & Sponheim, E. (2012). A randomized controlled trial of preschool-based joint attention intervention for children with autism. *J Child Psychol Psychiatry, 53(1)*, 97-105. doi:10.1111/j.1469-7610.2011.02450.x
- Kaat, A.J., Gadow, K., & Lecavalier, L. (2013). Psychiatric symptom impairment in children with autism spectrum disorders. *J Abnorm Child Psychol, 41(6)*, 959-69.
- Kaiser, A. P., Hancock, T. B., & Nietfeld, J. P. (2000). The effects of parent-implemented enhanced milieu teaching on the social communication of children who have autism. *Early Education and Development, 11(4)*, 423-446.

- Kanner, L. (1967). Autistic disturbances of affective contact. *Acta paedopsychiatrica*, 35(4), 100-136.
- 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.
- Kasari, C., Gulsrud, A., Paparella, T., Hellemann, G., & Berry, K. (2015). Randomized comparative efficacy study of parent-mediated interventions for toddlers with autism. *Journal of consulting and clinical psychology*, 83(3), 554.
- Kasari, C., Kaiser, A., Goods, K., Nietfeld, J., Mathy, P., Landa, R., ... & Almirall, D. (2014). Communication interventions for minimally verbal children with autism: A sequential multiple assignment randomized trial. *Journal of the American Academy of Child & Adolescent Psychiatry*, 53(6), 635-646.
- 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 Res*, 3(4), 162-173.
- Kitzerow, J., Teufel, K., Wilker, C., & Freitag, C.M. (2015). Using the brief observation of social communication change (BOSCC) to measure autism-specific development. *Autism Research*, doi: 10.1002/aur.1588.
- Koegel, R. L., Shirotova, L., & Koegel, L. K. (2009). Brief report: using individualized orienting cues to facilitate first-word acquisition in non-responders with autism. *Journal of autism and developmental disorders*, 39(11), 1587-1592.
- Lam, K. S., & Aman, M. G. (2007). The Repetitive Behavior Scale-Revised: independent validation in individuals with autism spectrum disorders. *J Autism Dev Disord*, 37(5), 855-866. doi:10.1007/s10803-006-0213-z
- Lam, K. S., Bodfish, J. W., & Piven, J. (2008). Evidence for three subtypes of repetitive behavior in autism that differ in familiarity and association with other symptoms. *Journal of Child Psychology and Psychiatry*, 49(11), 1193-1200.
- Lever, A.G. & Geurts, H.M. (2016). Psychiatric Co-occurring Symptoms and Disorders in Young, Middle-Aged, and Older Adults with Autism Spectrum Disorder. *J Autism Dev Disord*, Epub ahead of print.
- Levy, S., Mandell, D., & Schultz, R. (2009). Autism. *Lancet*, 374, 1627-1638.

- Lopata, C., Thomeer, M.L., Volker, M.A., Toomey, J.A., Nida, R.E., Lee, G.K....Rodgers, J.D. (2010). RCT of a manualized social treatment for high-functioning autism spectrum disorders. *J Autism Dev Disord*, 40(11), 1297-1310.
- Lord, C., Bishop, S., & Anderson, D. (2015). Developmental trajectories as autism phenotypes. *Am J Med Genet C Semin Med Genet*, 169(2), 198-208. doi:10.1002/ajmg.c.31440
- Lord, C., Corsello, C., & Grzadzinski, R. (2014). Diagnostic Instruments in Autistic Spectrum Disorders. In F. Volkmar, S. Rogers, R. Paul, & K. Pelphrey (Eds.), *Handbook of Autism and Pervasive Developmental Disorders* (pp.609-660). Hoboken, New Jersey: John Wiley & Sons, Inc.
- Lord, C., Luyster, R., Guthrie, W., & Pickles, A. (2012). Patterns of developmental trajectories in toddlers with autism spectrum disorder. *J Consult Clin Psychol*, 80(3), 477-489. doi:10.1037/a0027214
- Lord, C., Luyster, R. J., Gotham, K., & Guthrie, W. (2012). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Toddler Module*. Los Angeles, California: Western Psychological Services.
- Lord, C., Risi, S., DiLavore, P. S., Shulman, C., Thurm, A., & Pickles, A. (2006). Autism from 2 to 9 years of age. *Archives of general psychiatry*, 63(6), 694-701.
- Lord, C., Risi, S., Lambrecht, L., Cook, E. H., Leventhal, B. L., DiLavore, P. C., ... & Rutter, M. (2000). The Autism Diagnostic Observation Schedule—Generic: A standard measure of social and communication deficits associated with the spectrum of autism. *Journal of autism and developmental disorders*, 30(3), 205-223.
- Lord, C., Rutter, M., & Le Couteur, A. (1994). Autism diagnostic interview-revised: A revised version of a diagnostic interview for caregivers of individuals with possible pervasive developmental disorders. *J Autism Dev Disord*, 24(5), 659-685. doi:10.1007/BF02172145
- Lord, C., Rutter, M., DiLavore, P. C., Risi, S., Gotham, K., & Bishop, S. L. (2012a). *Autism Diagnostic Observation Schedule, Second Edition (ADOS-2) Modules 1-4*. Los Angeles, California: Western Psychological Services.
- Lord, C., Petkova, E., Hus, V., Gan, W., Lu, F., Martin, D. M., ... & Algermissen, M. (2012b). A multisite study of the clinical diagnosis of different autism spectrum disorders. *Archives of general psychiatry*, 69(3), 306-313.
- Mandy, W. P., Charman, T., & Skuse, D. H. (2012). Testing the construct validity of proposed criteria for DSM-5 autism spectrum disorder. *J Am Acad Child Adolesc Psychiatry*, 51(1), 41-50. doi:10.1016/j.jaac.2011.10.013



- Matson, J. (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*, 207-218.
- Matson, J. L., & Cervantes, P. E. (2014). Commonly studied comorbid psychopathologies among persons with autism spectrum disorder. *Research in Developmental Disabilities, 35*(5), 952-962.
- Matson, J. L., & Shoemaker, M. (2009). Intellectual disability and its relationship to autism spectrum disorders. *Research in Developmental Disabilities, 30*.
- Matson, J. L., Wilkins, J., Sharp, B., Knight, C., Sevin, J. A., & Boisjoli, J. A. (2009). Sensitivity and specificity of the Baby and Infant Screen for Children with aUtism Traits (BISCUIT): Validity and cutoff scores for autism and PDD-NOS in toddlers. *Research in Autism Spectrum Disorders, 3*(4), 924-930.
- McConachie, H., Parr, J., Glod, M., Hanratty, J., Livingstone, N., Oono, I., et al. (2015). Systematic review of tools to measure outcomes for young children with autism spectrum disorder. *Health Technology Assessment, 19*(41).
- Mullen, E. M. (1995). *Mullen scales of early learning*. Circle Pines, MN: American Guidance Service.
- Mundy, P., Block, J., Delgado, C., Pomares, Y., Van Hecke, A. V., & Parlade, M. V. (2007). Individual differences and the development of joint attention in infancy. *Child development, 78*(3), 938-954.
- Murphy, S. A. (2005). An experimental design for the development of adaptive treatment strategies. *Statistics in medicine, 24*(10), 1455-1481.
- Muthen, B., & Muthen, L. (1998-2012). *Mplus User's Guide*. (7th ed.). Los Angeles, CA: Muthen and Muthen.
- Narayanan, S., & Georgiou, P. G. (2013). Behavioral signal processing: Deriving human behavioral informatics from speech and language. *Proceedings of the IEEE, 101*(5), 1203-1233.
- 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*, 19-29.
- Norris, M., Lecavalier, L., & Edwards, M. (2012). The structure of autism symptoms as measured by the autism diagnostic observation schedule. *J Autism Dev Disord, 42*, 1075-1086. 10.1007/s10803-011-1348-0.

- Oosterling, I. J., Wensing, M., Swinkels, S. H., Van Der Gaag, R. J., Visser, J. C., Woudenberg, T., ... & Buitelaar, J. K. (2010). Advancing early detection of autism spectrum disorder by applying an integrated two-stage screening approach. *Journal of Child Psychology and Psychiatry*, *51*(3), 250-258.
- Owley, T., McMahon, W., Cook, E. H., Laulhere, T., South, M., Mays, L. Z., . . . Filipek, P. A. (2001). Multisite, double-blind, placebo-controlled trial of porcine secretin in autism. *J Am Acad Child Adolesc Psychiatry*, *40*(11), 1293-1299. doi:10.1097/00004583-200111000-00009
- Phelps-Terasaki, D., & Phelps-Gunn, T. (2007). *Test of Pragmatic Language*, Second Edition. East Moline, IL: Linguisticsystems.
- Pickles, A., Harris, V., Green, J., Aldred, C., McConachie, H., Slonims, V., . . . Consortium, P. (2015). Treatment mechanism in the MRC preschool autism communication trial: implications for study design and parent-focussed therapy for children. *J Child Psychol Psychiatry*, *56*(2), 162-170.
- Pijl, M. K., Rommelse, N. N., Hendriks, M., De Korte, M. W., Buitelaar, J. K., & Oosterling, I. J. (2016). Does the Brief Observation of Social Communication Change help moving forward in measuring change in early autism intervention studies?. *Autism*, 1362361316669235.
- Reichow, B., & Volkmar, F. R. (2010). Social skills interventions for individuals with autism: evaluation for evidence-based practices within a best evidence synthesis framework. *Journal of autism and developmental disorders*, *40*(2), 149-166.
- Reynolds, C.R. & Kamphaus, R.W. (2006). *BASC-2: Behavior Assessment System for Children*. 2nd ed. Upper Saddle River, NJ: Pearson Education, Inc.
- Reznick, J. S., Baranek, G. T., Reavis, S., Watson, L. R., & Crais, E. R. (2007). A parent-report instrument for identifying one-year-olds at risk for an eventual diagnosis of autism: the first year inventory. *Journal of autism and developmental disorders*, *37*(9), 1691-1710.
- Richler, J., Bishop, S. L., Kleinke, J. R., & Lord, C. (2007). Restricted and repetitive behaviors in young children with autism spectrum disorders. *Journal of autism and developmental disorders*, *37*(1), 73-85.
- Richler, J., Huerta, M., Bishop, S. L., & Lord, C. (2010). Developmental trajectories of restricted and repetitive behaviors and interests in children with autism spectrum disorders. *Development and Psychopathology*, *22*, 55-69.
- Robertson, J. M., Tanguay, P. E., L'Ecuyer, S., Sims, A., & Waltrip, C. (1999). Domains of social communication handicap in autism spectrum disorder. *Journal of the American Academy of Child & Adolescent Psychiatry*, *38*(6), 738-745.

- Robins, D. L., Casagrande, K., Barton, M., Chen, C. M. A., Dumont-Mathieu, T., & Fein, D. (2014). Validation of the modified checklist for autism in toddlers, revised with follow-up (M-CHAT-R/F). *Pediatrics*, *133*(1), 37-45.
- Robinson, E. B., Munir, K., Munafò, M. R., Hughes, M., McCormick, M. C., & Koenen, K. C. (2011). Stability of autistic traits in the general population: further evidence for a continuum of impairment. *Journal of the American Academy of Child & Adolescent Psychiatry*, *50*(4), 376-384.
- Rogers, S., & Vismara, L. (2008). Evidence-based comprehensive treatments for early autism. *Journal of Clinical Child and Adolescent Psychology*, *37*(1), 8-38. doi:10.1080/15374410701817808
- Rogers, S. J., Estes, A., Lord, C., Vismara, L., Winter, J., Fitzpatrick, A., . . . 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. *J Am Acad Child Adolesc Psychiatry*, *51*(10), 1052-1065. doi:10.1016/j.jaac.2012.08.003
- Roid, G. H., & Miller, L. J. (1997). Leiter-R Performance Scale, Revised. *Wood Dale, IL: Stoelting Co.*
- Ronald, A., Larsson, H., Anckarsäter, H., & Lichtenstein, P. (2014). Symptoms of autism and ADHD: A Swedish twin study examining their overlap. *Journal of abnormal psychology*, *123*(2), 440.
- Rutgers, A. H., Bakermans-Kranenburg, M. J., van Ijzendoorn, M. H., & van Berckelaer-Onnes, I. A. (2004). Autism and attachment: a meta-analytic review. *J Child Psychol Psychiatry*, *45*(6), 1123-1134.
- Rutter, M., Baily, A., Lord, C. (2003) *Social Communication Questionnaire*. Los Angeles, CA: Western Psychological Services.
- Rutter, M., & Pickles, A. (2015). Annual Research Review: Threats to the validity of child psychiatry and psychology. *Journal of Child Psychology and Psychiatry*.
- Salazar, F., Baird, G., Chandler, S., Tseng, E., O'sullivan, T., Howlin, P., . . . Simonoff, E. (2015). Co-occurring Psychiatric Disorders in Preschool and Elementary School-Aged Children with Autism Spectrum Disorder. *Journal of Autism and Developmental Disorders*, *45*(8), 2283-94.
- Sallows, G. & Gaupner, T. (2005). Intensive behavioral treatment for children with autism: Four-year outcome and predictors. *American Journal on Mental Retardation*, *110*(6), 417-438.
- Scahill, L., Aman, M. G., Lecavalier, L., Halladay, A. K., Bishop, S. L., Bodfish, J. W., . . . Dawson, G. (2015). Measuring repetitive behaviors as a treatment endpoint in youth with autism spectrum disorder. *Autism*, *19*(1), 38-52. doi:10.1177/1362361313510069

- Schopler, E., Lansing, M. D., Reichler, R. J., & Marcus, L. M. (2005). Examiner's manual of psychoeducational profile (Vol. 3rd). Austin, Texas: Pro-ed Incorporation.
- Schopler, E., Reichler, R. J., & Renner, B. R. (2002). *The childhood autism rating scale (CARS)*. Los Angeles, CA: Western Psychological Services.
- Seltzer, M. M., Krauss, M. W., Shattuck, P. T., Orsmond, G., Swe, A., & Lord, C. (2003). The symptoms of autism spectrum disorders in adolescence and adulthood. *Journal of autism and developmental disorders*, 33(6), 565-581.
- Sheinkopf, S. J., & Siegel, B. (1998). Home-based behavioral treatment of young children with autism. *Journal of autism and developmental disorders*, 28(1), 15-23.
- 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 Res*, 5(4), 267-276. doi:10.1002/aur.1238
- Shuster, J., Perry, A., Bebko, J., & Toplak, M. E. (2014). Review of factor analytic studies examining symptoms of autism spectrum disorders. *J Autism Dev Disord*, 44(1), 90-110. doi:10.1007/s10803-013-1854-3
- Siller, M., Hutman, T., & Sigman, M. (2013). A parent-mediated intervention to increase responsive parental behaviors and child communication in children with ASD: A randomized clinical trial. *Journal of Autism and Developmental Disorders*, 43(3), 540-555.
- Siller, M., & Sigman, M. (2008). Modeling longitudinal change in the language abilities of children with autism: Parent behaviors and child characteristics as predictors of change. *Developmental Psychology*, 44(6), 1691-1704.
- Silverman, J. M., Buxbaum, J. D., Ramoz, N., Schmeidler, J., Reichenberg, A., Hollander, E., ... & Kryzak, L. A. (2008). Autism-related routines and rituals associated with a mitochondrial aspartate/glutamate carrier SLC25A12 polymorphism. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics*, 147(3), 408-410.
- Sim, L. (2006). Effectiveness of a social skills training program with school age children: transition to the clinical setting. *Journal of Child and Family Studies*, 15, 409-418.
- Simonoff, E., Pickles, A., Charman, T., Chandler, S., Loucas, T., & Baird, G. (2008). Psychiatric disorders in children with autism spectrum disorders: prevalence, comorbidity, and associated factors in a population-derived sample. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(8), 921-929.
- Slaughter, V., & Ong, S. S. (2014). Social behaviors increase more when children with ASD are imitated by their mother vs. an unfamiliar adult. *Autism Res*, 7(5), 582-589.

- Smith, T. (2001). Discrete trial training in the treatment of autism. *Focus on autism and other developmental disabilities*, 16(2), 86-92.
- Solomon, M., Ono, M., Timmer, S., & Goodlin-Jones, B. (2008). The effectiveness of parent-child interaction therapy for families of children on the autism spectrum. *J Autism Dev Disord*, 38(9), 1767-76.
- Sparrow, S. S., Cicchetti, D. V., & Balla, D. A. (2005). *Vineland adaptive behavior scales, (Vineland-II)*. Circle Pines, MN: American Guidance Services.
- Sparrow, S. S., Cicchetti, D. V., & Saulnier, C. A. (2016). *Vineland adaptive behavior scales, (Vineland-3)*. Bloomington, MN: Pearson.
- Spence, S. & Thurm, A. (2010). Testing autism interventions: trials and tribulations. *Lancet*, 375, 2124-2125.
- Spreckley, M. & Boyd, R. (2009). Efficacy of applied behavioral intervention in preschool children with autism for improving cognitive, language, and adaptive behavior: a systematic review and meta-analysis. *Journal of Pediatrics*, 154, 338-344.
- Sprenger, L., Bühler, E., Poustka, L., Bach, C., Heinzl-Gutenbrunner, M., Kamp-Becker, I., & Bachmann, C. (2013). Impact of ADHD symptoms on autism spectrum disorder symptom severity. *Research in developmental disabilities*, 34(10), 3545-3552.
- Spiker, D., Lotspeich, L., Dimiceli, S., Myers, R., & Risch, N. (2002). Behavioral phenotypic variation in autism multiplex families: evidence for a continuous severity gradient. *Am J Med Genet*, 114, 129-136.
- Tager-Flusberg, H., & Kasari, C. (2013). Minimally verbal school-aged children with autism spectrum disorder: the neglected end of the spectrum. *Autism Research*, 6(6), 468-478.
- Tager-Flusberg, H., Paul, R., Lord, C. (2001). Language and communication in autism. In *Handbook of autism and pervasive developmental disorder: Vol. 1*. 3rd edition. Edited by Volkmar F, Paul R, Klin A, Cohen DJ. New York: Wiley; 2001:335–364.
- Thurm, A., Manwaring, S. S., Swineford, L., & Farmer, C. (2015). Longitudinal study of symptom severity and language in minimally verbal children with autism. *J Child Psychol Psychiatry*, 56(1), 97-104. doi:10.1111/jcpp.12285
- van Lang, N. D., Boomsma, A., Sytema, S., de Bildt, A. A., Kraijer, D. W., Ketelaars, C., & Minderaa, R. B. (2006). Structural equation analysis of a hypothesised symptom model in the autism spectrum. *J Child Psychol Psychiatry*, 47(1), 37-44.
- Wallace, G.L., Budgett, J., Charlton, R.A. (2016). Aging and autism spectrum disorder: Evidence from the broad autism phenotype. *Autism Research*, Epub ahead of print.

- Warren, Z., McPheeters, M. L., Sathe, N., Foss-Feig, J. H., Glasser, A., & Veenstra-VanderWeele, J. (2011). A systematic review of early intensive intervention for autism spectrum disorders. *Pediatrics*, *127*(5), e1303-e1311.
- Wetherby, A. (2008). Understanding and measuring social communication in children with autism spectrum disorders. *Social and communication development in autism spectrum disorders: Early identification, diagnosis, and intervention*, *3*.
- Wetherby, A. M., Guthrie, W., Woods, J., Schatschneider, C., Holland, R. D., Morgan, L., & Lord, C. (2014). Parent-implemented social intervention for toddlers with autism: an RCT. *Pediatrics*, *134*(6), 1084-1093. doi:10.1542/peds.2014-0757
- Wetherby, A. & Prizant, B. (2002). *Communication and Symbolic Behavior Scales Developmental Profile (First Normed Edition)*. Baltimore, MD: Paul H. Brooks.
- Wing, L., Leekam, S.R., Libby, S.J., Gould, J., & Larcombe, M. (2002). The Diagnostic Interview for Social and Communication Disorders: Background, inter-rater reliability, and clinical use. *Journal of Child Psychology and Psychiatry and Allied Disciplines*, *43*(3), 307-325.
- Wolery, M. & Garfinkle, A. (2002). Measures in Intervention Research with Young Children who have Autism. *Journal of Autism and Developmental Disorders*, *32*(5), 463-478.
- Wong, C., Odom, S. L., Hume, K. A., Cox, A. W., Fetting, A., Kucharczyk, S., ... & Schultz, T. R. (2015). Evidence-based practices for children, youth, and young adults with autism spectrum disorder: A comprehensive review. *Journal of autism and developmental disorders*, *45*(7), 1951-1966.
- Woolfenden, S., Sarkozy, V., Ridley, G., & Williams, K. (2012). A systematic review of the diagnostic stability of Autism Spectrum Disorder. *Research in Autism Spectrum Disorders*, *6*(1), 345-354.
- Yoder, P., 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. *Evid Based Commun Assess Interv*, *7*(4), 150-174.
- 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. doi: 10.1352/1944-7558-119.1.17.
- Zachor, D. A., & Itzhak, E. B. (2010). Treatment approach, autism severity and intervention outcomes in young children. *Research in Autism Spectrum Disorders*, *4*(3), 425-432.

Zhou, T., & Yi, C. (2014). Parenting styles and parents' perspectives on how their own emotions affect the functioning of children with autism spectrum disorders. *Fam Process*, 53(1), 67-79.

APPENDIX

IRB Approvals.



*Teachers College IRB*

*Exempt Study Approval*

To: Rebecca Grzadzinski  
From: Curt Naser, TC IRB Coordinator  
Subject: IRB Approval: 16-354 Protocol  
Date: 05/30/2016

Thank you for submitting your study entitled, "*Measuring Changes in Social Communication Behaviors*;" the IRB has determined that your study is **Exempt** from committee review (Category **4**) on 05/30/2016.

Please keep in mind that the IRB Committee must be contacted if there are any changes to your research protocol. The number assigned to your protocol is **16-354**. Feel free to contact the IRB Office by using the "Messages" option in the electronic Mentor IRB system if you have any questions about this protocol.

You can retrieve a PDF copy of this approval letter from the Mentor site.

Best wishes for your research work.

Sincerely,  
Curt Naser, Ph.D.  
TC IRB Coordinator  
curtn@axiomeducation.com





Weill Cornell Medical College

---

Institutional Review Board

Mailing Address:  
1300 York Avenue  
Box 89  
New York, NY 10065

Telephone: 646-962-8200  
E-mail: [irb@med.cornell.edu](mailto:irb@med.cornell.edu)

September 10, 2015

Catherine Lord, MD

Submission Type: Expedited Continuing Review  
Protocol Number: 1205012407R003  
Protocol Title: Developing a sensitive, cost effective measure of  
behavior change in children with ASD.(BOSCC)  
Status of IRB Protocol: Open  
Risk Level: Minimal Risk  
Expedited Category: 5

Dear Dr. Lord:

The renewal for the abovementioned protocol was reviewed and approved by a member of the Institutional Review Board via expedited review procedures as per 45CFR46.110.

The protocol and its relevant documents stand approved for the following period:

**Approved:** September 6, 2015

**Expires:** September 5, 2016

Please do not hesitate to contact the IRB office staff if you have any questions or need assistance in complying with the terms of this approval.

Sincerely,

A handwritten signature in cursive script that reads "Rosemary Kraemer".

Rosemary Kraemer, Ph.D.  
Director, Human Research Protections Program



## Weill Cornell Medical College

---

### Institutional Review Board

Mailing Address:  
1300 York Avenue  
Box 89  
New York, NY 10065

Telephone: 646-962-8200  
E-mail: [irb@med.cornell.edu](mailto:irb@med.cornell.edu)

August 5, 2015

Catherine Lord, MD

Submission Type:	Expedited Continuing Review with Amendment Response to Modifications Required
Protocol Number:	1206012570R003
Protocol Title:	Adaptive Interventions for Minimally Verbal Children with ASD in the Community
Status of the protocol:	Open
Risk Level:	Minimal Risk
Pediatric Risk Determination:	45 CFR 46.404
Expedited Category:	9
Nature of Amendment:	We would like to add Rebecca Gradzinski to our protocol as co-investigator. We have updated our oral consent script to represent accurate contact information for the study outreach coordinator. We have revised four of our recruitment documents to represent updated contact information for the study outreach coordinator: recruitment flyer, brochure, parent presentation; and the recruitment school letter. We also changed the parent child filmed play assessment from 10 to 15 minutes on one of the slides in the English Parent Presentation. We are attaching 2 new letters of approval for schools which granted us access to work with their students in the classroom. We changed the address on our consent forms to reflect the correct IRB address and the date the forms were revised.

Dear Dr. Lord:

The Institutional Review Board has conducted a review of your response to the modifications required letter issued on 7/30/2015 regarding the above mentioned protocol. The renewal and amendment for the abovementioned protocol was approved via expedited review procedures as per 45CFR46.110.

The protocol and its relevant documents stand approved for the following period:

- Consent Form for Clinical Investigation (English and Spanish) – revised 7/23/2015
- HIPAA Authorization Form (English and Spanish) – revised 7/23/2015
- Phone Screening (English and Spanish)
- AIM-ASD DSMB Report – 12/16/2014
- Adverse Event & IND Table – 6/26/2015
- Jawonio and Devereux Site Approval Letters – 12/27/2013 and 5/13/2015
- Recruitment Tools: Flyer (English and Spanish), AIM Brochure (English and Spanish), District Presentation, Parent Presentation (English and Spanish), Treatments & Research Evidence, School Letter


- Assessments: Clinical Global Impressions, Structured Play, Seizure Interview, Parent Ranking of Therapy, Parent Perceptions of Therapy, Parent Expectancies of Therapy, Language Sample, Examiner-Child Interaction, Caregiver-Child

Approved: 8/5/2015

Expires: 8/4/2016

Please do not hesitate to contact the IRB office staff if you have any questions or need assistance in complying with the terms of this approval.

Sincerely,



Rosemary Kraemer, Ph.D.  
Director, Human Research Protections Program

Please note the following important information about this approval:

- **Billing Compliance:** This approval is contingent upon continued adherence with institutional billing compliance policies.
- **Immediate Reporting:** Investigators must follow the Immediate Reporting Policy at [http://weill.cornell.edu/research/research\\_integrity/institutional\\_review\\_board/irb\\_adv.html](http://weill.cornell.edu/research/research_integrity/institutional_review_board/irb_adv.html)
- Failure to comply with IRB directives within specified time frames may result in federally mandated penalties, up to and including suspension or termination of IRB approval and mandatory reporting to the Federal government.
- **Human Gene Transfer:** If this is a human gene transfer protocol, it is a term and condition of IRB approval that the principal investigator obtains Institutional Biosafety Committee (IBC) approval of all amendments prior to initiation, reportable adverse events as per WCMC policy, and annual reports as per M-I-C-3 of the NIH Guidelines for Research Involving Recombinant DNA Molecules. View the IBC website at [http://weill.cornell.edu/research/research\\_integrity/ibc.html](http://weill.cornell.edu/research/research_integrity/ibc.html) or contact [ibc@med.cornell.edu](mailto:ibc@med.cornell.edu) if you require assistance in complying with these requirements.
- **Other reporting:** The reporting requirements of various regulatory bodies may differ with regard to both what must be reported and when. You are responsible for acquainting yourself with and abiding by all applicable federal and state regulatory reporting requirements.
- **Changes to this protocol:** If you want to change this research in any way or if any unanticipated hazardous conditions emerge affecting the rights or welfare of the human subjects involved in it, you must submit an amendment detailing these changes to the IRB for review and approval prior to implementing those changes. If the CTSC is used, the changes must also be submitted to the Translational Research Advisory Committee (TRAC). It is your responsibility to obtain approval for any such changes prior to initiating them.
- **Continuing approval:** You will receive a reminder via email for continuing review of this protocol in advance of the expiration date. The continuing review forms must be filed with the IRB sufficiently early to permit timely review and approval if the project is to continue beyond the period for which it was approved. Please note, no study related activities can continue beyond the WCMC IRB expiration date, including subject recruitment, enrollment, intervention and data analysis.
- **If your research study involves human tissues:** In addition to IRB approval, Section 4.4 of the hospital By-Laws "Specimens Removed During Resective Surgery" requires that all specimens removed during surgical diagnostic procedures that will be used for research must be approved by Pathology Service. Information about Pathology review can be found online at [http://www.med.cornell.edu/research/for\\_pol/forms/Pathology\\_Review\\_Instructions.pdf](http://www.med.cornell.edu/research/for_pol/forms/Pathology_Review_Instructions.pdf)
- **If the IRB is requiring that you obtain informed consent from subjects:** The signed IRB approved consent forms must be kept in the subject's hospital chart. If the subject has no New York Presbyterian Hospital chart, you are responsible for retaining such signed forms in your research files.
- **Information about the WCMC IRBs:** The Weill Cornell Medical College (WCMC) Institutional Review Board (IRB) is constituted as required by the Federal Office for Human Research Protections (OHRP). WCMC holds a Federalwide Assurance (FWA) with OHRP. The FWA number is FWA00000093. The WCMC IRB is registered on that FWA. The registration number for the IRB is: General IRB #1 IRB00009417, General IRB #2 IRB00009418, Cancer IRB#1 IRB00009420, Cancer IRB#2 IRB00009421 and Expedited IRB IRB00009419. Should you need additional information about the terms of the WCMC FWA or the WCMC IRBs, please refer to [http://weill.cornell.edu/research/research\\_integrity/institutional\\_review\\_board/index.html](http://weill.cornell.edu/research/research_integrity/institutional_review_board/index.html).