ABSTRACT

MCDONOUGH II, STEPHEN GERARD. Construct Validity in the Measurement of Autism Symptoms Across Raters. (Under the direction of Dr. Scott A. Stage.)

Measures of autism spectrum disorders (ASDs) provide critical information relevant to clinical, educational, and home settings, and evaluate the presence and severity of symptoms/traits (i.e., social, communication [SC], and restrictive repetitive behaviors [RRB]). As a matter of practice, such measures can confirm diagnosis, guide intervention selection, and evaluate symptom severity over time. Diagnosis of ASD requires integration of multiple reporters. However, to date, knowledge regarding the construct validity for ASD traits are rather limited. By using a multitrait-multirater (MTMR) matrix, this study evaluated the construct validity for RRB and SC related to the measurement of ASD. Comparisons of ratings from three reporters (i.e., caregiver, teacher, and independent observer) across two constructs (i.e., RRB and SC) were evaluated for construct validity. The measure of constructs included the Repetitive Behavior Scale Revised (RBS-R), the Social Responsiveness Scale (SRS), and the Autism Diagnostic Observation Scale (ADOS). The reports were based on a population of pre-school aged children with ASD. This study tested several hypotheses (a) if convergent validity (CV) coefficients were significantly larger than the discriminant validity (DV) coefficients of different raters, (b) if CV coefficients were significantly larger than DV coefficients of same raters and, (c) that DV coefficients of same raters were larger than DV coefficients of different raters. Results indicated that the tested traits by raters failed to show construct validity as described by Campbell and Fiske (1959) and that rater bias significantly accounted for the largest degree of association.

Keywords: Autism Spectrum Disorder, Asperger's Disorder, restrictive repetitive behaviors, social communication, motor stereotypies, preschool.

Construct Validity in the Measurement of Autism Symptoms Across Raters

by
Stephen Gerard McDonough II

A thesis submitted to the Graduate Faculty of
North Carolina State University
in partial fulfillment of the
requirements for the Degree of
Master of Science

Psychology

Raleigh, North Carolina

2013

APPROVED BY:

Mary E. Haskett

William P. Erchul

Scott A. Stage,
Chair of Advisory Committee

BIOGRAPHY

Stephen Gerard McDonough II was born on December 31, 1977 in Stoneham, Massachusetts. After attending St. John's Preparatory School in Danvers, Massachusetts, he attended St. Anselm College in Manchester, New Hampshire where he majored in Psychology. Stephen received his Bachelor of Arts degree in 2000 and began working at the North Shore Medical Centers' adolescent and pediatric psychiatric facility. Following this experience, Stephen attended Tufts University in Medford, Massachusetts where he earned a Master of Arts degree in Child Development with a concentration in clinical developmental psychology in 2004. Following the completion of this degree, Stephen accepted a job with the Academy North of Reading, Massachusetts where he worked closely with a team of psychologists who provided school based consultation, social skills groups, and therapeutic summer camps. During his time at the Academy North, Stephen developed many skills and received recognition for his ability to work with children with significant emotional and behavioral problems including children and adolescents on the autism spectrum. In 2007, Stephen moved to North Carolina where he worked as a research assistant at Frank Porter Graham Child Development Institute at the University of North Carolina - Chapel Hill. Stephen's work at UNC focused on early interventions for the treatment of autism. Upon completing his Masters of Science degree, Stephen intends to continue his studies in School Psychology at North Carolina State University until obtaining his doctoral degree.

ACKNOWLEDGEMENTS

I would like to thank my advisor Dr. Scott Stage for his direction, patience, and support as well as my committee members, Dr. William Erchul and Mary E. Haskett. I owe a great debt to many people at UNC at Chapel Hill who provided me the opportunity and professional mentorship to work on the research project, especially Dr. Brian Boyd, Dr. Kara Hume, and Dr. Sam Odom. I would also thank my parents Nancy and Steve as well as my sister Megan, without their support, I would not have felt as capable and competent.

TABLE OF CONTENTS

LIST OF TABLES vii

LIST OF FIGURES viii

Introduction 1

ASD Diagnostic Measurement 4

Criterion measure 6

Using multiple raters 7

Measuring Construct Validity 9

Restrictive Repetitive Behaviors 9

Clinical criteria 9

Typical and atypical development 10

Heterogeneity of repetitive behaviors 11

Measures of RRB 12

Measures of RRB rated by the caregiver 12

Measures of RRB rated by the teacher 14

Summary of RRB rated measures 15

Social and Communication Deficits 16

Clinical criteria 16

Typical and atypical development 17

Measures of social and communication deficits 18

Measures of SC rated by the caregiver 18

Measures of SC rated by teachers 19

Summary of SC rated measures 19

Purpose of the Study 20

Research Design 20

Research Hypotheses 23

Hypothesis 1 23

Hypothesis 2 23

Hypothesis 3 24

Method 24

Sample and Participant Selection 24

Sampling procedures 25

Participants 25

Materials and Procedures 26

Autism Diagnostic Observation Schedule (ADOS; 1989) 28

Repetitive Behavior Scale-Revised (RBS-R; 2000) 28

Social Responsiveness Scale (SRS; 2005) 29

Results 30

Descriptive Statistics 30

Multitrait Multirater Matrix 33

Reliability diagonal 35

Convergent validity coefficients 36

Discriminant validity coefficients 36

Heterotrait - heterorater coefficients 36

Heterotrait - monorater coefficients 37

Test of magnitude 37

Wilcoxon Signed-Rank Test 38

Convergent validity vs. heterotrait - heterorater 38

Convergent validity vs. heterotrait - monorater 40

Heterotrait - monorater vs. heterotrait - heterorater 40

Discussion 41

Limitations of the Present Study 46

Conclusions and Future Perspectives 48

References 50

LIST OF TABLES

Table 1 *Child and Teacher Demographic Data* 26

Table 2 *Descriptive Statistics of the Autism Diagnostic Observation Schedule, the Repetitive Behavior Scale-Revised, and the Social Responsiveness Scale* 31

Table 3 *Skew, Kurtosis, and Z-score Tests* 32

Table 4 *Multitrait-Multirater Matrix of Autism Symptoms by Raters* 34

Table 5 *Descriptive Statistics for the Wilcoxon Signed Ranks Tests* 38

Table 6 *Wilcoxon Signed Ranks Test* 39

LIST OF FIGURES

Figure 1. Example of a Multitrait-multirater matrix of ASD 21

Introduction

Autism is a pervasive developmental disorder characterized by atypical social and communication abilities and the presence of a variety of restricted repetitive behaviors (RRB), interests, and activities. The *Diagnostic and Statistical Manual of Mental Disorders* (4th ed., text rev.; *DSM-IV-TR*; American Psychiatric Association, 2000) requires that three criteria be met for the diagnosis of Autism Disorder. The first criterion involves the presence of at least six symptoms across the three atypical domains (i.e., impairments in social interaction, impairment in communications, and presence of RRB). The second criterion requires the onset of at least one symptom occurring prior to the age of 3. The third criterion requires that Rett's Disorder or Childhood Disintegrative Disorder does not better explain the symptoms. Although diagnostically similar to Autism, Asperger's Disorder requires the same social and RRB symptoms, but requires only three symptoms across the two domains of social interaction and RRB (4th ed., text rev.; *DSM–IV–*TR; American Psychiatric Association, 2000). Another similarly defined diagnosis is Pervasive Developmental Disorder – not otherwise specified (PDD-NOS), which is a diagnosis given when the severity and persistence of symptoms significantly impair functioning but fail to meet either of the above diagnostic criteria. Together, Autistic Disorder, Asperger's Disorder, and PDD-NOS make up the term Autism Spectrum Disorders (ASD), commonly used in the research literature, medical practice, and the community (Lord, 2010).

Proposed changes to the fifth edition of the *Diagnostic and Statistical Manual of Mental Disorders* (5th ed.; *DSM-V;* American Psychiatric Association, 2011) have re-

formalized ASD diagnoses in place of the disorders described (i.e., Autistic Disorder, Asperger's Disorder, PDD-NOS and Childhood Disintegrative Disorder). An ASD diagnosis requires that four criteria be met (a) deficits in social communication or social interaction (SC), (b) presence of RRB, (c) symptoms present in early childhood, and (d) symptoms together limit and impair everyday functioning. Therefore, the proposed ASD model requires establishing the presence of two separate constructs SC and RRB. Whereas, the third and fourth criteria are qualifiers applied to these constructs and consequently subsumed in the measure of SC and RRB.

The first criterion includes collapsing the social and communication deficit domains into one broad criterion, reciprocal social-communication-interaction (American Psychiatric Association, 2011). Meeting the proposed SC criterion requires the presence of deficits in the following areas: social-emotional reciprocity, nonverbal communicative behaviors, and development and maintenance of relationships (American Psychiatric Association). This is consistent with multiple studies that have demonstrated that social and communication domains are not differentiated (Boomsma et al., 2008; Constantino et al., 2004; Robertson, Tanguay, L'Ecuyer, Sims, & Waltrip, 1999; Snow, Lecavalier, & Houts, 2009). The second criterion focuses on RRB and requires establishing the presence of two out of four possible manifestations of RRB, previously one out of four. The categories of RRB manifestations remain relatively similar with fixated interests and persistent preoccupation with parts of objects combining to form one category. The second and third categories, stereotyped/repetitive behaviors and adherence to routines, now include behaviors that

involve language (i.e., stereotyped speech). The fourth proposed category for RRB is new and requires the presence of atypical hyper- or hypo-reactivity to sensory input or unusual interest in sensory aspects of the environment (American Psychiatric Association).

The proposed ASD diagnostic criteria would also include three degrees of severity that would reflect the degree to which symptoms affect daily functioning (Lord, 2010; Schneider, 2010). The symptom severity category would effectively indicate the level of intervention required for the child. Assessing symptom severity would also require developmental appropriateness. For example, a 30-month-old child's engaging in occasional arm flapping may present as mildly atypical or typical, given that the behavior is unlikely to present any significant impairment. However, a 13-year-old child engaging in the same occasional arm flapping is more atypical, when compared to a typical group of same aged peers, and therefore, more severe. In light of proposed changes in the diagnostic criteria, there likely will be an increased importance placed on valid measures that aid diagnosis, especially measures that are applicable across settings and can be used by multiple raters.

The proposed changes to the *DSM-V* (American Psychiatric Association, 2011) diagnosis/construct of ASD serves as a reminder to ensure that measures of ASD possess construct validity. Therefore, this study assessed the construct validity of commonly used measures of ASD traits across raters using the proposed diagnostic criteria. The study also assessed the assumption that validated measures of ASD will hold together. Given that ASD is a spectrum disorder, the measures ought to assess a comprehensive range of symptom (i.e.,

trait) expression if they are truly measuring the construct. Finally, the study aimed to provide a psychometric heuristic relative to the clinical application of multirater assessment for ASD.

The following literature review will address some measures used for diagnosis and the importance of using multiple raters. This review also will explain the two symptom areas of ASD in detail and measurement including the criterion measure of ASD. The first symptom area described is RRB and this description will address the following: (a) clinical definition, (b) descriptions of typical and atypical development, (c) heterogeneity of behaviors, (d) trait-based measures of RRB by rater, and (e) a summary of presented measures. The second symptom area, SC, will describe: (a) the clinical definition, (b) descriptions of typical and atypical development, (c) trait-based measures of SC symptoms by rater, and (d) a summary of presented measures. Following the symptom descriptions, the rationale explains the use of a multitrait-multirater (MTMR) matrix containing measures of symptoms across multiple raters (i.e., caregiver, teacher, and independent observer, or formally trained evaluator of ASD).

ASD Diagnostic Measurement

According to the National Institute of Mental Health (NIMH; 2008), an ASD diagnosis follows a two-stage process of developmental screening and comprehensive evaluation. Elevated scores on screenings result in comprehensive evaluations, which include measures expressly designed for the detection of ASD (NIMH, 2008). The best evidence for a comprehensive diagnosis includes the use of a caregiver report, standardized ASD measures, and direct observation (Allen, Robins, & Decker, 2008; Warren & Stone, 2011).

Generally, two standardized measures are used by the clinician: the *Autism Diagnosis Interview-Revised* (ADI-R) and the *Autism Diagnostic Observation Schedule-General* (ADOS; NIMH, 2008; Gotham, Pickles, & Lord, 2009). The ADI-R is a semi-structured interview conducted with caregiver, composed of 111 questions that address 3 domains: social, communication, and restrictive repetitive behaviors (Lord, Rutter, & LeCouteur, 1994). The ADOS is a semi-structured, standardized observational assessment of communication, social interaction, and play. It is separated into different modules based on language and developmental level, which is conducted with the individual (Lord et al., 1994). The ADOS is quite useful in diagnosis given it is a standardized measure and provides direct observation of the child. Given the importance of clinical judgment, the ADOS has clear advantages over the ADI-R. The ADOS requires extensive training for an assessor prior to him/her being considered proficient and is principally based on the clinician's interpretation of observed behavior. However, the ADOS or ADI-R alone is not sufficient to establish a diagnosis; the clinician must also obtain information regarding the child across different settings.

A diagnosis of ASD results from a clinician establishing the presence of a specific pattern of behaviors. In order to achieve this, the clinician acquires information from multiple sources and direct observation. The patterns of behaviors for ASD are theoretically linked to neurodevelopmental impairments, and these patterns constitute the construct or diagnostic criteria presented in the *DSM-IV-TR* (2000) and *DSM-V* (American Psychiatric Association,

2011). As noted above, the ASD diagnostic constructs are expected to change over time based on scientific evidence collected on assessment.

According to Cronbach and Meehl (1955), every psychological construct has a "nomological network," which is the explanation of constructs and variables and how they relate. The network describes observable properties of theoretical constructs in relationship to other observable properties of different theoretical constructs (Cronbach & Meehl). Essentially, the nomological network is the application of logic to a theoretical construct and provides a structure for measuring construct validity that requires a test that incorporates observations and exhibits clear steps in interpreting procedures.

Criterion measure. According to Cronbach and Meehl (1955), criterion validity can be established by the research community with a measure that has consensus as the standard that is predominantly accepted. The ADOS by design is that criterion-referenced assessment, which is based on the *DSM-IV-TR* (2000) criteria. The ADOS is widely accepted as the "gold standard" of diagnostic assessment (Luyster et al., 2009; Reaven, Hepburn, & Ross, 2008). The research community has long considered the ADOS as a criterion measure due in part to the high standards of training and inter-rater reliability required for researchers set by the publisher (Western Psychological Services, 2011). The most recent revision to the ADOS is reformulated into primary constructs based on a two-factor model that includes SC and RRB with both factors uniquely contributing to an ASD diagnosis (Gotham et al., 2008).

Traditional validity evidence to support the ADOS is present in the literature. However, the validity evidence requires some caution in that the ADOS in now being

interpreted with the two-factor model. Inter-rater reliability and internal consistency of the ADOS three-factor model (i.e., social, communication, and RRB) are strong using the previous algorithm (Lord, Risi et. al., 2000). Given that the two-factor model did not change the content or order of the items, it is likely these findings remain consistent. Kim and Lord (2010) examined RRB observed during the ADOS using a longitudinal study of toddler and preschool children ($N = 455$). This sample included children with ASD, PDD-NOS, and other developmental delays, as well as typically developing children. Kim and Lord addressed discriminant validity with regard to RRB and differentiated children with ASD from other developmental delays and typically developing children based on the severity and prevalence of RRB observed during the ADOS. The researchers noted the importance of using clinical observation over structured interviews with younger children to establish the presence of RRB. This is a strength of the ADOS and a limitation of the ADI-R because caregivers may not have critically examined RRB in younger children. Chawarska, Klin, Paul, and Volkmar (2007) found that parent reports of RRB were not consistent (i.e., not identifying RRB when present) with clinical observations of RRB in a population of children under two years of age. Despite the extensive use of the ADOS in research and the evidence of validity, it lacks evidence of construct validity. Campbell and Fiske (1959) suggested, the best measure of construct validity is through the use of a MTMR matrix when a significant source of variance can be expected due to multiple rater requirements

Using multiple raters. The ADOS alone cannot provide sufficient evidence to support a differential diagnosis because it fails to ascertain the severity or onset of symptoms.

Neurodevelopment disorders, such as ASD, will present across settings and to establish that they are present requires information from multiple sources. This practice increases the information available for a clinician to make a more informed decision (NIMH, 2008). For a trait such as RRB, a multi-rater assessment increases the accuracy of diagnosis by assessing the presence and severity of symptoms across settings (e.g., home and school) using various measures (e.g., ADOS, Social Responsiveness Scale [SRS], or Repetitive Behavior Scale Revised [RBS-R]). This is particularly true for diagnosis in children under 3. Stone et al. (1999) found strong evidence for the reliability and stability of an ASD diagnosis given prior to 3, using multiple clinicians during two consecutive annual evaluations. However, clinicians varied greatly in their ability to distinguish specific aspects of the ASD diagnosis (i.e., autism vs. PDD-NOS) and the authors suggested clinical experience and parental underreporting may contribute to this change over time (Stone et al., 1999).

Families present as one of the best sources to establish frequency and duration of RRB because they observe the child over long periods of time and across multiple settings. However, emerging or established behaviors may missed due to families' accommodation to the RRB or not having a normative reference for comparison. Caregivers have underreported symptoms of RRB in toddler samples when compared to clinical observations (Chawarska et al., 2007; Kim & Lord, 2010). In contrast to caregivers, teachers have access to a comparative sample of peers and experience working with many different children, which is likely to aid in identifying behaviors that parents may have missed. Therefore, it is also important that clinicians be aware of their own and each informant's knowledge of ASD. In

summary, an ASD diagnosis requires multiple informants to establish the presence of various symptoms, many of which may not be obvious to an untrained observer.

Measuring Construct Validity

Construct validity refers to how well an instrument or scale measures the theoretical construct it claims to measure. According to Campbell and Fiske (1959), the best method to establish construct validity is through the use of a MTMR matrix for constructs heavily reliant on multiple raters. The MTMR matrix procedure has humbled many a psychological construct (Campbell & Fiske, 1959). The constructs of ASD are the two symptom areas RRB and SC and measures of ASD aim to establish the presence of RRB and/or deficits in SC. The clinician compiling and interpreting diagnostic data determines an ASD diagnosis with data that include multiple measures and multiple raters. Therefore, establishing construct validity for ASD measures is better achieved through the use of a MTMR matrix. To date, there is no clear evidence of construct validity using multiple raters for ASD. Prior to an explanation of the MTMR matrix, the ASD constructs are discussed.

Restrictive Repetitive Behaviors

Evidence of several subtypes of RRB has emerged from the literature (Bodfish, 2011; Bodfish, Symons, Parker, & Lewis, 2000). These behaviors can change over time, making their evaluation essential to accurate clinical diagnosis and for determining ASD severity.

Clinical criteria. The *DSM-IV-TR* (2000) defines the RRB domain as "restricted repetitive and stereotyped patterns of behavior, interests, and activities" (pp. 75 & 84). The criteria for RRB are met when one of the following is exhibited: (a) an encompassing

preoccupation or interest that has an abnormal focus or intensity, (b) non-functional routines or rituals that are specific and followed with inflexible adherence, (c) repetitive movements, and (d) preoccupation with objects or parts that is persistent.

Typical and atypical development. Identification of RRB early in development is difficult given that some repetitive behaviors, specifically stereotypy, is a normal aspect of maturation prior to the age of three (Evans et al., 1997; Mahone, Bridges, Prahme, & Singer, 2004; Thelen, 1979; Werry, Carlielle, & Fitzpatrick, 1983; Zohar & Felz, 2001). The term stereotypy refers to movements that are repetitive and have no clear function or aim. For example, in the case of an infant rocking, the movement is repetitive and determining a cause or purpose of the behavior is difficult. Berkson and Tupa (2000) reviewed 50 years of research literature on early stereotypy development and identified two probable causes or functions of stereotypy: either self-stimulation or motor expression. Research in this area suggests that these behaviors start in early infancy and increase over the first two years and then decrease as locomotion develops (Evans et al., 1997; Thelen, 1979; Wehmeyer, 1991; Werry et al., 1983). Therefore, the optimal time for early identification of problematic or atypical RRB coincides with the entry into preschool.

Researchers hypothesize that the RRB of children with ASD likely serve many functions (Turner, 1999). Across the research, there is support for four categories of RRB functions, including: (a) biological, (b) behavioral, (c) homeostatic, and (d) sensory (Goodwin, Intille, Albinali, & Velicer, 2010; Turner, 1999). Biological theories suggest that RRB results from the development of atypical neurological systems, including the basal

ganglia (Canales & Graybiel, 2000), caudate nucleus (Turner, Frost, Linsenbardt, McIlroy, & Muller 2006), and dopaminergic pathways (Bodfish et al., 1995). Operant behavioral theory suggests that RRB is reinforced (Iwata, Dorsey, Slifer, Bauman, & Richman, 1982; Lovaas, Newsome, & Hickman, 1987; Vollmer, Marcus, LeBlanc, 1994). Homeostatic theories propose RRB serve to balance states of arousal, and initiation occurs when the individual is in a state of hyper-arousal (Hutt & Hutt, 1970; Zentall & Zentall, 1983). Theories of sensory functions of RRB offer that the individual engages in RRB to change auditory, visual, tactile, vestibular, and proprioceptive stimulation that result from one of three sensory patterns: hyperresponsiveness, hyporesponsiveness, or sensory seeking (Boyd et al., 2010).

Heterogeneity of repetitive behaviors. The majority of studies of RRB subtypes have identified two factors, "lower-order" and "higher-order" behaviors (Bishop, Richler, & Lord, 2006; Cucarro et al., 2003; Richler, Bishop, Kleinke, & Lord, 2007; Richler, Huerta, Bishop, & Lord, 2010; Szatmari et al., 2006). Lower-order behaviors consist of repetitive sensory and motor behaviors (i.e., self-injurious behavior, stereotyped movements, motor-stereotypies, and repetitive object use). Higher-order behaviors are more complex and cognitive (i.e., rituals and routines, compulsions, and insistence on sameness). Evidence of a third factor is also present in the literature. Honey, Leekam, Turner, and McConachie (2007) found support for circumscribed interest as an independent factor for children ages 2 through 4 with autism. Lam, Bodfish, and Piven (2008) found evidence to support an "interest / preoccupation / attachment" factor (i.e., odd fixation on odd topics, intense focused hobbies, or strong attachment to objects). They also found that 88% of their sample ($N = 316$)

revealed some circumscribed interest, suggesting that this is a common symptom in autism, unlike other RRB (e.g., motor stereotypic movements) that are observable in other psychopathological disorders (Lam et al., 2008). Therefore, establishing the presence of a RRB requires screening for a range of subtypes. Based on factor analysis of RRB measures, there appear to be five RRB subtypes: ritualistic/sameness, self-injurious, stereotypic, compulsive, and restricted interests (Lam, 2004; Lam & Aman, 2007; Mirenda et al., 2010).

Measures of RRB. Measures of RRB include paper-and-pencil rating scales, diagnostic observations, and clinical interviews. The design of these approaches often emanates from the rater doing the assessment (e.g., caregiver, teacher, or clinician). Measures designed for the clinician rely on observation of the child and involve presses (i.e., an attempt to elicit a behavior) for specific behaviors (e.g., ADOS), or interview of the caregiver (e.g., ADI-R). Approaches designed for caregivers and teachers include paper and pencil measures that screen for the presence or absence of RRB. Measures designed for caregivers and teachers often share similar or identical questions and differ in how questions factor into scores for each group. For example, teachers do not likely observe routines related to bedtime. Most approaches require the simple identification of a specific behavior (e.g., the child engages in self-injury), whereas other measures require greater specificity with endorsement of frequency, duration, latency, and intensity of the given behavior.

Measures of RRB rated by the caregiver. Few validated rating measures look exclusively at RRB within an ASD population (Lam & Aman, 2007). Measures of RRB based on caregiver report often include a semi-structured parent interview, such as the ADI-R

and the Children's Yale-Brown Obsessive Compulsive Scale modified for PDD (Scahill et al., 2006). These measures are able to establish the presence of RRB but do not provide comprehensive exploration of various sub-types. Traditional rating scale measures for caregivers' report of RRB often require an endorsement of whether the behavior occurs and, either a rating of the frequency of the behavior or a rating of the degree to which the behavior interferes with functioning, or both. These measures can exclusively focus on RRB, or combine other behavioral, traits or an ASD diagnosis.

The Child Symptom Inventory (CSI) is a 126-item measure with items based on the *DSM-IV-TR* (2000) that are rated from 0 to 3 by the primary caretaker (Gadow & Sprafkin, 1994). However, this measure also screens for several other *DSM-IV-TR* diagnoses and does not differentiate RRB subtypes. Mirenda et al. (2010) noted how measures designed to differentiate RRB subtypes, such as the *Restrictive Repetitive Behaviors Scale - Revised* (RBS-R; Bodfish, Symons, & Lewis, 1999) possess greater clinical application in the assessment of and treatment planning for RRB. For example, the ADI-R measures RRB to establish their presence but not the specific type; although this aids in determining diagnosis, it may be insufficient for determining appropriate interventions (Mirenda et al.).

In light of the proposed *DSM-V* (American Psychiatric Association, 2011) changes (i.e., identification of two RRB), measures that identify the presence or absence of RRB will need follow up information in order to specify the RRB sub-type. The RBS-R measures severity for a variety of RRB subtypes and has applicability for clinic, home, and school settings. Previous research indicates that the RBS-R is reliable for caregivers as demonstrated

by high Cronbach's alphas for all of the subscales (.78 to .91, mean $\alpha = 0.83$; Lam & Aman, 2007). Mirenda et al. (2010) found evidence to support the validity of the RBS-R with a preschool aged population using a five subtypes (ritualistic/sameness [ST], self-injurious [SIB], restrictive stereotypy [RS], compulsive [CO], and restricted interests [RI]). Each RBS-R subtype was compared to the RRB score on the ADI-R using a non-parametric Spearman's rank order correlation because the data were not normally distributed (ST $r = .297$, $p < .01$; SIB $r = .129$, RS $r = .362$, $p < .01$; $p < .05$; CO $r = .330$; $p < .01$; RI = .351, $p < .01$). Mirenda et al. also suggested that the RBS-R is psychometrically sound and provides a better measure of RRB than the ADI-R. In summary, according to Mirenda et al., the five subtype model demonstrates the breadth of the construct validity of RRB for preschool aged children with ASD.

The SRS is a commonly used standardized rating scale that provides a clinical subscale for RRB (Constantino & Gruber, 2005). The SRS was standardized with a sample of 1,600 children from the general population, ages 4-18 using parent and teacher reports. The SRS provides *T*-scores for children by gender and rater (Constantino & Gruber).

Measures of RRB rated by the teacher. Validated measures of RRB for teachers include multi-symptom measures such as the SRS and unitary symptom measures such as the RBS-R. Lam and Aman (2007) found support for inter-rater reliability for the RBS-R in outpatient settings (median reliability of subscales = .67; Lam, 2004). The SRS was normed for teachers and parents and shows discriminate, concurrent, and structural validity (Constantino & Gruber, 2005). However, validity studies presented in the SRS manual do not

provide separate validity coefficients for teachers. Constantino, Przybeck, Friesen, and Todd (2000) found discriminate validity for the SRS between children with and without ASD for teacher as well as parents. Based on the normative sample, there appears to be variability by gender of the child for both parent and teacher ratings (Constantino & Gruber). For example, female children were on average rated lower than males by both parents and teachers (Constantino & Gruber). Constantino and Gruber indicated that teachers' ratings of RRB (M = 6.6, SD = 7.2) were higher for males compared to parents' ratings (M = 4.6, SD = 4.4) and similar for female children with teachers' ratings (M = 3.6, SD = 5.1) being slightly higher than parent ratings (M = 3.2, SD = 3.4). Inter-rater agreement coefficients are presented in the manual based on a clinically referred sample (N = 62). Correlations between teacher and mother ratings (r = .82) and teacher and father ratings (r = .75) suggest strong inter-rater agreement (Constantino & Gruber). The SRS manual provides a clinical subscale score for RRB with an alpha coefficient of .90 as well as T-score norms for teachers based on gender of the child (Constantino & Gruber).

Summary of RRB rated measures. RRB present certain challenges when attempting to measure or evaluate them. Such challenges include the variability in expression of RRB, latency of occurrence, duration, intensity, and developmental appropriateness of behavior. Diagnostic measures, such as the ADOS, are extensively used to identify RRB (Gotham, Risi, Pickles, & Lord, 2007) and may identify RRB not reported by parents, especially when the child is less than two years of age (Kim & Lord, 2010). Broad screening measures, such as the SRS, screen for the presence of RRB in conjunction with other symptoms of ASD and

were normed for both teacher and caregiver groups (Constantino & Gruber, 2005). Establishing the presence of symptoms is essential to accurate diagnosis but broad-band screening instruments do not provide sufficient specificity of RRB subtypes. Therefore, collecting increased specificity of RRB requires the use of more specific measures, such as the RBS-R. The reviewed assessments share the same limitation that they alone cannot exhaustively screen for all expressions of RRB. Much of the validity data available on measures of ASD relate to concurrent validity.

Social and Communication Deficits

Abnormal patterns of social and communication development are at the core of ASD symptomatology since the earliest reports. Leo Kanner (1943) noted these patterns of social and communication deficits across the 11 cases presented in his seminal work. This was also true for Hans Asperger's report (1944) but with greater focus on the social defects, including a lack of empathy, limited ability to develop or maintain peer relationships, and one-sided conversations. Measurement of social and communication symptoms became increasingly important due in part to the inconsistent manner in which clinicians came to a formal diagnosis of this new disorder (Lord, 2010). Although the presence of RRB was an indicator of ASD, social and communication deficits are unique to ASD (Lord).

Clinical criteria. The *DSM-IV-TR* (2000) diagnostic criteria for ASDs consider social and communication deficits as separate domains. The social domain is defined as a "qualitative impairment in social interaction" (pp. 75 & 84), which is demonstrated by the presence of two of the following: (a) impairment in several non-verbal behaviors; (b) peer

relationships fail to develop; (c) efforts to share enjoyment, interests, or achievements is lacking; and (d) social or emotional reciprocity is limited. The definition of communication domain is as a "qualitative impairment in communication" (p. 75). This impairment is demonstrated by the presence of two of the following: (a) development of spoken language is delayed; (b) inability to initiate or sustain conversation when speech is present; (c) language is repetitive or idiosyncratic; and (d) development of play appears delayed and lacks variety of play, spontaneous make-believe, or imitative play.

The diagnostic changes proposed for the *DSM-V (*American Psychiatric Association, 2011) include a two factor-model that considers social and communication deficits as one symptom or domain area, defined as a reciprocal SC deficit (American Psychiatric Association, 2011; Lord, 2010). This proposed change is based on research findings by Snow Lee, and Ashford (2009) who found that a two-factor model with RRB and SC deficits uniquely represented separate factors compared to the traditional three factor model in the diagnosis of ASDs using the ADI-R.

Typical and atypical development. There is a great deal written regarding social and communication development for both typical and atypical development. Given that this symptom area is more associated with the absence of specific skills, it is easier to define atypical development. Social behaviors are often notably impaired such as making eye contact, soliciting caregivers for social attention, and facial expressions. There appears to be a relationship between impairments in basic behaviors and more advanced social skills (Gotham et al., 2007; Lord et al., 1994).

Measures of social and communication deficits. Methods used to measure social and communication deficits include rating scales, diagnostic observations, and clinical interviews. Each of these approaches requires establishing the child's language/communication and social skills. A trained independent observers (e.g., a clinician) would complete assessments involving diagnostic observation (e.g., ADOS) or clinical interview (e.g., ADI-R). Rating scale measures are thus left to caregivers and teachers.

Measures of SC rated by the caregiver. There are many measures of SC for caregivers. The SRS measures impairment on a quantitative scale and provides a total score of SC impairment. Four subscales from this instrument provide measures of impairment for receptive, cognitive, expressive, and motivational aspects of social behavior (Constantino & Gruber, 2005). Evidence of validation is primarily based on research studies in which clinical interviews were compared to the SRS (Conway & Venn, 2007). Constantino et al. (2000) found evidence of discriminant validity using SRS scores indicating social deficits were continuously distributed and differentiated children with ASDs from other psychiatric disorders. Constantino, Hudziak, and Todd (2003) found evidence of concurrent validity with the ADI-R and discriminant validity with mean scores differed between those with an ASD diagnosis and those with developmental disorders ($F = 72.95$, $df = 2.58$, $p < .0001$). Using twin comparisons, Constantino and Todd (2000) found evidence of structural validity with consistent deficits in SC and RRB across twins. Much like RRB measures, families are one of the best sources to establish a child's abilities. Inter-rater reliability is high between mother-father ($r = .91$), mother-teacher ($r = .82$) and father-teacher ($r = .75$; Constantino et

al., 2003). Concurrent validity of the SRS was strong for mothers', fathers', and teachers' scale ratings (i.e., *rs* > .64; Constantino et al., 2003), when correlated with items from the ADI-R that were derived from the ASD *DSM-IV-TR* (2000) criterion.

Measures of SC rated by 7. Measures of SC are similar for teachers and caregivers and often differ only in question construction or which questions are more relevant to each group. Constantino et al. (2003) found moderate evidence for concurrent validity by comparing total scores based on teacher and caregiver assessment on the SRS with the ADI-R subscales for SRS based on ASD behavioral symptoms (e.g., social deficits, $r = .67$; verbal communication, $r = .65$; and nonverbal communication, $r = .52$). Using a preschool-age sample of typically developing children and children diagnosed with ASD, Pine, Luby, Abbacchi, and Constantino (2006) found good inter-rater reliability between parents and teachers ($r = .79$) between the SC items on the SRS.

Summary of SC rated measures. Measures of SC are readily available and more abundant when compared to measures of RRB (Lam & Aman, 2007). The SRS as a measure of SC is well suited for multirater comparison, given that its standardization is for both caregivers and teachers. The SRS parent and teacher versions are identical in questions asked as well as order of questions, which allows for item analysis. The ADOS SC composite also presents as a strong independent measure for SC, given that it elicits targeted behaviors with behavioral presses that is not dependent on report. Measures of SC are given to determine the presence of developmentally appropriate SC skills through questionnaires (i.e., caregiver and teacher reports) or by attempting to elicit the behavior (i.e., ADOS).

Purpose of the Study

The primary purpose of this study was to further develop the evidence base for construct validity of common measures of ASD traits across raters. Given the challenges associated with accurate diagnosis, evidence for construct validity is fundamental to interpreting and making clinical judgments. The diagnosis of ASD requires psychometrically valid assessment, yet little is known regarding how well these measures compare across raters. The present study used a MTMR matrix similar to that described by Campbell and Fiske (1959) to examine the construct validity of RRB and SC traits. The MTMR matrix included caregiver, teacher, and independent observer ratings on commonly used measures of ASD with a sample of preschool-aged children.

Research Design

The coefficients in the MTMR matrix are divided into four primary categories (see Figure 1). Due to unequal number of measures for the SC and RRB constructs. one additional category of coefficients is presented that depicts the RRB trait by the same rater (monotrait - monorater) for caregivers and teachers only. These coefficients are treated as a measure of inter-rater reliability (Campbell & Fiske, 1959).

Multitrait-multirater matrix of ASD symptoms by Independent Observer, Caregiver, and Teacher

			A_1	B_1	A_2	B_2	B_3	A_3	B_4	B_5
Independent Obs.	ADOS (SC)	A_1	(α)							
	ADOS (RRB)	B_1	DV	(α)						
	SRS (SC)	A_2	CV	DV	(α)					
Caregiver	RBS-R	B_2	DV	CV	DV	(α)				
	SRS (RRB)	B_3	DV	CV	DV	R	(α)			
	SRS (SC)	A_3	CV	DV	CV	DV	DV	(α)		
Teacher	RBS-R	B_4	DV	CV	DV	CV	CV	DV	(α)	
	SRS (RRB)	B_5	DV	CV	DV	CV	CV	DV	R	(α)

Figure 1. On the diagonal are the reliability coefficients (α). Convergent validity (CV) coefficients are the correlations between measures of like constructs. Multimethod reliability coefficients (R) are the correlations between different measures of RRB by the same rater (monotrait - monorater). Discriminant validity (DV) coefficients in dark gray, outlined boxes are correlations between measures of different constructs by the same rater (heterotrait - monorater) whereas; light gray boxes are correlations between measures of different traits and raters (heterotrait -heterorater). Obs. is observer, ADOS is the Autism Diagnostic Observation Scale, SRS is the Social Responsiveness Scale, RBS-R is the Repetitive Behavior Scale - Revised, SC is social communication, and RRB is restrictive repetitive behaviors.

The first category of coefficients comprises the reliability diagonal, which runs along the top diagonal of the matrix and displays the reliability coefficients for each measure. Using Cronbach's alpha (α) as a measure of reliability, the internal consistency of each measure shows the strength of association of the scale item scores correlated with each other. In addition, there are two measures of the RRB trait provided by the caregivers and teachers. Campbell and Fiske (1959) suggest this is a form of reliability as it measures the rater consistency of items purported to measure the same trait by the same raters. An alpha value of.70 or higher indicates a scale with acceptable internal consistency (DeVellis, 1991).

The second category of coefficients measure CV that is composed of coefficients that measure the same trait (i.e., either SC or RRB) across different raters (i.e., monotrait – heterorater). The third and fourth coefficient categories are of discriminant validity (DV) and include coefficients of different traits across different raters (i.e., heterotrait - heterorater) or across the same raters (i.e., heterotrait - monorater).

In the MTMR matrix, construct validity is determined by measures of larger magnitude in CV coefficients versus DV coefficients. Additionally, CV is determined by a high degree of association between measures that purport to measure the same trait without the confounding shared variance of rater bias, which gives evidence for the construct. Therefore, the CV coefficients are expected to be significantly greater than the heterotrait-heterorater DV coefficients. CV coefficients are expected to be significantly greater than the heterotrait – monorater DV coefficients when statistically tested. In the MTMR matrix, DV coefficients are expected to be positively correlated, as they are comparisons of different

traits using the same rater type (i.e., heterotrait - monorater) and different traits and raters (i.e., heterotrait - heterorater). Heterotrait – monorater coefficients potentially share rater judgment bias. Therefore, the heterotrait – monorater coefficients will be significantly higher in magnitude than the heterotrait – heterorater coefficients. Given the MTMR matrix interpretation logic described above, the research hypotheses are as follows:

Research Hypotheses

Hypothesis 1. The 11 CV coefficients will be greater than the 10 heterotrait-heterorater DV coefficients when statistically tested. The test statistic is the Wilcoxon Signed Rank ordered test, which is the equivalent to a repeated measures analysis of variance although with nonparametric data. The statistical test requires an equal number of cases to be rank ordered, so the analysis will require two steps. The first step will test the 10 highest rank ordered CV coefficients against the 10 DV coefficients. The second step will replace the highest rank ordered coefficient with the last coefficient testing the 10 smallest CV coefficients against the 10 DV coefficients.

$\mathbf{H_1}$ 11 CV (monotrait - heterorater) >10 DV (heterotrait - heterorater)

$\mathbf{H_{1a}}$ 10 largest CV (monotrait - heterorater) > 10 DV (heterotrait - heterorater)

$\mathbf{H_{1b}}$ 10 smallest CV (monotrait - heterorater) > 10 DV (heterotrait-heterorater)

Hypothesis 2. The 11 CV coefficients will be significantly greater than the 5 heterotrait – monorater DV coefficients when statistically tested. The first step will test the five largest CV against the five DV coefficients. Because the number of comparisons is

unequal, the statistical analysis will require two steps. The second step will test the five smallest CV coefficients, against the five DV coefficients.

H_2 11 CV (monotrait – heterorater). > 5 DV (heterotrait - monorater)

H_{2a} 5 largest CV (monotrait – heterorater) > 5 DV (heterotrait-monorater)

H_{2b} 5 smallest CV (monotrait – heterorater) > 5 DV (heterotrait-monorater)

Hypothesis 3. The heterotrait – monorater coefficients will be significantly higher in magnitude than the heterotrait – heterorater coefficients. It is expected that rater bias will produce a larger amount of shared variance than different raters will across different trait assessments. Again, a two-step statistical analysis will be used. The first step will test the five largest heterorater DV coefficients against the five monorater DV coefficients. The second step will test the five smallest heterorater DV coefficients against the five monorater DV coefficients. The second step will require one replacement group, which will be the lowest heterorater DV coefficient from the first step.

H_3 5 DV (heterotrait - monorater) > 10 DV (heterotrait - heterorater)

H_{3a} 5 DV (heterotrait - monorater) > 5 largest DV (heterotrait - heterorater)

H_{3b} 5 DV (heterotrait - monorater) > 5 smallest DV (heterotrait - heterorater)

Method

Sample and Participant Selection

The data in this study come from a research project entitled, *Comparison of Two Comprehensive Treatment Models for Preschool-aged children with Autism and their Families* (principal investigator S. Odom, U.S. Department of Education, Institute for

Education Sciences; IES: [R324B070219]). The larger study was a four-year, national multi-site study with data collectedfrom four states: Colorado, Florida, Minnesota, and North Carolina. The project's larger aim was to contribute to the improvement of the cognitive, communication, academic, social, and behavioral outcomes of preschool-aged children identified with ASD and their families by comparing the outcomes of preschool children enrolled in the Treatment and Education of Autistic and Communication-Handicapped Children (TEACCH) model and the Learning Experiences: Alternative Program for Preschoolers and Parents (LEAP) model.

Sampling procedures. Following approval from institutional review board at the University of North Carolina at Chapel Hill, the University of Miami, the University of Colorado, Denver, and the University of Minnesota and participating school districts, screening and recruitment of pre-school teachers for participation in the larger treatment study began. School district personnel were instrumental in identifying teachers who met the study criteria, which required a minimum of two consecutive years working with children with autism in either an inclusive or mostly separate classroom setting. Teacher recruitment occurred by phone, mail, or e-mail and, after consenting study personnel observed classroom practices to ensure teachers' practices were of sufficient fidelity. Approval from the institutional review board at North Carolina State University was granted for analysis of the data.

Participants. A total of 76 teachers (75 female, 1 male) were recruited for participation in the larger study, including 15 teachers (19.7%) from Colorado, 24 (31.5%)

from Florida, 16 (21%) from Minnesota, and 21 (27.6%) from North Carolina. Teacher compensation for their participation was $500. Participant characteristics for this study are shown in Table 1.

Table 1

Child and Teacher Demographic Data

	Gender		Race				Ethnicity	
	Female	Male	Asian	Black	Mixed	White	Hisp./Latino	Not Hisp./Lat.
Child								
n	35	165	10	26	7	157	69	131
Percentage	17.5%	82.5%	5%	13%	3.5%	78.5%	34.5%	65.5%
Teacher								
n	74	1	0	2	1	65	11	57
Percentage	98.7%	1.3%	0%	2.7%	1.3%	86.7%	14.7%	76.0%

Note. Seven teachers did not provide race or ethnicity data, which accounts for 9.3% not reported.

Materials and Procedure

After completing the screening and consent process for teachers, families in selected classrooms received information packets regarding the study. Family members could contact

study personnel to learn more or return their informed consent. Upon receipt of informed consent, an additional phone conversation was used to ensure the child did not have significant untreated hearing, vision, or traumatic brain injury. To meet full inclusion, the child also had a clinical or an educational diagnosis of ASD and met the cut-off-score on the ADOS.

Collection of data from caregivers and teachers included the RBS-R as a measure of RRB and the SRS as a measure of SC. Teachers and caregivers received and completed the RBS-R and SRS during the beginning of the school year. Efforts were made to ensure that teacher and caregivers completed assessments within two weeks of the other. Study personnel met with the family most often in their home to collect additional data and answer questions regarding protocols. The home visit provided the opportunity for parents to ask questions and allowed study personnel to follow up on any missing data. Teachers also had the opportunity to talk with study personnel. Administration of the ADOS occurred in the school setting. Study personnel who conducted the ADOS completed the clinical training workshop as well as the research-training workshop. Ten percent of the ADOS assessments were videotaped and then re-scored by another ADOS trained researcher. In the event of inter-rater disagreement of more than one point, the two researchers viewed the taped assessment and resolve the scoring discrepancy using the assessment manual. For example, a child may have displayed a social smile to the researcher conducting the ADOS, but this was not visible on the video, and therefore, would not be counted as a disagreement.

Autism Diagnostic Observation Schedule (1989). The ADOS is a semi-structured assessment that consists of presses or activities designed to elicit behaviors (Lord, Rutter, Dilavore, & Risi, 2000). These activities press for specific skills or behaviors with a standardized administration. A trained examiner first presents a child with an activity; if the child fails to respond, the examiner presents graduated presses for the target behavior. These activities allow the assessor an opportunity to observe a child's social and communication behaviors, which relate to the diagnosis of ASD. The ADOS consists of four modules. Administration includes only one module, which depends on the examinee's developmental and language level. Each module on the ADOS includes a standardized use of toys and activities that allow an examiner to observe and rate specific behaviors. The revised version includes diagnostic algorithms used to assess two behavioral domains, SC and RRB (Gotham et al., 2007). Validity of the new algorithms was established with factor analytic results (Oosterling et al., 2010).

Reliability for the ADOS is well documented (Lord, Rutter, et al., 2000). The interclass correlations between the four different modules for the ADOS SC total score averaged .92 and for the RRB total score the average interclass correlation was .82 (Lord, Rutter, et al.).Test-retest reliability for the SC modules was .82 and, for the RRB, it was .59 (Lord, Rutter, et al.) after a 2 week period between testing.

Repetitive Behavior Scale - Revised (2000). The RBS-R is an empirically based rating scale comprised of 43 items that provide a measure of overall RRB in addition to 6 subscales of specific domains of the RRB. Each item describes a specific behavior (e.g., bites

self, insists on sitting at the same place, and strongly attached to one specific object) and is rated on a scale from 0 - 3 (0 = behavior does not occur to 3 = behavior occurs and is a severe problem). The six subscales of the RBS-R represent the broad categories of RRB: (a) compulsive behavior, (b) restricted behavior, (c) ritualistic behavior, (d) sameness behavior, (e) self-injurious behavior, and (f) stereotypic behavior (Bodfish et al., 1999). Lam and Aman (2007) found a statistically significant five-factor model for the RBS-R using factor analysis. The five-factor model loaded Ritualistic Behavior and Sameness Behavior on the same scale becoming Ritualistic/Sameness Behavior (validity coefficient =.55). The inter-rater reliability across caregivers was .60 and across teachers was .60, which is considered marginal for reliability (DeVellis, 1991).

Social Responsiveness Scale (2005) The SRS is a 65-item rating scale for measuring the severity of autism spectrum symptoms in children from 36 months to 18 years of age (Constantino & Gruber, 2005). The preschool version of the SRS for children age 36–48 months, and the standard SRS version for children 48 months and older were used in this study. The two versions share similar items and differ only in the wording for rating the behaviors of children, reflecting developmental differences between the different age groups. By design, the intended use of the SRS is for a primary caregiver or teacher who possesses knowledge of the child to complete the form. It takes about 15 to 20 minutes to complete. The instrument provides an assessment of the child's social behavior as it occurs in natural social settings by the rater. The SRS subscales evaluate social awareness, social information

processing, capacity for reciprocal SC, social anxiety/avoidance, and autistic preoccupations and traits (Constantino & Gruber).

Inter-rater reliability is high between mother-father (r = .91), mother-teacher (r = .82), and father-teacher (r = .75; Constantino et al., 2003). Concurrent validity of the SRS was strong for mothers, fathers, and teachers scale ratings when correlated with the derived scores for the ASD *DSM-IV-TR* (2000) criterion established with items from the ADI-R (i.e., rs > .64; Constantino et al. 2003). Validity was further supported in epidemiological twin studies with an intraclass correlation of .73 for monozygotic twins and .37 for dizygotic twins **(**Constantino & Todd, 2000).

Results

The results of this study are presented in three sections. The first section will present descriptive statistics and each measure by trait and by rater. The second section will present the results of the MTMR matrix that presents Spearman's rho correlations between various raters and trait measures of ASD. Included in the matrices are Cronbach alpha coefficients. The third section will show the statistical tests of the hypotheses by testing coefficients in the matrices using Wilcoxon signed-rank tests.

Descriptive Statistics

Prior to addressing the hypotheses, descriptive statistics and measures of kurtosis, and skewness are presented to assess the distribution of scores (see Tables 2 and 3).

Table 2

Descriptive Statistics of the Autism Diagnostic Observation Schedule, the Repetitive Behavior Scale-Revised, and the Social Responsiveness Scale

	M	*SD*	Median	Min.	Max.
Independent Observer					
ADOS SC	12.85	3.87	13	5	20
ADOS RRB	3.94	1.98	4	0	8
Caregiver					
SRS SC	60.02	14.71	58	24	97
RBS-R	21.99	15.91	18	0	92
SRS RRB	13.52	7.19	12	0	30
Teacher					
SRS SC	63.54	13.24	65	31	97
RBS-R	18.07	13.23	16	0	64
SRS RRB	14.61	7.32	15	0	32

Note. ADOS (n = 200); Caregiver SRS (n = 184), RBS-R (n = 188); Teacher SRS (n = 197), RBS-R (n = 197). ADOS is the Autism Diagnostic Observation Scale, SRS is the Social Responsiveness Scale, and RBS-R is the Repetitive Behavior Scale - Revised. Raw scores are reported.

To test the assumption of the scales being normally distributed a skew and kurtosis Z-score test was performed using the criterion of $z = \pm 1.96$. Most measures did not show an uneven normal distribution of scores with the absolute value of z_{skew} score test ± 1.96 (see Table 3). The SRS SC ratings from both caregiver and teacher and the ADOS RRB were within normal limits. However, the RBS-R distributions showed a positive skew for both caregiver and teacher ratings. Kurtosis scores that violated the assumption of normality included both the RBS-R ratings by caregiver and teacher. Platykurtic distributions included SRS RRB ratings for both caregiver and teacher as well as on the ADOS SC. Therefore, Spearman's Rho, a rank ordered correlation, was used rather than Pearson's *r* to determine the validity coefficients.

Table 3

Skew, Kurtosis, and Z-score Tests

	Skew	*SE*	z_{Skew}	Kurtosis	*SE*	$z_{kurt.}$
Independent Observer						
ADOS SC	-0.03	0.17	-0.17	-1.04	0.34	-3.04**
ADOS RRB	0.23	0.17	1.35	-0.67	0.34	-1.95
Caregiver						
SRS SC	0.35	0.18	1.95	-0.49	0.36	-1.36
RBS-R	1.19	0.18	6.72**	1.75	0.35	4.95**

Table 3 Continued

SRS RRB	0.27	0.18	1.48	-0.80	0.36	-2.24*
Teacher						
SRS SC	-0.26	0.17	-1.50	-0.29	0.34	-0.84
RBS-R	1.11	0.17	6.38**	1.14	0.34	3.32**
SRS RRB	0.02	0.17	0.10	-0.69	0.34	-2.00*

Note. ADOS (n = 200); Caregiver SRS (n = 184), RBS-R (n = 188); Teacher SRS (n = 197), RBS-R (n = 197). ADOS is the Autism Diagnostic Observation Scale, SRS is the Social Responsiveness Scale, and RBS-R is the Repetitive Behavior Scale - Revised. Raw scores are reported.

* $p < .05$, ** $p < .01$.

Multitrait Multirater Matrix

To interpret the MTMR matrix, it is helpful to group the coefficients in four ways. The first group of coefficients run the diagonal of the matrix and represents the reliability for the trait by rater. The second group consists of convergent validity coefficients. There are 11 CV coefficients, 3 for the SC trait (i.e., A_1A_2, A_1A_3, and A_2A_3) and 8 for the RRB trait (e.g., B_1B_2). Coefficients of different traits by the same rater (heterotrait-monorater) account for five of the DV coefficients (i.e., $A_1\,B_1$, $A_2\,B_2$, $A_2\,B_3$, $A_3\,B_4$, and $A_3\,B_5$). The remaining 10 coefficients are of different traits by the different raters (heterotrait-heterorater).

Presented in Table 4 is the MTMR matrix. The results show the reliability, CV, and two aspects of DV same raters (heterotrait-monorater) and different raters (heterotrait-heterorater).

Table 4

Multitrait-Multirater Matrix of Autism Symptoms by Raters

Rater	Trait		A_1	B_1	A_2	B_2	B_3	A_3	B_4	B_5
Independent Obs.	ADOS (SC)	A_1	.74[a]							
	ADOS (RRB)	B_1	.40**	.26[b]						
Caregiver	SRS (SC)	A_2	-.11	-.07	.80					
	RBS-R	B_2	-.03	.07	.55**	.94				
	SRS (RRB)	B_3	.04	.10	.70**	.76**	.84			
Teacher	SRS (SC)	A_3	.19*	.27**	.05	-.01	.08	.76		
	RBS-R	B_4	.39**	.37**	.09	.15*	.26**	.46**	.91	
	SRS (RRB)	B_5	.40**	.42**	.06	.14	.21**	.67**	.75**	.83

Note. Spearman's rho correlations using listwise deletion, $n = 184$. Obs. is observer, ADOS is the Autism Diagnostic Observation Scale, SRS is the Social Responsiveness Scale, RBS-R is the Repetitive Behavior Scale – Revised, SC is social communication, and RRB is restrictive repetitive behavior. Raw scores are reported.

Table 4 Continued

a. Mean of Cronbach alphas for the four ADOS modules: module 1, no-words, $n = 51$, $\alpha = .70$; module 1, words, $n = 73$, $\alpha = .76$; module 2, $n = 59$, $\alpha = .71$; and module 3, $n = 17$, $\alpha = .78$.

b. Mean of Cronbach alphas for the four ADOS modules no-words, $n = 51$, $\alpha = .30$; module 1, words, $n = 73$, $\alpha = .11$; module 2, $n = 59$, $\alpha = .31$; and module 3, $n = 17$, $\alpha = .32$

* $p < .05$, ** $p < .01$.

Reliability diagonal. The reliability diagonal in Table 4 starts in the top left and runs the diagonal of the matrix. A Cronbach's alpha (α) provides estimates of the internal consistency of each measure. The reliability coefficients presented for each ADOS trait are the mean of four alphas. Cronbach alpha was conducted for each ADOS module-algorithm that used different items for the composite. Each module's algorithm used the same number of items for SC (i.e., 10 items) and RRB (i.e., 4 items); however, individual items varied based of developmental expectations of the child. Therefore, a child with phrased speech received module three and a child with no words was given module one. For example, module one assesses communication by observing the frequency of vocalizations directed to others, whereas module two and three assess the amount of reciprocal social communication observed. Therefore, these modules receive unique measures of internal consistency. In addition to the reliabilities on the diagonal, there are two monomethod-monorater coefficients (i.e., B_2-B_3 and B_4-B_5) with coefficients of .76 and .75, respectively.

DeVillis (1991) suggested that an alpha value of.70 or higher indicates a scale with acceptable internal consistency. All but one measure , the ADOS, met this threshold. Each of the modules for RRB were below the acceptable alpha modules one-no-words (α = .30), one-some-words (α = .11), two (α = .31), and three (α = .32). This is likely a result of the limited number of items measuring this trait. Each ADOS module has four RRB items and across modules items fall within the following categories: stereotyped/idiosyncratic use of words, phrases, or vocalizations; unusual sensory interests in play, material, or person; complex mannerisms; unusually repetitive interests or stereotyped behavior.

Convergent validity coefficients. The MTMR matrix includes 11 CV coefficients of the same traits correlated with different measures by different raters. Only 6 of the 11 CV coefficients were significant: ADOS SC with teacher SRS-SC (r = .19), ADOS RRB with teacher RBS-R (r = .37), and ADOS RRB with teacher SRS-RRB (r = .42), caregiver with teacher RBS-R (r = .15), caregiver with teacher SRS-RRB (r = .21), and caregiver SRS-RRB with teacher RBS-R (r = .26). The magnitude of the CV coefficients ranged from .05 to .42 suggesting a very small to medium association.

Discriminate validity coefficients. The DV coefficients in Table 4 are comparisons of different traits using the same rater type (i.e., heterotrait - monorater) and different traits and rater type (i.e., heterotrait - heterorater).

Heterotrait – heterorater coefficients. There were 10 coefficients that compared different traits and different raters. They are shown in Table 4, three of which were significantly correlated. The three coefficients were independent observer ADOS-RRB with

teacher ratings on the SRS-SC ($r = .27$, $p < .01$), the ADOS-SC with the teacher SRS-RRB ratings ($r = .40$, $p < .01$) and the ADOS-SC with the teacher RBS-R ratings ($r = .39$, $p < .01$). The magnitude of the coefficients ranged from small and negative ($r = -.07$) to medium ($r = .39$ and $r = .40$).

Heterotrait– monorater coefficients. The MTMR matrix includes five coefficients of different traits correlated with different measures by the same rater. These coefficients include one for the independent observer using the ADOS traits and two each for the caregiver and teacher. All five coefficients were significant ADOS SC with ADOS RRB ($r = .40$); teachers SRS-SC by SRS-RRB ($r = .67$), teachers SRS-RRB with RBS-R ($r = .46$); caregivers SRS-SC with SRS-RRB ($r = .70$), and caregivers SRS-RRB with RBS-R ($r = .55$). The magnitude of these coefficients range from medium to large with the large magnitude coefficients yielded by the SRS-SC by SRS-RRB correlations for teacher and caregiver ($r = .67$ and $r = .70$, respectively).

Test of magnitude. In the MTMR matrix, construct validity is determined by larger measures of magnitude in CV coefficients versus DV coefficients. This requires establishing a high degree of association between measures that purport to measure the same trait, regardless of rater type. Wilcoxon signed rank ordered tests were conducted to rigorously evaluate the degree of association between validity coefficients and the statistical hypotheses proposed.

Wilcoxon Signed-Rank Test

Testing included the 11 CV coefficients against the 10 heterotrait – monorater DV coefficients. Wilcoxon signed rank ordered test required matching an equal number of cases. Therefore, the 10 largest CV coefficients were compared against the 10 DV coefficients. Following this step, the smallest coefficient was substituted for the largest CV coefficient and compared with the 10 DV coefficients, which requires nine redundant cases.

Convergent validity vs. heterotrait – heterorater. A Wilcoxon Signed Ranks Test showed that the largest 10 CV coefficients (monotrait – heterorater) were significantly greater in magnitude than the 10 heterotrait – heterorater DV coefficients ($z = 2.40$, $p = .017$). The rank tests for the smallest 10 CV were not significantly different than the 10 heterotrait – heterorater DV coefficients ($z = 1.17$, $p = .285$). Table 5 provides the descriptive statistics for the Wilcoxon signed rank ordered test and Table 6 provides the rank comparisons.

Table 5

Descriptive Statistics for the Wilcoxon Signed Ranks Tests

	Mean	*SD*	Median	Minimum	Maximum
CV vs. heterotrait – heterorater					
Largest 10 CV	.20	.12	.17	.07	.42
Smallest 10 CV	.15	.13	.15	-.11	.37
Heterotrait – heterorater	.12	.17	.07	-.07	.30

Table 5 Continued

CV vs. heterotrait - monorater					
Largest 5 CV	.29	.10	.26	.19	.42
Smallest 5 CV	.07	.11	.10	.07	.15
Heterotrait - monorater	.56	.13	.55	.40	.70
Heterotrait - monorater vs. HT – heterorater					
Heterotrait - monorater	.56	.13	.55	.40	.70
Largest 5 heterotrait – heterorater	.25	.15	.27	.08	.40
Smallest 5 heterotrait – heterorater	-.001	.05	-.01	-.07	.06

Note. CV is convergent validity, HT is heterotrait.

Table 6

Wilcoxon Signed Ranks Test

	Negative Ranks			Positive Ranks		
	N	Mean	Sum of	*N*	Mean	Sum of
CV vs. heterotrait – heterorater						
Largest 10 CV vs. HT-HR	2	2	4	8	6.38	51
Smallest 10 CV vs. HT-HR	4	4	16	6	6.50	39

Table 6 Continued

CV vs. heterotrait - monorater						
Largest 5 CV vs. HT-MR	5	3	15	0	0	0
Smallest 5 CV vs. HT-MR	5	3	15	0	0	0
Heterotrait - monorater vs.						
heterotrait – heterorater						
HT-MR vs. Largest 5 HT-HR	0	0	0	5	3	15
HT-MR vs. Sm. 5 HT–HR	0	0	0	5	3	15

Note. CV is convergent validity, HT is heterotrait, HR is heterorater, MR is monorater, Sm. is smallest.

Convergent validity vs. heterotrait - monorater. A Wilcoxon Signed Ranks Test showed that the five heterotrait – monorater DV coefficients were significantly greater in magnitude than the largest five CV coefficients (z = -2.023, p = .04). The smallest CV were also significantly less in magnitude than the five heterotrait - monorater coefficients (z = 2.023, p = .04).

Heterotrait - monorater vs. heterotrait - heterorater. A Wilcoxon Signed Ranks Test showed that the five heterotrait –monorater DV coefficients were significantly greater than the heterotrait – heterorater DV coefficients. When tested, the five heterotrait – monorater DV coefficients were significantly greater than the largest five heterotrait–

heterorater DV coefficients (z = 2.02, *p* = .04). The second step comparing the heterotrait – monorater DV coefficients were significantly greater than the smallest five heterotrait - heterorater DV coefficients (z = 2.02, p =.04).

Discussion

According to Campbell and Fiske (1959, p. 88), one way to validate measurement of psychological constructs is to use a variant of the multitrait-multimethod matrix by examining multiple raters (i.e., multirater), if the assessment procedures typically require this method of assessment. Therefore, this study used a MTMR matrix in attempts to validate the two behavioral constructs required for the diagnosis of ASD described in the new *DSM-V* (American Psychiatric Association, 2011) definition. However, the results showed that the heterotrait-monorater DV coefficients showed the highest magnitude of association due to rater bias on the instruments used.

Each hypothesis is addressed with a possible explanation for the results in turn. The first hypothesis was that the 11 CV coefficients would be significantly greater than the 10 heterotrait-heterorater DV coefficients. The results showed that the magnitude of the correlation coefficients measuring both CV and heterotrait-heterorater DV coefficients were relatively equivalent. One possible explanation is that SC and RRB traits are better explained as different dimensions of behavior with one underlying construct of SC and RRB. This would be the case if the CV and heterotrait-heterorater DV coefficients showed a pattern of relative equivalence in magnitude. However, visual observation of the first column of coefficients under A_1 (ADOS-SC) in the MTMR matrix shows that none of the caregiver

measures (i.e., A_2, B_2, and B_3) were significantly correlated with A_1, whereas the teacher versions of the same measures were. This suggests that in comparison of the different validity coefficients, the degree of rater bias exceeds other sources of variance.

In addition, there is considerable research that supports a two-factor model over a one-factor model. Snow et al. (2009) conducted an exploratory factor analysis using the ADI-R and found that the two-factor model best demonstrated factor loadings that indicated a clear distinction between SC and RRB. Their factor analysis had adequate goodness-of-fit ratings with a Comparative Fit Index (CFI) of 0.97 and a RMSEA of 0.07 (Snow et al., 2009). Gotham et al. (2007) also found that the confirmatory factor analysis model found a better fit for a two-factor model than a one factor model for SC and RRB CFI range .94 (.90-1 indicates a good fit). These results were confirmed in a replication study using a sample of 1,282 children (Gotham et al., 2008).

The second hypothesis was that the 11 CV coefficients would be significantly greater than the 5 heterotrait – monorater DV coefficients. The results showed just the opposite with the magnitude of the 5 heterotrait – monorater DV coefficients were greater than the largest and smallest CV coefficients. One possible explanation is that the rater better accounts for the variance observed in measures of ASD than the individual traits assessed. The heterotrait – monorater coefficients in Table 4 include the largest of all coefficients in the matrix and suggest that a rater bias is far greater than initially predicted.

The third hypothesis was that the 5 heterotrait – monorater DV coefficients would be significantly greater than the 10 heterotrait-heterorater DV coefficients. This hypothesis

confirmed that the heterotrait-monorater DV coefficients showed a larger magnitude than the heterotrait-heterorater DV coefficients. As predicted, the rater bias produced a larger amount of shared variance than different raters did across different trait assessments.

The literature on rater bias has a long history, going back to Thorndike (1920) in which he depicted the "constant error of the halo" (p. 28) better known as the halo effect. This effect supposes that a rater views the person that they are rating holistically. It is not unreasonable to assume that raters under or over report symptom severity by focusing on one symptom that becomes the holistic judgment of the child's abilities. For example, a child with severe self-injurious behaviors may receive increased SC severity ratings as the result of the presence of the self-injurious behaviors. Likewise, a high functioning child with strong verbal skills may have their restricted interest overlooked as a quirk and have both SC and RRB items be underrated. The monorater coefficients for different traits further support that the rater is consistent in how they report traits relative to the other, which could indicate that bias is applied in the same direction for both traits.

Research in clinical psychology has shown various factors that can influence clinical ratings and diagnosis. For example, Mumma (2002) found that clinicians' judgments of symptom severity in depressed patients were affected by knowledge that the patient had a history of depression or if they displayed depressive non-verbal behaviors. These factors increased the likelihood of a clinician's diagnosis of a major depression episode. This indicates that trained professionals were influenced by other factors altering their holistic interpretation and clinical diagnosis. A halo effect for symptom severity in all symptom areas

resulted in elevated symptom diagnosis unrelated to the major depression episode criteria (Mumma).

In the assessment of children with ASD, Kanne, Abbacchi, and Constantino (2009) found that ratings between teachers and caregivers showed that caregivers more often endorsed severe psychiatric symptoms. Additionally, caregivers' ratings of their child without ASD and their teacher's ratings of these children found greater inter-rater reliability (Kanne et al., 2009), suggesting that assessment of children without severe behaviors such as ASD were evaluated by the raters more reliably. The authors indicate that it reflected rater bias in that caregivers endorsed other psychiatric symptoms given when they endorsed behavior symptoms indicative of ASD (Kanne et al.). This suggests that parents' holistic view of their children with severe behaviors may overgeneralize to other sorts of severe behaviors.

Another explanation of the results considers the instruments and the inherent limitations of how the items represent the constructs. The conceptualization of autism being on a spectrum certainly appears accurate when one considers the variability in the expressions, frequencies, and intensity of behavioral symptoms. This creates a challenge when attempting to measure the specific traits.

Several measurement limitations may contribute to the lack of construct validity. The first limitation presented is that there were several outlier scores on a couple of scales. A few children received scores of zero on the RRB trait, indicating that the rater did not endorse any of the symptoms of the trait; this occurred on the ADOS and the RBS-R. On the ADOS, only three children received a score of zero for RRB. This is likely due to the limited number of

items related to RRB on the ADOS (there are only four questions representing RRB). Additionally, the limited window of time to observe the child's behavior makes it difficult to determine if a behavior is indeed repetitive. The outliers on the ADOS are likely attributed to the lack of measurement sensitivity as opposed to the absence of the trait. On the RBS-R, two caregivers and three teachers did not endorse any RRB; however, they did indicate the presence of RRB on the SRS. There was no overlap between these teachers and caregivers so these outliers are likely attributed to inconsistency in reporting on these measures.

Another measurement limitation is that the ADOS had poor inter-rater agreement on the RRB portion. This is likely attributed to two factors. The first is the limited number of items, four, that contribute to the RRB score on each ADOS module. The second factor is that each RRB item on the ADOS represents a different RRB subtypes and there is no research to suggest that different RRB subtypes correlate.

In order to address these limitations, each measure would have to include more items. The ADOS would benefit from including items that address routines and circumscribed interest. However, it would seem unlikely that many of these behaviors would be detectable in the observation portion of the ADOS. To address this, the ADOS would likely benefit from the addition of a short semi-structured interview that focuses on RRB. Adapting portions of the ADI-R (Lord et al., 1994) or Children's Yale-Brown Obsessive Compulsive Scale modified for PDD (Scahill et al., 2006) would certainly increase the cohesiveness and scope of the ADOS. The RBS-R would benefit from the inclusion of items that cover a broader range of RRB, which would allow the measure to include the variability across this trait.

Additionally, both the RBS-R and the SRS would benefit from the addition of a metric that assesses internal consistency of the rater such as the F-scale on the *Behavior Assessment System for Children*, Second Edition (BASC-2). This metric would aid in determining the presence of rater bias as well as help the diagnostician's interpretation of the rater's perceived severity of child's behavior.

Limitations of the Present Study

There are several limitations to this study to consider. The sample was restricted in range in both age and population, given all the children met the diagnostic criteria for ASD according to the diagnostic criteria used in this study, which reduces the variability expected on these instruments, because these children's behavior is on a far end of the continuum. As a result of the range restriction, it is estimated that the construct validity is lower than the true construct validity (Hunter & Schmidt, 2004).

Another limitation of this study was the measure differences for each trait. It is arguably easier to identify a deficit in SC, given that questions can focus on a specific, observable developmental skill. The SC behaviors can be thought of as a negative trait, given that are they are behaviors expected to be present in a typically developing child and absent in a child with ASD. Conversely, RRB can be thought of as positive behaviors, given that their presence is an addition to what is expected. The ADOS is a good example of a measure that predominantly focuses on SC, given it is easier to create presses to determine the absence of a specific skill. Arguably, the ADOS underrepresents the RRB trait, using only four items in comparison to ten items for SC. Although the time allotted in administering an

ADOS is insufficient to get a comprehensive account of all RRB, the scope of the behaviors that are covered is quite limited.

The measurement of RRB is impeded by two challenges, item specificity and rater instructions. The expression of RRB can take the form of many discrete behaviors that fall within several subtypes, which makes item construction difficult. Making a comprehensive evaluation, which included highly specified items depicting discrete behaviors, would be impractical. Therefore, measures must not be too specific nor should they over generalize, as the repetition of some behaviors is appropriate (e.g., daily showering or brushing teeth). The SRS uses 23 items that oscillate between general and specific items. For example, some questions ask in a general manner about focusing on just one topic, while others items to specifics about odd behaviors like hand flapping. Therefore, RRB item inclusion is "hit or miss."

The measurement of RRB becomes more complicated given that the same behavior was at one time developmentally appropriate for the child. RRB only becomes a positive trait relative to developmental expectations. Therefore, the instrument must be developed with a larger sample that covers the range of RRB and is tested rigorously with discriminant validity methodology to determine that children with ASD are accurately identified. The items will need to evaluate the child's behaviors to take into consideration the latency, frequency, and intensity of specific behavior that determine if the behaviors are problematic. The RBS-R covers a broad range of specific behaviors, yet it is assumed that raters can accurately judge with little instruction compared to typical child development. In addition to being subject to

interpretation, the repetitive nature of these behaviors requires longer periods of observation to establish the repetition. In general, people will go to the doctor because their child is not talking before they go because of odd repetitive behaviors.

Conclusions and Future Perspectives

Despite the limitations identified, this study adds to ASD research by demonstrating that measures of ASD across different raters do not reflect strong ratings of convergence. The results further suggest the use of caution when assuming that different measures of ASD have CV between them. The results lend support to the idea that the evaluators of ASD have experience with ASD and closely consider the measures used and consider how they are interpreted by raters. Such evaluators should consider factors such as the halo effect as well as the limitations of specific measures. Additionally, rater bias, while not extensively investigated in ASD, should be considered by all raters including the potential bias of the clinician.

Future studies should investigate factors that influence raters' evaluations of ASD symptoms. This area is likely to influence the accuracy of an ASD diagnosis. Research focusing on specific factors relative to caregiver and teacher bias will certainly be applicable in improving measures as well as helping clinicians evaluate the use of such measures. It is also recommended that measures of ASD include an internal reliability score or some other statistic that indicates when a rater is significantly skewed in their reporting (e.g., F-scale on the BASC-2). Such a statistic will aid the ASD evaluator in weighing the accuracy of ratings.

Additionally, the use of measures of ASD assume a degree of competence by the rater. Given the complex nature and expression of RRB and the absence of specific skills in SC, greater attention focused on the instruction of raters would be helpful. One idea would to create measures that concisely demonstrate RRB or SC skills. This is likely most applicable with RRB, given that it can be difficult to briefly describe the duration, latency, and frequency of RRB. A paper and pencil assessment such as the RBS-R adapted into a computer-based instrument could quickly qualify targeted RRB with short video clips of RRB. Instruments that both inform the rater and provide opportunity to rate may increase the accuracy of ratings.

Additionally, measures should increase the scope of behaviors to accurately reflect the individual trait in each diagnostic trait. An ASD diagnosis includes SC and RRB traits but measures like the ADOS do not fully represent RRB.

References

Allen, R. A., Robins, D. L., & Decker, S. L. (2008). Autism spectrum disorders: Neurobiology and current assessment practices. *Psychology in the Schools, 45*(10), 905-917. doi:10.1002/pits. 20341

American Psychiatric Association. (2000). *Diagnostic and statistical manual of mental disorders* (4th ed., text rev.). Washington, DC: Author.

American Psychiatric Association. (2011, January 24). *Proposed draft revisions (updated January 24, 2011) to the Diagnostic and statistical manual of mental disorders* (5th ed.). Retrieved from: http://www.dsm5.org

Asperger, H. (1944). Die "autistichen Psychopathen" im Kindersalter [Autistic psychopathy of childhood]. *Archive für Psychiatrie und Nervenkrankheiten, 117*, 76–136.

Berkson, G. & Tupa, M. (2000). Early development of stereotyped and self-injurious behaviors. *Journal of Early Intervention*, *23*(1), 1-19. doi:10.1177/10538151000230010401

Bishop, S., Reichler, 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. doi:10.1080/09297040600630288

Bodfish, J. W. (2011). Repetitive behaviors in individuals with autism spectrum disorders. In D. G. Amaral, G. Dawson, & D. H. Geschwind (Eds.), *Autism spectrum disorders*. New York, NY: Oxford University Press.

Bodfish, J. W., Crawford, T. W., Powell, S. B., Parker, D. E., Golden, R. N., & Lewis, M. H. (1995). Compulsions in adults with mental retardation: Prevalence, phenomenology, and comorbidity with stereotypy and self-injury. *American Journal Mental Retardation, 100*(2), 183-192.

Bodfish, J. W., Symons, F., & Lewis, M. (1999). *The Repetitive Behavior Scale: Atest manual.* Morganton, NC: Western Carolina Center Research Reports.

Bodfish, J. W., Symons, F. J., Parker, D. E., & Lewis, M. H. (2000). Varieties in repetitive behavior in autism. *Journal of Autism and Developmental Disorders*, *30*(3), 237–243. doi:10.1023/A:1005596502855

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. *Journal of Child Psychology and Psychiatry, 49*(8): 809–816. doi:10.1111/j.1469-7610.2008.01897.x

Boyd, B. A., Baranek, G. T., Sideris, J., Poe, M. D., Watson, L. R., Pattern, E., & Miller, H. (2010). Sensory features and repetitive behaviors in children with autism and developmental delays. *Autism Research*, *3*(2), 78-87. doi:10.1002/aur.124

Campbell, D. T. & Fiske, D. W. (1959). Convergent and discriminant validation by multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. doi: 10.1037/h0046016

Canales, J. J., & Graybiel, A. M., 2000. A measure of striatal function predicts motor stereotypy. *Nature Neuroscience, 3*(4), 377–383. doi:10.1038/73949

Chawarska, K., Klin, A., Paul, R., & Volkmar, F. (2007). Autism spectrum disorder in the second year: Stability and change in syndrome expression. *Journal of Child Psychiatry*, *48*(2) pp. 128-138. doi:10.1111/j.1469-7610.2006.01685.x

Constantino, J. N., Davis, S. A., Reich, W., Schindler, M. K., Gross, M. M., Brophy, S. L., … Todd, R. D. (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. doi:10.1023/A:1025014929212

Constantino, J. N., & Gruber, C. P. (2005). *Social Responsiveness Scale* (SRS). Los Angeles, CA: Western Psychological Services.

Constantino, J. N., Gruber, C. P., Davis, S., Hayes, S., Passanante, N., & Przybeck, T. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry*, *45*(4), 719–726. doi: 10.1111/j.1469-7610.2004.00266.x

Constantino, J. N., Hudziak, J. J., & Todd, R. D. (2003). Deficits in reciprocal social behavior in male twins: Evidence for a genetically independent domain of psychopathology. *Journal of the American Academy of Child and Adolescent Psychiatry, 42*(4), 458-467. doi:10.1097/01.CHI.0000046811.95464.21

Constantino, J. N., Przybeck, T., Friesen, D., & Todd, R. D. (2000). Reciprocal social behavior in children with and without pervasive developmental disorders. *Journal of Developmental Behavioral Pediatrics*, *21*(1), 2-11. doi:10.1097/00004703-200002000-00001

Constantino, J. N., & Todd, R. D. (2000). Genetic structure of reciprocal social behavior. *American Journal of Psychiatry, 157*(12), 2043-2045. doi:10.1176/appi.ajp.157.12.2043

Conway, F., & Venn, J. J. (2007). [Review of the Social Responsiveness Scale]. In K. F. Geisinger, R. A. Spies, J. F. Carlson, & B. S. Plake (Eds.), *The seventeenth mental measurements yearbook* (pp. 743-748). Lincoln, NE: Buros Institute of Mental Measurements.

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin, 52*(4), 281-302. doi: 10.1037/h0040957

Cuccaro, M. L., Shao, Y., Grubber, J., Slifer, M., Wolpert, C. M., Donnelly, S. L., … Pericack-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. doi:10.1023/A:1025321707947

DeVellis, R. F. (1991). *Scale development*. Newbury Park, CA: Sage Publications.

Evans, D. W., Leckman, J. F., Carter, A., Reznick, J. S., Henshaw, D., King, R. A., & Pauls, D. (1997). Ritual, habit, and perfectionism: The prevalence and development of compulsive-like behavior in normal young children. *Child Development, 68*(1), 58-68.

Gadow, K. D., & Sprafkin, J. (1994). *Child symptom inventories manual*. Stony Brook, NY: Checkmate Plus.

Goodwin, M. S., Intille, S. S., Albinali, F., & Velicer, W. F. (2010). Automated detection of stereotypical motor movements. *Journal of Autism and Developmental Disorders*, *41*(6), 770-782. doi:10.1007/s10803-010-1102-z

Gotham, K., Pickles, A., & Lord, K. (2009). Standardizing ADOS scores for a measure of severity in autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *39*(5), 693-705. doi:10.1007/s10803-008-0674-3

Gotham, K., Risi, S., Dawson, G., Tager-Flusberg, H., Joseph, R., Carter, A., … Lord, C. (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. doi:10.1097/CHI.0b013e31816bffb7

Gotham, K., Risi, S., Pickles, A., & Lord, K. (2007). The Autism Diagnostic Observation Schedule: Revised algorithms for improved diagnostic validity. *Journal of Autism and Developmental Disorders*, *37*(4), 613-627. doi:10.1007/s10803-006-0280-1

Honey, E., Leekam, S., Turner, M., & McConachie, H. (2007). Repetitive behaviour and play in typically developing children and children with autism spectrum disorders. *Journal of Autism and Developmental Disorders*, *37*(6), 1107-1115. doi:10.1007/s10803-006-0253-4

Hunter, J. E. & Schmidt, F. L. (2004). *Methods of meta-analysis: Correcting error and bias in research findings (2nd ed.)*. Thousand Oaks, CA: Sage Publications, Inc.

Hutt, C., & Hutt, S. (1970). Stereotypies and their relation to arousal: A study of autistic children. In: S. Hutt, & C. Hutt (Eds.), *Behavior studies in psychiatry* (175-204). New York, NY: Pergamon Press,.

Iwata, B. A., Dorsey, M., Slifer, K., Bauman, K., & Richman, G. (1982). Toward a functional analysis of self-injury. *Analysis and Intervention in Developmental Disabilities*, *2*(1), 3-20. doi:10.1016/0270-4684(82)90003-9

Kanne, S. M., Abbacchi, A. M., & Constantino, J. N. (2009). Multi-informant ratings of psychiatric symptom severity in children with autism spectrum disorders: The importance of environmental context. *Journal of Autism and Developmental Disorders*, *39*(6), 856-864. doi: 10.1007/s10803-009-0694-7

Kanner, L. (1943). Autistic disturbances of affective contact. *The Nervous Child*, *2*, 217–250.

Kim, S. H., & Lord, C. (2010). Restrictive and repetitive behaviors in toddlers and preschoolers with autism spectrum disorders based on the Autism Diagnostic Observation Schedule (ADOS). *Autism Research*, *3*(4), 162-173, doi:10.1002/aur.142

Lam, K. S. (2004). *The Repetitive Behavior Scale Revised: Independent validation and the effects of subject variables* (Doctoral dissertation, The Ohio State University). Retrieved from: http://etd.ohiolink.edu/send-pdf.cgi/Lam%20Kristen.pdf?osu1085670074

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

Lam, K. S., Bodfish, J. W., & Piven, J. (2008). Evidence for three subtypes of repetitive behaviors in autism that differ in familiarity and association with other symptoms. *Journal of Child Psychology and Psychiatry, 49*(11), 1193-1200. doi:10.1111/j.1469-7610.2008.01944.x

Lord, C. E. (2010). Autism: From research to practice. *American Psychologist*, *65*(8), 815-826. doi:10.1037/0003-066X.65.8.815

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., DiLavore, P., & Risi, S. (2000). *The Autism Diagnostic Observation Schedule (ADOS)*. Los Angeles: Western Psychological Corporation.

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. *Journal of Autism and Developmental Disorders, 24*(5)*,* 659–685. doi:10.1007/BF02172145

Lovaas, I., Newsom, C., & Hickman, C. (1987). Self-stimulatory behavior and perceptual reinforcement. *Journal of Applied Behavior Analysis*, *20*(1), 45-68. doi:10.1901/jaba.1987.20-45

Luyster, R., Gotham, K., Guthrie, W., Coffing, M., Petrak, R., Pierce, K., … Lord, C. (2009). The Autism Diagnostic Observation Schedule—toddler module: A new module of a

standardized diagnostic measure for autism spectrum disorders. *Journal of Autism and Developmental Disorders, 39*(9)*,* 1305-1320. doi:0.1007/s10803-009-0746-z

Mahone, E. M., Bridges, D., Prahme, C., & Singer, H. (2004). Repetitive arm and hand movements (complex motor stereotypies) in children. *The Journal of Pediatrics*, *145*(3), 391-395. doi:10.1016/j.jpeds.2004.06.014

Meng, X., Rosenthal, R., & Rubin, D. B. (1992). Comparing correlated correlation coefficients. *Psychological Bulletin, 111*(1), 172-175. doi:10.1037/0033-2909.111.1.172

Mirenda, P., Smith, I. M., Vaillancourt, T., Georgiades, S., Duku, E., Szatmari, P., … The Pathways in ASD Study Team. (2010). Validating the Repetitive Behavior Scale-Revised in young children with autism spectrum disorder. *Journal of Autism and Developmental Disorders*, *40*(12), 1521-1530. doi:10.1007/s10803-010-1012-0

Mumma, G. H. (2002). Effects of three types of potentially biasing information on symptom severity judgments for major depressive episode. *Journal of Clinical Psychology, 58*(10), 1327-1345. doi:10.1002/jclp.10046

National Institute of Mental Health. (2008). *Autism spectrum disorders: Pervasive developmental disorders* (NIH Publication No. 08-5511). Retrieved from http://permanent.access.gpo.gov/lps126143/nimhautismspectrum.pdf

Oosterling, I., Roos, S., de Bildt, A., Rommelse, N., de Jonge, M., Visser, J., … Buitelaar, J. (2010). Improved diagnostic validity of the ADOS revised algorithms: A replication

study in an independent sample. *Journal of Autism and Developmental Disorders*, *40*(6), 689-703. doi:10.1007/s10803-009-0915-0

Pine, E., Luby, J., Abbacchi, A., & Constantino, J. N. (2006). Quantitative assessment of autistic symptomatology in preschoolers. *Autism*, *10*(4), 344-352. doi:10.1177/1362361306064434

Reaven, J. A., Hepburn, S. L., & Ross, R. G. (2008). Use of the ADOS and ADI-R in children with psychosis: Importance of clinical judgment. *Clinical Child Psychology and Psychiatry*, *13*(1), 81-94. doi:10.1177/1359104507086343

Richler, J., Bishop, S. L., Kleinke, J., & 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*(1), 55-69. doi:10.1017/S0954579409990265

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 and Adolescent Psychiatry*, *38*(6), 738–745. doi:10.1097/00004583-199906000-00022

Scahill, L., McDougle, C. J., Williams, S. K., Dimitropoulos, A., Aman, M. G., McCracken, J. T., … Vitiello, B. (2006). Children's Yale-Brown Obsessive Compulsive Scale

modified for pervasive developmental disorders. *Journal of the American Academy of Child and Adolescent Psychiatry*, *45*(9), 1114-23. doi:10.1097/01.chi.0000220854.79144.e7

Schneider, M. E., (2010, March 29). Proposed DSM-5 stresses symptom range, severity. *Pediatric News*, *44*(3), 28. Retrieved from http://www.lexisnexis.com.prox.lib.ncsu.edu/lnacui2api/api/version1/getDocCui?lni=7Y2B-GTH0-YC2S-5367&csi=349931&hl=t&hv=t&hnsd=f&hns=t&hgn=t&oc=00240&perma=true

Snow, A. V., Lecavalier, L., & Houts, C. (2009). The structure of the Autism Diagnostic Interview- Revised: Diagnostic and phenotypic implications. *Journal of Child Psychology and Psychiatry, 50*(6)*,* 734–742. doi:10.1111/j.1469-7610.2008.02018.x

Stone, W. L., Lee, E. B., Ashford, L., Brissie, J., Hepburn, S. L., Coonrod, E. E., & Weiss, B. H. (1999). Can autism be diagnosed accurately in children under 3 years? *Journal of Child Psychology and Psychiatry*, *40*(2), 219-266. doi:10.1111/1469-7610.00435

Szatmari, P., Georgiades, S., Bryson, S., Zwaigenbaum, L., Roberts, W., Mahoney, W., … Tuff, L. (2006). Investigating the structure of the restricted, repetitive behaviours and interests domain of autism. *Journal of Child Psychology and Psychiatry, 47*(6), 582-590. doi:10.1111/j.1469-7610.2005.01537.x

Thelen, E. (1979). Rhythmical stereotypies in normal human infants. *Animal Behaviour*, *27*(3), 699– 715. doi:10.1016/0003-3472(79)90006-X,

Thorndike, E. L. (1920). A constant error in psychological ratings. *Journal of Applied Psychology 4*(1), 25–29. doi:10.1037/h0071663

Turner, M. A. (1999). Annotation: Repetitive behavior in autism: A review of psychological research. *Journal of Child Psychology and Psychiatry*, *40*(6), 839-849. doi:10.1111/1469-7610.00502

Turner, K. C., Frost, L., Linsenbardt, D., Mcilroy, J. R., & Muller, R. (2006). Atypically diffuse functional connectivity between caudate nuclei and cerebral cortex in autism. *Behavioral and Brain Functions*, *2*(1), 34-46. doi:10.1186/1744-9081-2-34

Vollmer, T. R., Marcus, B. A., & LeBlanc, L. (1994). Treatment of self-injury and hand mouthing following inconclusive functional analyses. *Journal of Applied Behavior Analysis*, *27*(2), 331-344. doi:10.1901/jaba.1994.27-331

Warren, Z., & Stone, W. L. (2011). Best practices: Early diagnosis and psychological assessment. In D. G. Amaral, G. Dawson, & D. H. Geschwind (Eds.), *Autism spectrum disorders* (pp. 1271-1282). New York, NY: Oxford University Press. doi:10.1093/med/9780195371826.001.0001

Wehmeyer, M. L. (1991). Typical and atypical repetitive motor behaviors in young children at risk for severe mental-retardation. *American Journal of Mental Retardation*, *96*(1), 53-62.

Werry, J. S., Carlielle, J., & Fitzpatrick, J. (1983). Rhythmic motor activities (stereotypies) in children under five: Etiology and prevalence. *Journal of the American Academy of Child Psychiatry*, *22*(4), 329–336. doi:10.1016/S0002-7138(09)60667-1

Western Psychological Services. (2011, January 10). ADOS frequently asked questions. Retrieved from http://portal.wpspublish.com/portal/page?_pageid=53,84992&_dad=portal&_schema=PORTAL#What is required for someone

Zentall, S. S., & Zentall, T. R. (1983). Optimal stimulation: A model of disordered activity and performance in normal and deviant children. *Psychological Bulletin*, *94*(3), 446-471. doi:10.1037/0033-2909.94.3.446

Zohar, A. H., & Felz, L. (2001). Ritualistic behavior in young children. *Journal of Abnormal Child Psychology*, *29*(2), 121-128. doi:10.1023/A:1005231912747