The Development of the Minnesota Visual Autism Symptom Scale (MN-VASS)

A DISSERTATION
SUBMITTED TO THE FACULTY OF THE GRADUATE SCHOOL
OF THE UNIVERSITY OF MINNESOTA
BY

Diane Estelle Halpin

IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

Brian H. Abery and Susan C. Hupp, *Faculty Co-Advisors*

May, 2010

© Diane Estelle Halpin 2010

Acknowledgements

I would like to thank the many individuals who have been instrumental in the completion of this dissertation. First to my daughters, Becca and Jenna who have been my inspiration through this process: thank you for your understanding.

To my advisors, Brian Abery and Sue Hupp, I am grateful for your support, your guidance and your inspiration. To Frank Symons and Scott McConnell, thank you for helping me to stay in the program. To Mark Davison, thank you for your patience and guidance through the technical development of the scale.

To my parents who supported me going back to school throughout my life, thank you. And to Dean Restorff, who has been a source of encouragement and strength, thank you.

Finally to Debbie Thomas and the men and women at Partners in Excellence, your participation was instrumental. To the teachers who completed the scale, I am also indebted.

Dedication

This dissertation is dedicated to children with autism, their teachers, and their parents.

Abstract

The development and psychometric characteristics of the Minnesota Visual Autism Symptom Scale (MN-VASS) was described. The relationship between the literature surrounding autism symptoms, the diagnostic criteria for autism, and the resulting content of subscales was presented. Item analyses were conducted using item total correlations. All of the item total correlations were above .30, suggesting adequate item functioning. Two internal consistency estimates of reliability were computed for the MN-VASS: a split-half coefficient expressed as a Spearman-Brown corrected correlation and coefficient alpha. For the split-half coefficient, the scale was split between whole subscales so that the traits of autism measured on the scale would be equally divided between the halves. The value was .89 for split half and .90 for coefficient alpha, suggesting satisfactory reliability. Test Retest reliability was reported as a correlation coefficient of .90 for the entire scale (n=22), with subscale correlations ranging from .63 to .93. Inter rater agreement was between 60% and 87%. Convergent validity was investigated between the Childhood Autism Rating Scale and the MN-VASS. A correlation of .89 was reported between the two scales. Teacher/therapist evaluations of the scale suggest that the MN-VASS is a useful and valid measure of the important and teachable behaviors which can be addressed through a program of instruction for children on the autism spectrum.

Table of Contents

List of Tables

List of Figures

CHAPTER 1

Introduction

Infantile autism was identified by Dr. Leo Kanner in the early 1940's (Kanner, 1943). Today, the Diagnostic and Statistical Manual IV-TR (APA, 2000) describes Autistic Disorder as one of five pervasive developmental disorders (PDD). These disorders comprise what is commonly referred to as Autism Spectrum Disorders (ASD) (Wing, 1996; NIMH, 2007). The framework of a spectrum of disorders is especially appropriate to ASD because of the great variability of the individuals diagnosed with the disorders. Indeed, one line of genetic research theorizes that autistic symptoms are normally distributed in the population and those individuals with a clinical diagnosis represent those individuals in whom the normally distributed attributes have crossed a particular threshold (Baron-Cohen et al., 2000; Constantino et al., 2003; Ring, Woodbury-Smith, Watson, Wheelwright & Baron-Cohen, 2008). This heterogeneity of the presentation of symptoms of individuals with ASD is well documented (Chawarska, Klin, Paul & Volkmar, 2007; Filipek, et al., 1999; Ronald, et al., 2006; Rutter & Schopler, 1987).

According to the DSM-IV-TR, ASD affects an individual in three areas of functioning: verbal and non-verbal communication, social interaction, and restricted and repetitive patterns of behavior (APA, 2004). These three broad areas form a structure within which individuals may vary in their related strengths and challenges in skill subsets and behaviors. Understanding the individual differences among children with autism has been studied for a number of purposes, such as clarifying the borders of the disorder and contributing to the study of the genetic underpinnings of autism.

Because autism has no biological markers, but relies entirely on behavioral observations, it is not surprising that a proliferation of assessments has been created to gather particular types of data that support the area of interest under study.

Another important reason for extracting individual differences within the spectrum is in helping teachers to design effective instructional programs which are tailored to the unique needs of the individual.

In March of 2007, the Autism and Developmental Disabilities Monitoring Network (ADDMN), a group sponsored by the Centers for Disease Control to determine the prevalence of ASDs in the United States, reported the prevalence of an ASD as one in every 150 children in some areas of the United States (United States Center for Disease Control, 2009). The implication of this finding to special educators is significant because it implies that a large number of children with ASDs are in or will enter the special education system.

The sheer number of children in the system will pose a challenge to schools. In addition, because each child with autism presents a unique profile of strengths and challenges in many areas of functioning, educators are faced with a great deal of information to synthesize about these children regarding how best to support and teach them.

Most of the assessments related to autism attempt to quantify the degree of impairment in a set of domains, and then to select a scale point where a determination is made as to the clinical significance of the symptom. This is indeed the primary function of a screening and diagnostic instrument, however, as noted by Filipek et al. (1999), "As a practical matter, the assessment should be concerned not only with diagnosis as

such but with obtaining information on patterns of strengths and weaknesses important to intervention" (p.456).

In contrast to the numerous instruments which assist a clinician in ruling out an ASD, an instrument which assumes a baseline of clinical significance in areas which define autism and attempts to elucidate a finer degree of distinction of impairment within the spectrum has yet to be developed. Further, no scales currently exist which also address areas where children with autism experience challenges that are not necessarily a specific symptom of autism.

Many current diagnostic instruments have yielded valid and reliable results regarding the presence of an ASD. The degree to which they help to inform instruction varies widely. Further, individuals with ASD may or may not have co-morbid impairments in areas of cognitive, adaptive, physical and behavioral functioning which each vary on a continuum of mild to severe. Clearly these individuals form a complex group for whom very little can be generalized to support their educational programming.

In order to fully understand a student, a teacher must gather information on a wide variety of behaviors and skills. A scale which could help teachers to identify the unique differences of individuals on the spectrum could potentially alleviate the amount of work associated with synthesizing all of the information about an individual with an ASD. To this end the Minnesota Visual Autism Symptom Scale (MN-VASS) is proposed. This scale could potentially fill a gap between assessment and the development of an individualized instructional program.

In this study the development and testing of the MN-VASS is described. In Chapter two, 12 current assessments for children with autism are reviewed and described. The purpose of each assessment and how the authors established evidence of reliability and validity are included. Chapter 3 describes the methods for constructing the subscales of the MN-VASS and the literature upon which each scale was developed. Chapter 3 also describes the methods employed for gathering evidence of reliability and validity for the MN-VASS. Chapter 4 describes the results of the tests for reliability and validity, as well as the results of item analyses. In Chapter 5, the results of the reliability and validity measures are discussed in terms of how the MN-VASS compares to other assessments used for children on the autism spectrum. Limitations of the instrument and the study are also discussed with recommendations for future applications of the MN-VASS.

The Standards for Educational and Psychological Testing (American Educational Research Association, 2004) (hereafter referred to as "the Standards") describe the sound and ethical uses of assessment. The two most important considerations in evaluating an assessment are its reliability and validity.

In this study, reliability is determined through a number of measures. Because the MN-VASS employs subscales, many of the analyses are applied to the subscales as well as the aggregated scale. Chronbachs's alpha is computed to determine the internal consistency of the instrument and the subscales. Split half reliability is another measure which takes the assessment and divides it into two equivalent halves. A correlation coefficient is produced between the halves. Test-retest reliability was measured with a subgroup of individuals to determine reliability.

Validity is determined, not by a score, but by the interpretation of that score in a meaningful context (Thorndike, 2005). The Standards (AERA, 2004) identify validity as "the most fundamental consideration in developing and evaluating tests" (p. 9). The subscales of the MNVASS are constructed from a rich literature surrounding the presentation of symptoms of individuals with autism, which focuses on, but is not limited to the autism triad presented in the DSM-IV-TR (APA 2004).

Two further methods were employed to gather evidence of validity of the scales. First, a subset of teachers/therapists who completed the instrument was asked to identify the results of their assessment from a field of three profiles. This test was designed to determine if the output generated from the instrument could be matched to the child. The other measure of validity was through a review by end users and various domain experts to assess the usefulness of the output in the development of instructional programming for children with autism.

The MN-VASS can be used for four purposes: 1) to assist public school teachers to streamline the process of developing IEP's; 2) to compare the perception of strengths and weaknesses between school personnel and parent/guardians; 3) to provide accurate documentation on a student's profile during a transition, such as from elementary to middle school, or from one lead teacher to another; and 4) to collect large bodies of data that could potentially yield descriptive profile patterns which are common within this population of children. While the potential for rich and varied applications is apparent, the first step in the process is to develop a reliable and valid instrument.

CHAPTER 2

Review of the Literature

In this chapter a brief history of autism and measurement is provided, and types of measurement instruments are described. Statistical procedures used for evaluating the reliability and validity of these instruments are briefly described. Twelve instruments currently used with ASD are then examined in terms of their intended uses, the reliability and validity measures that were undertaken in their development, and the results of those measures. It is interesting to note that the development processes for these instruments vary widely. There is no standard set of procedures that is applied across the development of the scales; therefore, certain statistics are reported on some scales and not on others. The chapter concludes with a synthesis of the procedures used to establish the reliability and validity of the instruments.

A Brief History of Measuring ASD

Since the 1980's the study of ASDs has spawned numerous assessments. Most of the early scales were diagnostic and were meant to help researchers and clinicians rule out what was then considered a rare disorder. Three related events could be responsible for the proliferation of autism assessments: the rise in prevalence, the evidence supporting early intervention, and the increasingly early ages at which autism could be reliably identified.

At the time of the DSM-III (1980) autism was thought to occur very rarely, 1 in 10,000 cases. Current prevalence estimates are 1 in 150 (CDC, 2009). It is logical to see how increases in the number of children would drive a movement to refine and improve diagnostic technology. Positive results from early intervention (Heflin &

Simpson, 1998; Hurth, Shaw, Izeman, Whaley, & Rogers, 1999; Lovaas, I.O, 1987; Mundy, Sigman & Kasari, 1990) were demonstrated through increases of language and IQ and decreases in maladaptive behaviors associated with ASDs. Better instruments could therefore detect ASDs more reliably thus enabling children to access early intervention. Currently, autism can be reliably detected at 18 months of age (Baird, et al., 2000; Dawson, Estes, Munson, Schellenberg, Bernier, & Abbott 2007; Robins, Fein, Barton, & Green, 2001; Stone, Coonrod, Turner & Pozdol, 2004; Ventola et al., 2007). This line of research has been responsible for a surge in the development of instruments in the most recent years. Because of the recent nature of their development, they constitute a large portion of this review. They illustrate the most recent efforts for establishing reliability and validity in autism measurement instruments.

In 1999, Filipek and a consensus panel consisting of medical doctors, psychologists, autism experts and representatives from parent groups published practice parameters for the diagnosis and screening of ASDs. The panel recommended a two level approach. The first level consists of routine developmental surveillance and the second level consists of diagnosis and evaluation of autism. To implement this model, two broad kinds of measurement instruments are required: screening instruments and diagnostic instruments (Fillipek et al., 1999).

Further, there are two levels of screening. The first level of screening involves all children and is meant to detect some sort of deviance from the course of normal development that could indicate a problem (Robins, 2008; Williams & Brayne, 2006). The second level of screening is meant to focus on those children who “failed” the first level of screening and specifically targets a possibility of an ASD (Filipek, et al., 1999;

Charak & Stella, 2002). After a second level screening, a child would then be referred for a diagnostic evaluation for autism.

In addition to diagnosis and screening, a third type of assessment is common in the autism literature. These assessments measure either co-morbid behavioral manifestations associated with autism, or are intended to further delineate subtypes of the ASDs within the spectrum. The MN-VASS would fall into this category of assessment because it is not meant to diagnose or screen for autism. For the purposes of this review, these assessments will be referred to as ancillary assessments for autism.

Lecavalier (2005) suggests that all autism related measurement instruments can be divided between those that are completed by trained experts and those that are completed by caregivers. Using this framework, a number of diagnostic instruments, screening instruments, and ancillary assessments for autism can be reviewed. In all cases, the measurement instrument should be both reliable and valid.

The Standards (AERA, 2004) define reliability as "The degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker; the degree to which scores are fee of errors of measurement for a given group" (p. 25). Validity on the other hand is "the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test" (p.9). Evidence of reliability and validity help us determine the overall quality of an assessment. There are numerous statistical methods that we use to gather evidence of reliability and validity. To the extent that these methods have been used in the assessments reviewed, they are briefly described here.

Measures of Test Reliability Used in Autism Assessments

A key construct in measuring test reliability is measurement error. A person's score on a test consists of a true score and error. A person's true score is the conceptual score of an individual without any measurement error. The degree to which a test can control for error is therefore directly related to the degree to which the test will yield a true score over successive administrations of the test. Thus, error is a central concept in reliability studies.

Reliability is determined by gathering evidence that a test reflects a consistent true score for an individual over successive measurements. A number of statistical procedures exist to garner evidence of the reliability of a test instrument. The procedures used in the instruments reviewed are test-retest, internal consistency, and inter-rater reliability. Alternate forms and generalizability theory are also methods for establishing reliability evidence; however, neither of these techniques is reported in this literature. In addition, while Item Response Theory (IRT) offers another set of tools for collecting evidence of reliability, IRT methods have not been applied to date on any of the existing diagnostic and related ASD tests.

Test-retest reliability is a procedure where the exact same test is administered to the exact same individuals with a time period in between the administrations. Under the premise that the individual groups members have not changed specifically in the time between test administrations, a reliability coefficient is determined by testing whether the test ranks the individuals in the same overall order between testing. While reliability scores are relative to other acceptable instruments used for the same purposes, a generally acceptable reliability is above .80 (Thorndike, 2005).

Split half procedures are accomplished with a single administration of the test where the test is divided into two parts with each part functioning as a parallel form. A problem with split half procedures is that the manner in which the test is split will produce different reliability estimates. In part to overcome this problem, other measures of internal consistency were developed (Thorndike, 2005).

Coefficient alpha is another measure of internal consistency. This measure is derived by comparing the variance of the test scores with the variances of the separate items. The coefficient is the average of all possible split half coefficients, and therefore, solves the problem associated split half procedures and the variability of the different reliability coefficients due to how the test is split (Thorndike, 2005).

Measures of reliability and internal consistency are heavily influenced by the number of items in the measure. As tests become longer, these coefficients increase. Therefore, interpretations of the coefficients must always be made with regard to the length of the test. In some cases where subscales are used, reliability coefficients from the shorter subscales can be more meaningful than when all of the subscales are combined and measured in the entire assessment (AERA, 2004).

When measurement scales involve the observation of behavior, interrater reliability is crucial to the quality of the test. This type of reliability is generally established by having two raters use the same instrument to measure the same individual. When the proportion of agreement also takes into consideration the probability of chance, a kappa statistic can be calculated (Thorndike, 2005). For tests that use Likert-type scales, a weighted Kappa statistic is appropriate. Weighted Kappa allows partial credit for agreement between scale items such as “strongly disagree” and

"disagree" where the responses are not exact, but do constitute some level of agreement between raters (Hartman, 1977; Mun & VonEye, 2004; Sim & Wright, 2005).

Some of the instruments use an intraclass correlation coefficient (ICC) to measure interrater reliability (Shrout & Fleiss, 1979). The ICC is typically a ratio of the variance of interest over the sum of the variance of interest plus error.

Measures of Test Validity Used in Autism Assessments

Reliability measures are rather straightforward compared to measuring validity. The Standards (AERA, 2004) point out that validity is the degree to which accumulated evidence and theory support specific interpretations of test scores entailed by proposed uses of a test. It is not the test itself which is judged as valid as much as the uses of the scores associated with the test. It is important to note that one does not prove validity, but rather accumulates a body of evidence which can support an argument that the uses of the results of the assessment are valid. There is a subtle interplay between reliability evidence and validity evidence. Measures of internal consistency are generally reliability measures, but they can also be used to support a validity argument.

In terms of diagnostic and screening instruments, validity is somewhat more straightforward in terms of whether the assessment yields an appropriate diagnosis or screen for ASDs. We may therefore consider assessment "sensitivity" and "specificity."

Sensitivity is the ability of the instrument to yield true positives (Thorndike, 2005). In the current frame of reference it is how well the assessment picks up true cases of ASD's. Instruments which lack sensitivity are of little value in a screening capacity, since this implies that they would miss children who indeed had the disorder.

On the other hand, in terms of autism, specificity is how well an assessment picks up ASD apart from other disorders such as Attention Deficit Hyperactivity Disorder (ADHD) or Specific Language Impairment (SLI).

In some cases, particularly the cases of level 1 and 2 screeners, a good argument for the validity of a screening or assessment instrument is its "Positive Predictive Value" (PPV). This is a ratio of how many children originally identified by the assessment are found to have the disorder after a period of time. An instrument's "Negative Predictive Value" is the opposite statistic that keeps track of the number of false negatives and is also a measure of sensitivity (Altman & Bland, 1994). Some of the instruments reviewed in this chapter provide a PPV or an NPV value. It is worthwhile to note that deriving a positive or negative predictive value implies a long-term study and commensurate resources, where children are followed for a period of time to see if they develop an ASD or not.

Some scales rely on convergent validity to establish that the scale measures ASD with the same reliability as another scale. Convergent validity provides support that the new scale measures the same construct as the scale to which it is being compared. However, it also raises the question of the purpose of the new scale. If the purpose of the new scale is not significantly different than the first, why invent the new scale?

Deeper consideration must be given to the role of factor analysis in the development and use of assessments for children with autism. Factor analysis is used in a number of studies to justify the composition of subscales, to organize and reduce

items which measure the same construct, and to reveal the number of constructs being measured in the instrument.

Factor analysis is used in autism assessments as a validity argument in two ways. First, factor structures are used to confirm the organization of instruments into subscales and to identify items which do not contribute to the model. These studies are discussed within the review of the assessments which follows. The other use of a factor analysis is in supporting the underlying construct of measurement, such as Constantino's (2004) work on the Social Responsiveness Scale. Again, the use of factor analysis is discussed under the reviews of the instruments to which they are applied.

Evidence of reliability and validity help us to determine the overall quality of an instrument, yet establishing reliability and validity is far from a uniform endeavor. The following section looks at a variety of instruments designed specifically for various uses within this population with an emphasis on how the authors of these instruments established the psychometric properties of their scales and some of the accompanying research by other individuals who have either reviewed, extended, or used the instruments in their work.

With such broad applications and extensive proliferation of assessments with this group of individuals, it is prudent to examine their purposes, their development and the techniques employed when establishing the psychometric properties of the assessments. As a result of this, one might ask whether yet another assessment is necessary and if so, what types of evidence should be used to establish its reliability and validity.

Autism and Measurement

Autism was identified in 1943. Since that time, a number of diagnostic instruments and screening tools have been developed to help in the identification of the disorder. These tests offer a broad range of options and varying degrees of psychometric soundness. Lecavalier (2005) suggests that these scales can be divided between those that are completed by trained experts and those that are completed by caregivers.

Within the Lecavalier framework, we examine three selected expert-completed, diagnostic instruments, The Childhood Autism Rating Scale (CARS) (Schloper, Reichler, & Renner, 1986), The Autism Diagnostic Interview-Revised (ADI-R) (Lord, Rutter & LeCouteur, 1994); the Autism Diagnostic Observation Schedule (ADOS) (Lord, et al., 2000), two expert completed screening instruments, The Screening Tool for Autism in Two Year Olds (STAT) (Stone, Coonrod & Ousley, 2000) and the Autism Observation Scale for Infants (AOSI) (Bryson, Zwaigenbaum, McDermott, Rombough and Brian, 2008); and two expert-completed ancillary instruments, the Broader Phenotype Autism Symptom Scale (BPASS) (Dawson, Estes, Munson, Shellenberg, Bernier & Abbott, 2007) and the Children's Social Behavior Questionnaire (CSBQ) (Luteijn, Luteijn, Jackson, Volkmar & Mindera, 2000). Next, we examine two parent/caregiver-completed diagnostic instruments, the Autism Behavior Checklist (ABC) (Krug, Arick & Almond, 1980) and the Gilliam Autism Rating Scales – 2 (Gilliam, 2006), two parent/caregiver-completed screeners, the Q-CHAT (Quantitative Checklist for Autism in Toddlers) (Alison, et al., 2008) and The Modified Checklist for Autism in Toddlers (M-CHAT) (Robins, Fein, Barton & Green,

2001), and one caregiver-completed ancillary instrument, The Social Responsiveness Scale (SRS) (Constantino, 2004).

These particular instruments were chosen for a number of reasons. The CARS, ADOS, ADI-R, ABC and GARS-2 have been used extensively in the field (Allen, Robins, & Decker, 2008; Charak & Stella, 2002; Fillipek, 1999; Matson, 2008). These instruments have a long history in the field and are the subject of numerous outside evaluations and studies. The autism screeners are a relatively new breed of assessment, spawned by the drive to identify symptoms of autism in very young children, toddlers and even infants. The STAT, AOSI, Q-CHAT, M-CHAT provide an interesting look at the science of screening instruments which must be both quick and accurate to be truly valuable for the purposes for which they are constructed. Finally, three ancillary instruments, the BPASS, CBSQ, and the SRS, are reviewed to examine how the methods employed in establishing the reliability and validity of these scales differs from other types of scales developed for this population. While there are a number of instruments that are not reviewed here, these that were selected represent a cross section of purposes, styles, formats, and results that are representative of most of the instrumentation that currently exists in the field.

The oldest of the expert-completed autism diagnostic instrument is the Childhood Autism Rating Scale (CARS) (Schopler & Reicheler, 1980; Schopler, Reichler & Renner, 1986). Allen, Robins, and Decker (2008) conducted a small survey (n=117) to ascertain, among other questions, which instruments were being used by school psychologists to diagnose autism spectrum disorders. The CARS was the overwhelming leader. The CARS consists of 15 scales. Each scale address a symptom

commonly associated with autism; however one scale is reserved for the "General Impressions." Clinical judgment is often associated with an accurate diagnosis of autism (Mahoney, et al.,1998; Volkmar, Charwarska, & Klin, 2005), and the CARS builds this feature into the instrument.

The CARS requires direct observation, parent report and in some cases, a review of records. A person trained to use the CARS observes a child and selects from a Likert scale anchored by 1 for "normal" and 4 for "severely impaired." Interestingly, the CARS allows for a .5 report which allows the examiner to score a midpoint between the scale points. The scale yields an overall score with three possible results: Severely autistic (over 36.5), moderately to mildly autistic (30-36.5), and not autistic (below 30).

The strengths of the instrument are its strong psychometric properties, its enduring positive predictive value, the short time of administration, and the minimal training required to use the instrument reliably.

The reliability of the CARS has been established through numerous studies and replication. Three types of evidence are available on the CARS: internal consistency, test-retest reliability, and inter-rater reliability. Alpha is reported as .94 (n=537) and interrater reliability was reported as a correlation coefficient of .71 (Schopler, et al., 1986). A Japanese version of the CARS has also demonstrated good reliability with alpha reported as .87, and an average inter rater reliability of .62 (Tachimori, Osada, & Kurita, 2003).

Validity was established through a correlation between CARS scores and independent clinical judgments yielding $r = .84$, $p<.001$ with clinician's ratings and $r =$

.80, p<.001 when compared to expert judgments made with an independent assessment by a child psychologist and a child psychiatrist (Schopler, et al., 1986).

In addition to the direct studies of psychometric properties by the authors, the CARS has performed well in numerous outside studies and reports (Charak & Stella, 2002; Dilalla & Rogers, 1998; Garfin & McCallon, 1988). Rellini and colleagues (2004) found the CARS to be more reliable than the Autism Behavior Checklist (Krug, et al., 1980) and recommended its use as the diagnostic protocol in Italy. A five site study involving 274 preschool children reported excellent specificity and sensitivity to expert diagnosis (Perry, Condillac, Freeman, Dunn-Geier & Belair, 2005). These more recent studies of the validity of the CARS are important since the DSM description of autism has changed significantly since the CARS was written, and there was some question about whether the CARS could accurately diagnose children under the new criteria (Rellini, et al., 2004), which has since been resolved, with the CARS still performing at acceptable levels in diagnosing children with an ASD.

The manual for the CARS reports an administration time of about 5 to 10 minutes, which is not an insignificant benefit. While some instruments such as the ADI-R require 90 minutes or more to administer and extensive formal training, the CARS can be useful when time is an important factor in the diagnostic process, such as moving children from waiting lists to services where the presence of disability is not in doubt.

The Autism Diagnostic Interview-Revised (ADI-R) and the Autism Diagnostic Observation Schedule (ADOS) are two complementary diagnostic instruments which have evolved over twenty years of development and refinement. The original Autism

Diagnostic Interview (ADI) was developed in 1989 (LeCouteur, Rutter, Lord & Rios). The original Autism Diagnostic Observation Schedule (Lord, Rutter, DeLavore & Risi, 1989) was developed in the same year as a complement to the ADI. These instruments remain closely related. The ADI (currently revised to be the ADI-R) is a parent/caregiver interview. The ADOS is a schedule of activities which allows the evaluator to directly observe behavior which is indicative of autism spectrum disorders.

The ADI-R (Lord, Rutter, & LeCouteur, 1994) is a semi-structured parent/caregiver interview. Unlike the ADOS, the ADI-R does not rely on interactions between the child and the examiner. It is, therefore, easier to administer and does not require the engagement of a child in a task using toys or other stimuli under controlled conditions. However, the ADI-R does require specific training to administer the instrument and can take up to two and one half hours to administer (Western Psychological, 2009).

The ADI-R was revised in 1994 to align with the DSM-IV and ICD 10 descriptions of autism. The original ADI was intended for research purposes for children and adults over the age of five. The demand for an instrument which could be used clinically caused the authors to work to reduce the length of the interview. Additionally, with the advent of clinical use, the ADI-R needed to be more sensitive to possible ASDs in a younger population (Lord, Rutter, & LeCouteur, 1994).

The ADI-R has five sections. The first section is a series of opening questions followed by a section on communication, social development and play, repetitive and restricted behaviors, and general behavior problems.

Surprisingly, the psychometric data were collected on a very small sample of children. The original reliability study of the instrument involved 20 children, 10 of whom had a diagnosis of autism and 10 of whom were identified as "mentally handicapped or language impaired." Each group contained eight male and two female children.

Like the ADOS, extensive training is involved with individuals who administer the assessment. In the 1994 study, graduate students or medical students had spent six weeks engaged in various training activities associated with the ADI-R. Each of the raters for the study achieved agreement with at least one of the original authors at over 90% on three consecutive scorings of the assessment.

Interrater reliability was reported as a weighted kappa statistic, but only on those items which were new to the assessment, (i.e., those items which were written to adapt the assessment to children under the age of five). Weighted kappa statistics were reported for each of the new items and the items which were modified from the original ADI. Exact inter-rater agreement was also provided. The authors note that on three items in particular, even though the weighted kappa statistic was somewhat low (.52, .54, and .59), the exact agreement was over 87%. Thus sensitivity to the range in scores resulted in a lower kappa statistic even though the actual agreement was high (Lord, Rutter & LeCouteur, 1994).

Internal consistency using Chronbach's alpha was conducted on the three domains of communication (.85), restrictive and repetitive behaviors (.69), and social functioning (.95) (Lord, Rutter, & LeCouteur, 1994). It is interesting to note that the low alpha statistic in restrictive and repetitive behaviors has been supported in

subsequent research as being associated with older children (Wiggins & Robins, 2008). It should also be noted that the communication scale was only analyzed on 11 subjects who had three word phrase speech. Overall, especially considering the amount of training that each observer had undertaken, the internal consistency is not particularly high. Also, considering the stature that the ADI-R holds in the community, the small sample size upon which the psychometrics were based is surprising.

The validity study for the ADI-R added 30 additional children: 15 with autism and 15 who were "mentally handicapped with language impairment." Each of the three domains were subjected to a one way ANOVA for between group differences and these differences were reported significant at the <.05 level. Sensitivity of the ADI-R in this study was high (96%) with only one child who was autistic not scoring high enough on the algorithm. Specificity was high with only 2 false positives in the MH/LI group (92%). A recent study of the ADI-R, ADOS and GARS yielded only 75% agreement between the ADI-R and ADOS and a diagnosis of autism from a team of professionals (Mazefsky & Oswald, 2006). (The GARS performed significantly worse than the ADOS and ADI-R.).

Noting the scarcity of psychometric studies of the ADI-R and its rather unearned status as the gold standard in autism diagnostics, a large cross disciplinary team undertook a validity study of the ADI-R and conducted an exploratory factor analysis. The team also examined internal consistency and convergent validity with 226 children with pervasive developmental disabilities. Their findings indicate that while the ADI-R is a useful instrument for differentiating children with autism from children with disabilities that are not on the autism spectrum, they cautioned regarding the use of the

scales for a diagnosis of autism. In particular, measures of internal consistency with this larger group supported the relative strength of the communication domain and also mirrored the original weakness of the repetitive behavior domain. The factor analysis conducted by this group suggested that the emphasis of the instrument is on social impairment. Convergent validity with a variety of instruments reinforced the concept of a multi-disciplinary approach to a diagnosis of autism, relying on both expert opinion and a variety of measures (Lecavalier, et al., 2006).

Research conducted comparing the ADI-R, the CARS and the ADOS-G with clinical expertise, found high concordance between the ADOS and CARS, but not with the ADI-R (Ventola, Kleinman, Pandey, Barton, Allen, Green, et al., 2006).

Another deficit associated with the ADI-R is its sensitivity to autism in very young children (Ventola, et al., 2006). Since the scoring algorithm associated with the test takes into account stereotyped behavior, the ADI-R will miss autism in young children who have as not yet begun to engage in either social relationships or to display circumscribed interests. A comparison between the CARS and ADI-R was conducted by Saemundsen, Magnusson, Smari, & Sigurdardottir (2003). Among their conclusions were that the two instruments measured the construct differently, with the CARS applying a unidimensional framework for autism and the ADI-R measuring three core deficits of autism and applying more weight to the area of social deficit. Their conclusion was that the ADIR/CARS group identified children with more symptomology and lower IQ than a group of children diagnosed with the CARS alone. While their conclusion was that the CARS was a better instrument for the purposes of identification of children with autism in Iceland than the ADI-R, their study highlights

the fact that different measurement instruments are particularly suited to different situations.

Considered by many to be the gold standard of autism diagnostic instruments (Chawarska, Klin, Paul & Volkmar, 2007; deBildt, et al., 2003; Filipek, et al., 1999) the ADOS; (Lord, et al., 2000; Lord, et al., 1989) is a professionally-administered diagnostic instrument for children on the autism spectrum. According to the authors, the original intent of the ADOS was for research purposes; however, a demand from clinicians to use the ADOS in clinical practice precipitated some modifications to the original instrument (Lord, et al., 2000). Lord and colleagues attribute the current form of the ADOS to the original version and another related instrument, the PL-ADOS (Pre-linguistic ADOS) (DiLavore, Lord, & Rutter, 1995).

In contrast to the ADI-R, the psychometric properties of the ADOS are well documented, and the procedures for establishing the reliability and validity of the assessment are detailed and thorough. The ADOS consists of four modules. The modules vary depending on the expressive language abilities of the subject. The examiner is guided to select the module which most closely matches the subject's language ability. Each module was separately analyzed for reliability and validity. While psychometric integrity varies across the modules, each module has strong psychometric properties. It is worth noting that the ADOS does not provide a module which adequately matches with adults or adolescents who are severely affected by the disorder (Lord, et al., 2000).

The ADOS has established reliability through interrater agreement on items and scales, intraclass correlations for inter-rater reliability, internal consistency and test

retest. Items on the ADOS are scaled 0, 1, and 2. On some items, a "3" is available to indicate particularly severe abnormalities. Interrater agreement on items was established both through percent exact agreement and weighted kappa. Inter-rater exact agreement was reported to range from 81% to 100% and weighted kappa from .46 to 1.0 across the four modules. Test-retest was performed on a relatively small sample of children (27) within a nine month time frame. It is reported as an intraclass correlation coefficient across all four domains ranging from a low of .59 on the restricted and repetitive behaviors to a high of .82 in social communication.

Item analysis was accomplished using item total correlations between individual test items, domain scores, and between items and the domain score. Additionally intercorrelations were constructed between items and chronological age and verbal mental age or verbal IQ.

The ADOS presents an interesting consideration in the study of test reliability. Those who use the ADOS are required to be trained on the instrument.

> Prior to their participation in the study, the examiners had observed and coded many "live" and videotaped ADOS sessions. Weekly practice coding sessions were held at the major testing site, in which videotapes were scored and consensus codings for each item were reached. Before officially beginning to collect data, each examiner reached 80% or greater exact agreement with other examiners who were already considered reliable raters. The inexperienced raters continued scoring ADOS sessions and having their codings compared, item by item, with the experienced raters' codings of the same sessions. The 80% agreement criterion had to be met for three consecutive scorings of Modules 1 or 2 (including each module at least once) and three consecutive scorings of Modules 3 or 4 (including each module at least once); the three sessions for each pair of modules had to include at least one session conducted by the inexperienced rater and at least one conducted by another person."(p. 113).

Clearly, care was taken to ensure that fidelity to the scoring algorithms was maintained. There is absolutely no doubt that the ADOS yields a reliable and valid

diagnosis of autism, but the question might be posed as to whether the reliability lies with the instrument itself, or is a result of the extensive training that one receives before he or she is allowed to use it.

The ADOS uses a series of algorithms to determine one of three possible outcomes: Autistic Disorder, PDD-NOS, and nonspectrum diagnoses. Validity was established by comparing the diagnosis of autism from the algorithms with a diagnosis of autism by an expert panel. The authors used a sample of 74 children: 40 with autism, 17 with PDD-NOS and 17 nonspectrum. The 40 children with autism were further divided into two groups: one with low language abilities and one with matched nonverbal intellectual functioning abilities to the PDD-NOS and the Nonspectrum group. The purpose of the intellectual matching was to control for the contribution of cognitive functioning to the overall diagnosis of autism. The authors retained the 20 non-matched group to explore how well the algorithms performed when identifying children who were not matched on cognitive ability. A one way analysis of variance was conducted to test for significant group differences in ADOS scores on all four modules with significant differences across all four groups included in the study.

Cutoff scores for assigning a diagnosis of Autism, PDD-NOS, and autism spectrum disorders was accomplished through an analysis of receiver operating characteristic (ROC) curves. As could be expected, the authors found clearer support for cutoffs between autism and nonautism disorders than between autistic and pdd-nos cutoffs due to the continuous distribution of scores once an autism spectrum disorder was indicated. Sensitivity was reported to range between 86 and 100 across modules and specificity between 68 and 100. Lowest levels of sensitivity were in module four

between the PDD-NOS and the nonspectrum group, while lowest levels of specificity were in Module 3 between the PDD-NOS and the autism group (Lord, et al., 2000).

Since the study in 2002, the algorithms have been revised with improved diagnostic predictive validity being reported (Gotham, Risi, Pickles, & Lord, 2007; Gotham, et al., 2008).

A strength of this instrument is the ability of the ADOS to identify children with an autism spectrum disorder from children without autism and even to differentiate children with classic autism from those with PDD-NOS. However, because of the extensive time and costs associated with administering the ADOS, one would seriously have to justify the necessity of that level of validity in the diagnosis or separation of autism from pdd as important to rely on this instrument as opposed to a CARS or a GARS which can yield an equally reliable indication of autism while requiring far fewer resources. An example of a setting where the ADOS is justified could be in research settings where phenotypic specificity is an important part of the study. On the other hand, using the ADOS for a school-psychological evaluation where behavioral observations and a simpler instrument qualify a child for special education services might not be the most prudent use of resources.

Within the expert/clinician completed domain of assessments, a number of scales function as "screeners." The purpose of a screener is to flag children who are suspected of having an ASD and refer them for further evaluation. Screeners, by nature, should be shorter and easier to administer than diagnostic instruments because their practical use is in widespread applications such as well-child checkups in the

American health care system, or kindergarten readiness screenings in educational or health care settings under IDEA.

According to Fillipek et al., (1991), all children should receive a screener, but only those who fail the screener should take a diagnostic assessment. The Screening Tool for Autism in Two-Year Olds (STAT) (Stone, Coonrod, & Ousley, 2000) and the Autism Observation Scale for Infants (AOSI) (Bryson, Zwaignebaum, McDermott, Rombough and Brian, 2008) are two clinician-administered screening instruments. These instruments are also noteworthy because they are designed to detect autism in very young children.

The Screening Tool for Autism in Two Year Olds (STAT) (Stone, et al., 2000) was developed to detect autism at a very early age. There are two levels of screening instruments used for children with autism (Siegel, 1998; Williams & Brayne, 2006) the first, level one, is used for differentiating children with an ASD from the general population. A level two screening instrument, on the other hand, is one which is used to further identify children from a group of children considered "at risk" due to some anomaly in expected developmental progression. The STAT is a level two screening instrument. The impetus for the development of the STAT was to promote early identification of children with autism.

Reliability for the STAT was measured using a sample of 104 children, 50 with a diagnosis of autism, 15 with PDD-NOS, and 39 with Developmental Delay/language impairment. The authors used test-retest reliability and inter-rater reliability. Interestingly, the test-retest reliability was conducted with different raters from time period one to time period two. In expert/clinician rated evaluations, one must consider

the role of changing evaluators as an extra source of variation within the test/retest framework. Additionally, the authors report a mean time between the test and retest with a range from 4 to 44 days. On the positive side, however, kappa statistics for the interrater agreement were conducted during both the test and the retest and resulted in a coefficient of 1.00 and .90. It is interesting that the authors eschewed traditional measures of reliability such as coefficient alpha and a Pearson correlation between the test and retest measures.

Validity for the STAT was established through convergent/concurrent validity between the ADOS-G and the STAT. Again, a kappa statistic was applied.

The authors make the statement that the STAT has reasonable psychometric properties; however, only one coefficient (kappa) was derived for both test-retest reliability and for inter-rater reliability. On the positive side, the STAT had excellent sensitivity for those students who were identified as having a diagnosis of autism using the ADOS-G. Every student who was diagnosed with autism was correctly identified as at high risk by the STAT. However, in contrast, of 15 children who were diagnosed with PDD-NOS, seven (47%) were identified in the "low risk" category of the STAT. This indicates that the sensitivity of the STAT is less than what would be useful as a general population screener.

A recent entry to the assessment literature is the Autism Observation Scale for Infants (AOSI) (Bryson, et al., 2008). This scale was developed to detect early signs of autism in siblings of autistic probands. The central problem in the development of this scale is to reliably identify behaviors which are indicative of autism at such an early age. This was accomplished through retrospective analysis of videotapes of autistic

children prior to their first year. Initial measures of reliability were conducted on children as young as six months of age. The AOSI measures behaviors across the domains of social development, sensory-motor responses and repetitive behaviors. The scale is intended mostly for research purposes to document the earliest signs of autism (Bryson, et al., 2008).

The AOSI sample size is small, and while the study reports that interrater reliability was measured on 32 infants at age six months, 34 infants at 12 months and 26 infants at age 18 months, it is important to understand that many of these infants were the same child. Interrater agreement was reported as an unweighted kappa statistic and ranged from perfect agreement on a number of subtests including visual tracking at six months to -.05 on social interest/affect at age six months. The majority of kappa coefficients were above .60. Test retest reliability was measured using 20 infants two weeks apart after their 12 month visit. Test retest reliability was reported as an intraclass correlation coefficient and ranged from .61 to .68. A unique feature of this scale is the combination of binary (present/absent) scoring for some items and a Likert-scale scoring of 0-3 on other items. The authors report that the total scale score is the most robust for predicting autism. Perhaps with a bigger sample, item analysis could reveal the contribution of items to the total score in the form of item total correlations to enhance our understanding of the individual item contributions and more fully understand the contribution of items which can be weighted more heavily than others.

It should also be noted that this scale has not been tested on nonautistic infants. It would seem that this would be a necessary step in confirming the validity of the scale, beyond just a positive predictive value based on this high risk population. In all cases,

while the identification of children with autism as early as six months could have important and powerful effects upon early intervention opportunities and outcomes, all of the standard measures of reliability when applied to a very small population of children with a higher probability of achieving clinically significant scores than the normal population should be interpreted with caution.

A third type of assessment used with children (and in this case adults) on the autism spectrum are ancillary assessments which measure either the co-morbid presentation of behavior associated with autism spectrum disorders or those whose purpose is to further refine the subtypes of autism within the spectrum. Two such measures are reviewed here: the Broader Autism Phenotype Symptom Scale (BPASS) (Dawson, et al., 2007) and the Children's Social Behavior Questionnaire (CSBQ) (Luteijn, Luteijn, Jackson, Volkmar & Minderaa, 2000).

A scale which measures the degree to which characteristics of the broader autism phenotype exist within the immediate families of autistic probands is the Broader Phenotype Autism Symptom Scale (BPASS) (Dawson, et al., 2007). This scale was developed specifically to measure the autism-related traits in nonautistic family members in families who have more than one child with autism. The purpose of this scale is for the specific contributions that it can make toward the definition of an endophenotype of autism for use in genetic studies. Among the scales reviewed the BPASS is notable for using some of the more sophisticated statistical techniques to establish the psychometric properties of the scale, such as HLM analysis for between group differences on the basis of sex and controlling for IQ in the analysis of scale performance.

The BPASS consists of an interview, but is conducted by a trained investigator who can probe symptoms more deeply when a suspected atypical response is provided to the questions. The BPASS was administered to 690 individuals from 201 multiplex families of children with autism. Inter rater reliability was reported as an ICC on four domains ranging from a low of .71 on Conversational skill to a high of .95 for social motivation. Internal consistency was reported as Chronbach's alpha, again on the four domain scores, yielding a range of .76 for social motivation and .91 for expressiveness.

In order to study the heterogeneity of individuals who receive a diagnosis of PDD-NOS. an interesting assessment is the Children's Social Behavior Questionnaire (CSBQ) (Luteijn, et al., 2000). The current CSBQ is a revision to an initial version of the scale produced in 1994. This scale proposes to measure subtypes within the broader category of PDD-NOS. The study employed five groups of children: 240 children with a diagnosis of PDD-NOS who did not meet the criteria for either Aspergers Disorder or Autistic Disorder, 95 children described as "high functioning autistic" and181 children with ADHD, a clinical control group of 400 children with a variety of psychiatric disorders other than an ASD, and 234 normal controls. Children with intellectual disability were excluded from the samples.

The authors subjected a 96 item, 3-point Likert scale to a factor analysis and found a 5 factor solution to be the best fit for the data. The factors were "Acting Out" "Social Contact Problems," "Social Insight Problems," "Anxious/Rigid" and "Stereotypical."

Reliability measures included an estimate of internal consistency (Chronbach's Alpha) for each scale which ranged from .76 for Scale 5, "Stereotypical," and a high of

.92 for Scale 1, "Acting Out." Interrater reliability was reported as a Pearson r obtained by having both parents of one child independently completing the scale. Correlation coefficients ranged from .64 to .85. Test retest reliability was computed using a sample of 21 mothers at an interval of 4 weeks. Intraclass correlation coefficients ranged from .32 to .90. The scale, "Stereotypical" was the low performer. Each of these statistics was reported on both the subscales and the overall scale.

Evidence of validity for the scale was established through concurrent validity measures (Pearson's' *r*) between the CSBQ and the Autism Behavior Checklist (ABC). Further a discriminate analysis was used to examine the extent to which the scales could correctly place children from each of the 5 groups. Specificity was 50%, indicating that about half of the children could be correctly identified. While this statistic is somewhat disappointing, nonetheless, the scale has some merit in parsing out subgroups within the broad and highly diverse group of children who are diagnosed with a PDD-NOS. Good subtyping of students in this category could contribute to research on specific phenotypes of PDD-NOS as well as to focus on interventions which could be more applicable to one group than another.

Caregiver Completed Diagnostic Scales

There are numerous advantages to collecting data using caregiver scales. According to Lecavalier (2005), these scales can save time and money and can provide normative information. Like the instruments which require a trained clinician to administer, this group of instruments is comprised of diagnostic instruments, screeners and ancillary assessments.

Among the oldest of these instruments is the Autism Behavior Checklist (ABC) (Krug, Arick & Almond, 1980). While Charak and Stella (2001) classify the ABC as a screener, the authors published the ABC as a diagnostic instrument. The ABC is also notable as one of the first scales to use teacher observations (Volkmar, et al. 1989). The ABC consists of 57 questions grouped in five areas: sensory, relating, body/object use, language and social and self-help. The ABC is also interesting because although all of the questions are yes/no scored, some questions are weighted more heavily in the scoring scheme than others. A total score (based on weighted sums) yields a score and these scores are divided into ranges which fall into high probability of being autistic, questionably autistic and unlikely autistic.

As a testament to the longevity of the ABC, no less than four separate, factor analytic studies have been undertaken on its behalf (Eaves & Williams, 2006; Miranda-Linne & Melin, 2002; Volkmar, et al., 1988; Wadden, Bryson, & Rodger, 1991). Wadden, Bryson and Rodger's (1991) study suggested a three factor solution and suggested that the 57 items of the ABC could be reduced to approximately the 38 with highest weightings. Miranda-Linne and colleagues (2002) suggested five factors, but not the same ones that comprise the current subscales, and Eaves and Williams (2006) found support for both a four and five factor solution. These studies are important for a number of reasons: first, they highlight the fact that even as our analytical procedures become more sophisticated, the design of the study can have dramatic effects on the solutions. Further, as researchers continue to define the borders of the disorder and phenotypic variation with the spectrum, how researchers investigate the attributes which yield "factors" has a strong influence on how we view the disorder in general.

Validity of the ABC has been established through discriminant analysis. The discriminant analysis is reflected in a specificity and sensitivity statistic. These studies used various cutoff scores to measure the number of students correctly identified with an ASD. When the categories of both "probably" and "questionably" autistic are combined, the ABC yielded a sensitivity of 80%.

Volkmar and colleagues were the first to identify significant problems with the ABC reliability and validity (1988). In a study of 157 individuals, 94 with autism and 63 without, the ABC over identified 36% of students who did not have an ASD from a population of students with some sort of disability (Volkmar, et al., 1988). These problems with over identification prompted the authors to recommend the ABC as a screening instrument, but recommended strong reservations with using the ABC as a diagnostic instrument.

Rellini and colleagues (2004) compared the ABC with the CARS in a sample of 65 children using DSM-IV criteria and found that the ABC failed to identify 46% of the sample that did have an ASD. These children were correctly identified by the CARS. This study is notable because the ABC was developed when the DSM-III-R criteria for autism was employed by those making diagnoses. This particular version of the DSM was found to be over-inclusive of autism in general (Sponheim, 1996), thus it is not surprising that when the more stringent criteria of the DSM-IV were applied, the instrument would not perform as well.

A second, well-known caregiver completed diagnostic instrument for children with ASD is the Gilliam Autism Rating Scale (GARS) (Gilliam, 2006). The Gilliam Autism Rating Scale-2 is a revised edition of the original GARS published in 1995.

Since the publication of the original GARS, numerous studies have challenged the psychometric properties of the original scales (Lecavalier, 2005; Mazefsky & Oswald, 2006; South, et al., 2002). The GARS-2 has been largely reworked to address problems with over-identifying children as having an ASD who didn't.

The assessment is constructed in three parts: a four choice Likert scale which can be completed by parents or another individual with current knowledge of the person with autism, a parent interview, and a series of open ended questions which are designed to guide the application of one of the possible ASD diagnoses outlined in the DSM-IV-TR.

The Likert-type scales are structured around the autism triad of the DSM-IV, communication, social interaction and restricted and repetitive interests or behaviors. There are a total of 42 items ranging from "Never observed" to "Frequently Observed." Each Likert response is assigned a score of 0 to 3 and these are totaled to yield an Autism Index which provides an overall score reflecting the severity of autism symptoms displayed by the individual.

The second part of the GARS-2 is a parent interview. This section of the assessment collects developmental history which can be helpful in arriving at a specific ASD diagnosis. This part of the assessment is a series of yes or no questions. The final section collects information about medical history and particular parental concerns.

Validity for the GARS-2 is established through three lines of evidence: convergent validity, theoretical foundations of the DSM-IV criteria, and an evaluation of the discriminant ability of the assessment with a group of non-ASD children with other disabilities, and a non-disabled control group.

Convergent validity was assessed by comparing the results of the GARS-2 with the Autism Behavior Checklist (ABC). Both assessments were administered to 63 parents of children. Five subscales of the ABC were matched to the GARS-2 subscales yielding correlations ranging from .56 to .78 with significance at the <.01 level.

Unlike most of the assessments reviewed here, the GARS-2 has a section which lists instructional objectives which can address the symptoms of autism noted by the assessment. These objectives are meant to be used in instructional planning and provide a much needed link between assessment and intervention (Montgomery, Newton, & Smith, 2006). The GARS-2 is relatively new and further studies of its reliability and its validity (especially with regard to the use of instructional objectives) is warranted.

While the ABC and the GARS-2 function as diagnostic instruments, screeners which utilize parent/caregiver report are increasing. One of the first screening tools developed for use in the primary care setting of the British health system is the Checklist for Autism in Toddlers (CHAT) (Baron-Cohen, Allen, & Gillberg, 1992; Baron-Cohen, et al., 1996). It is worth noting that the differences in the health care systems in the two countries have a substantial influence on the use of screeners. The managed-care movement in American health systems severely limits the amount of time that physicians are encouraged to spend with patients, and therefore time restraints are of considerable importance in the American system and somewhat less important in the British health care system. While the original CHAT did not report a specific administration time, it was estimated that the CHAT required approximately 10 minutes to administer (Charak & Stella, 2002).

The original instrument had a 38% sensitivity, considered very low, but a high positive predictive value (83%). These statistics indicate that while those children who were identified as being at risk for an ASD with the CHAT were correctly identified, there were a significant number who were not picked up by the screener who later went on to get an ASD diagnosis.

Largely in response to the under-sensitivity, the CHAT underwent extensive reorganization in 2008 and has been since renamed the Q-CHAT (Quantitative Checklist for Autism in Toddlers) (Alison, et al., 2008). Interestingly, one of the first revisions was to remove the "expert" part of the assessment, because the authors found that the assessment was equally reliable with only the parent report, and the clinician observation could therefore be shortened without loss of sensitivity. However, rather than shortening the instrument, a number of questions were added so that the total number of questions rose from 14 to 25. The second major revision was to change the yes/no responses to a five point Likert-scale with scoring of 0-4. It is noteworthy that scale point titles differ depending on the question, with such anchors as "very easy" to "impossible," and "many times a day" to "never." This implies that the frequency of a behavior as well as the quality of the behavior are both measured in the same scale.

One of the more impressive qualities about the initial Q-CHAT study is the number of individuals in the study. Questionnaires were sent to 2,360 families in two health districts in Great Britain. Of these questionnaires, 754 were returned and included for study in the non-diagnosed group. A comparison group of 160 children with an ASD diagnosis made up the second group. The authors employed a t test for between group differences and were found significant ($t(912)=-.31.1$, $p = <.0001$).

For reliability, Chronbach's alpha was computed for the non-ASD group (.67) and for the ASD group (.81). Item total correlations were measured with 4 items falling below an acceptable level. Test retest for both groups was reported as an ICC of .82 ($p < .0001$). Interestingly, test-retest was also measured using a paired samples t-test with no significant difference between test pairs.

At this time, positive predictive value and negative predictive values are not available for this instrument. Further follow up studies must be conducted in order to see how well the initial screening either identified or failed to identify ASD in this very young population (Allison, et al., 2008).

The authors of the Q-Chat point out an important point in interpreting reliability and validity statistics on autism assessments, and that is the composition of the test group. Sensitivity will naturally be higher in a group where the traits are always present, as opposed to groups where large numbers of neuro-typical children are also included. Assessments which use only clinically referred individuals on which to gather the sensitivity data should be viewed with caution.

The Modified Checklist for Autism in Toddlers (M-CHAT) is a screening instrument which relies on parent report on 23 yes/no items (Robins, Fein, Barton & Green, 2001). The M-CHAT is notable for a number of reasons. First, it was developed to screen for autism at an early age, generally between 16 and 30 months. The M-CHAT is also interesting because the first 9 questions of the 23 item scale are taken directly from the Childhood Autism Scale for Toddlers (CHAT) (Baron-Cohen, 2000). The original CHAT was developed for use as a screening instrument in the

British health system, relying on a home health component that does not have a corollary in the American system.

The M-CHAT was developed for use in the well-child checkup system recommended by the American Academy of Pediatrics. The instrument consists of 23 yes/no questions and relies on parent report versus physician observation. Children who scored above the cutoff for suspicion of an ASD received a follow-up phone call to reduce the number of false positives. With the follow up phone call, reliability for the M-CHAT was established through a large scale (n= 1,293) study. Of these children, 58 were identified by the instrument as requiring further evaluation for potential developmental delay. Of these 58, 39 were eventually diagnosed with an autism spectrum disorder.

Internal consistency was reported as coefficient alpha = .83. The authors employed a discriminant functional analysis to analyze items which contributed the most information toward an ASD diagnosis. The authors found the M-Chat to have a specificity of .99, sensitivity of .89 and a positive predictive power of .80. The M-CHAT has many features which make it a good screening instrument. The psychometric properties are within an acceptable range, the instrument is quick and easy to complete, and the instrument's sensitivity is high enough to function as a good safety net for children at risk of an ASD.

Paralleling the information gathered through expert/clinician observed scales, there are ancillary scales which rely on parent/caregiver information. The Social Responsiveness Scale (SRS) (Constantino, 2004) is an example of a scale that was designed not so much as a direct diagnostic measure, but as a measure of the single

construct of reciprocal social behavior. The author has conducted a good deal of research on the theory that autism is a unidimensional construct of social reciprocity which is normally distributed in the population. Additionally, he views the other DSM characteristics of communication deficits and repetitive and restricted patterns of interest as being largely explained by the degree to which an individual's social behavior is impacted (Constantino, et al.,2003; Constantino & Todd, 2000). This somewhat unidimensional theory of autism has some support in the autism measurement literature (Baron-Cohen, 2002; Spiker, Lotspeich, Dimiceli, Myers, & Risch, 2002).

The SRS illustrates the link between theory and validity in scale development. Constantino's theory of autism is one of a normally distributed trait which reaches clinical significance in a few individuals (those diagnosed with an ASD). However, using the scale for purposes other than investigating the unidimensional theory of autism entails a buy-in on the part of the end user to this particular theory. Thus, while the SRS is a very interesting scale with acceptable psychometric properties, its uses are circumscribed by the underlying theory on which it is constructed.

Another interesting feature of the SRS is the treatment subscales. The instrument generates scores in five domains: receptive, cognitive, expressive, and motivational aspects of social behavior. The manual claims that these scores can be helpful in designing treatment programs; however, the narrow focus of the SRS on social aspects of autism could ignore some of the other pervasive areas of need such as self help/adaptive skills.

The literature surrounding the measurement of autism provides a body of evidence supporting the use of screening instruments, diagnostic instruments, and related measures of autism symptomology. Most of the instruments reviewed here have adequate psychometric properties in terms of reliability. Additionally, the extent to which screeners can detect children with autism from a normal population and diagnostic instruments can detect autism spectrum disorders from other developmental disorders and the extent to which autism spectrum disorders can be sub classified accurately support the validity of the use of the scores from these instruments.

Summary

The assessments in this review demonstrate that while the construction of assessments and the establishment of psychometric properties is a penultimate goal, there is wide variability in how the authors approach test development, reliability and validation.

Common Measures of Reliability

Twelve assessment instruments for use in a population of children with autism were reviewed. Approximately half of the instruments (7/12) provided some measure of internal consistency. Most instruments reported this as coefficient alpha, but the STAT used a kappa statistic. In addition, for those tests that employed subscales, an internal consistency statistic was reported on each subscale.

Inter-rater reliability was reported for seven of the instruments. Methods for computing inter-rater reliability varied between kappa, weighted kappa, intra class correlation coefficients and percent of exact agreement. Some of the instruments employed both.

Test-retest was reported for five instruments. The reason for the limited use of this powerful reliability coefficient could lie more in resources than in any deliberate avoidance of gathering the data. Mostly, test-retest reliability requires two contacts with the same individual after a period of time. It is evident that to accurately assess the psychometric properties of an assessment for wide scale use, resources are not an insignificant factor.

Validity

All of the assessments relied on the literature on ASDs to formulate the basis of the assessments. Most of the assessments considered the DSM/ICD formulations of the disorder that are composed of a triad of deficits in communication, social reciprocity, and restricted and repetitive areas of interest. An interesting finding from the synthesis of these psychometric data is that subscales which measure restrictive and repetitive behavior generally suffer from lower performance on all psychometric measures (internal consistency, inter-rater reliability, sensitivity, and specificity). Specifically, both the ADI-R and the AOSI subscales measuring restrictive and repetitive behaviors have noticeably lower coefficients than the rest of the subscales within the respective instruments. This points to an age effect with the diagnosis of ASD's and that as the age of diagnosis is getting younger, the use of measuring restricted and repetitive behaviors contributes less to the reliability and validity of early diagnosis.

Sensitivity and specificity were reported for most of the screeners. This area is where the performance of screeners face the greatest challenge, especially for instruments like the STAT, CHAT, M-CHAT and AOSI that are attempting to push the envelope in terms of how young symptoms of autism can be detected. Screeners that

are overly sensitive lose their value if they consistently identify children for further evaluation when further evaluation is not warranted. A more serious problem, however, are screeners that fail to identify children who eventually do receive an ASD diagnosis. The development of these instruments is likely to continue as we improve in our ability to trace the development of ASDs in very young children.

Convergent validity was used to develop validity arguments for six of the scales. In some cases, such as the GARS-2 and the STAT, convergent validity was established through comparing scores on other assessments. The other instruments compared the results of the assessments with expert clinician opinions.

Construct Representation

Until biological markers for autism are identified, behavioral observations must serve as proxies for definitive diagnoses. What traits these scales choose to measure is the area where the greatest amount of variability and in some cases, creativity, exists. The CARS casts the widest net with 14 separate questions which do not form subscales. Most of the assessments have some questions relating to communication and social reciprocity, many also include questions which address unusual sensory responding. Only the ABC has a specific group of questions that address self help skills, and the SCBQ is unique in its "Acting Out" scale. Three scales have questions that address anxiety (CARS, SCBQ, and the SRS). In summary, our understanding of the disorder is both illuminated and limited by the constructs within which we choose to measure it.

Samples

The samples that were used to establish the psychometric properties of the scales reviewed varied from an *n* of 26 infants for the 18 month old portion of the AOSI to 912

reported for the Q-CHAT. Larger samples generally included a number of children who were not affected by autism. Some studies included children who were only affected by the disorder, such as the AOSI which was only administered to infants in multiplex families of autism. Other studies such as the Q-Chat included vast numbers of children who were screened during routine physical examinations.

The composition of samples is not inconsequential. Samples such as those used for the BPASS and the AOSI consisted purely of individuals who came from first degree relatives of individuals with autism. The sample for the revised items of the AID-R was drawn from a group of ASD children and a group of children with other non spectrum disabilities. One the one hand, including children without a diagnosis of an ASD is very important to test the validity of screeners; however, in terms of yielding a reliability statistic, it is a much more difficult task to place students in the same rank order when the differences among them are smaller (Thorndike, 2005). Therefore, it is important when evaluating an assessment to consider not only the size of the sample, but also its composition and to consider that in relation to the overall purpose of the assessment.

Sources of Data

Data were collected from observation and report. All of the expert-administered assessments relied on direct observation, while the other scales used teachers, parents or other caregivers as data sources. The CARS and the ADI-R rely on both sources of information. Interestingly, the original CHAT was revised to exclude the expert observation portion since the authors found the instrument to be just as reliable without this section. The expert-administered scales also required training, and in many cases

(ADOS, ADI-R, AOSI, BPASS) that training is extensive. An interesting study would be to look at the difference in sensitivity and specificity between scales which require extensive training of experts and those that rely on parent/caregiver report. This cursory investigation does not reveal any significant differences in reliability.

Scaling

Scaling was generally in a binary format (yes/no, present/not present) and on Likert Scales ranging from three to five choices, and in some cases both. Frequency of behavior was measured as well as severity of behavior. All of the scales employed classical test theory methodology of adding up total scores. (The ABC applied weights to different score items.) Two scales (ADOS and BPASS) used ROC estimates to determine scale cut offs. Virtually no IRT methods were used in scaling, scoring, or development of any of the measures.

Relationship to Instruction

Of the instruments reviewed here, only two, the GARS-2 and the SRS have instructional material provided to remediate identified deficits. For the most part, while some of the instruments can provide data to be used in instructional planning, that is not their primary purpose. In terms of Fillipek et al.'s (1999) observation that assessment should in some way have a bearing on instruction, it would seem that as of yet, that connection has not been adequately addressed.

The Need for the MN-VASS

With the proliferation of tests to measure ASDs, it is prudent to revisit the question of whether yet another instrument should be developed for this population. Clearly the existing instruments currently provide a number of reliable choices for

autism diagnostics; however, there appears to be a need for additional scales that could help address the gap between assessment and intervention.

The MN-VASS is a care-giver completed assessment that helps to summarize the strengths and challenges of an individual child on the autism spectrum on one page using a visual representation. The main purpose of the MN-VASS is to help educators to prioritize and establish areas for remediation and intervention across 14 domains of behavior commonly reported to be atypical in children with ASD. The MN-VASS is comprised of 14 subscales (one for each behavioral domain). The user rates a series of descriptions arranged on a Likert scale. The MN-VASS computer program converts responses to a mean response for each scale and produces an elliptical shape representing the child's strengths or challenges in that area. It produces a small ellipsis for what is perceived to be a strength for the child and a large ellipsis in an area that is a challenge for the child.

The MN-VASS is different from the instruments reviewed in this Chapter in a number of important ways. First, the MN-VASS is neither a screener nor a diagnostic instrument. As such, the emphasis of the scales and the subscales is not to discriminate between a symptom that children have or do not have, but rather, assuming that all symptoms can be present to some degree, characterize the individual according to a unique profile of strengths and challenges in regard to the specific presentation.

In terms of validity, the use of scores from the instruments reviewed here is mainly for the purpose of identification and diagnosis. The purpose of the MN-VASS is for educational programming and skill remediation. To the extent that the MN-VASS accurately identifies skill deficits and strengths and can be used to communicate those

strengths to stakeholders (teachers, parents, speech pathologists, occupational therapists, behavior therapists), the use of the scale output could be considered valid.

Educators have the daunting task of serving the whole child. The needs of a child with autism are not just contained in the essence of those attributes which are a function of his autism, but also because of the nature of the many co-occurring behaviors and individual presentations. Because the MN-VASS is different in its purpose and because the skill set covered in the MN-VASS includes not only those behaviors that separate children with ASD, but also the co-morbid symptoms that educators must also be responsible for remediating and teaching, the MN-VASS is not a duplicate of any assessment or scale; it is a unique tool for a growing population of children who require individualized intervention and education.

CHAPTER 3

Methods

Participants

Sixty-two individuals who work with children with autism took part in this study. Cumulatively, these professionals rated 105 children diagnosed with an autism spectrum disorder. Eighty-eight boys and seventeen girls were rated using the MN-VASS.

The individuals who participated in the study were behavior therapists or special education teachers. For purposes of clarity, these individuals are referred to as "respondents." All teachers who responded were licensed special educators. Behavior therapists were recruited through two large providers: Partners in Excellence and the Minnesota Autism Center. All behavior therapists are supervised by licensed psychologists. Table 1 illustrates the characteristics of the respondents including their position, gender, education, and years of experience with children on the autism spectrum.

Table 1

Respondent Information on Position, Gender, Level of Education
n = 62

Respondent Information	Number	%
Position		
Licensed Teacher	14	23%
Behavior Therapist	48	77%
Gender		
Male	9	15%
Female	53	85%
Level of Education		
Bachelor's Degree	51	82%
Higher than Master's	11	18%
Years of Experience with Children with ASD		
Less than 1 Year	7	11%
1 to 3 Years	23	37%
4 to 5 Years	18	29%
5 to 10 Years	11	18%
Over 10 Years	3	.05%

The objects of measurement in this study were children with an autism spectrum disorder between the ages of 2 and 11 who were receiving either special education services through a public school district or participating in an intensive early intervention program through a private provider. Children who were measured from private behavior therapy organizations have a medical diagnosis of an autism spectrum disorder, such as Autistic Disorder, PDD-NOS, or Asperger's Disorder. Children who were rated from the public school programs have an educational diagnosis of one of the Autism Spectrum Disorders. An educational diagnosis is a less stringent qualification and does not require the evaluation of a licensed mental health professional or medical

doctor. Table 2 describes the diagnostic categories of the children measured by the MN-VASS, as well as the number of students who were drawn from private behavior therapy and the number of students who were drawn from school-based special education programs.

Analyses of the psychometric properties of the scales under study are reported with aggregate statistics for the entire group (n = 62).

All of the participants listed in Table 1 participated in the study by completing the MN-VASS on line. From this initial pool of participants, a small number participated in further, associated measures of validity and reliability (e.g., test-retest reliability, interrater reliability, convergent validity). All of these participants were recruited from Partners in Excellence. Twenty two respondents participated in the inter-rater reliability study. Eleven respondents participated in the convergent validity study and 12 respondents participated in the test-retest study. A survey was administered as part of the on line MN-VASS program. Forty-nine respondents participated in this portion of the study.

A between groups analysis of teachers and behavior therapists was conducted to assess if differences in responses could be attributed to the type of position that the respondent held. Additionally, an analysis of variance was conducted to determine if differences in the level of experience with children with ASD had an influence on how respondents rated children using the MN-VASS.

Table 2

Child Information – Age, Gender, Diagnosis, & Placement
n = 105

Child Information	Number	%
Age		
2	3	3%
3	11	10%
4	30	29%
5	17	16%
6	14	13%
7	16	15%
8	8	8%
9	6	6%
Gender		
Male	88	84%
Female	17	16%
Diagnosis		
Autism	28	27%
PDD-NOS	9	9%
Asperger's	1	1%
Autism Spectrum Disorder	55	52%
Other	12	11%
Placement		
Public School	14	13%
Behavior Therapy Program	91	87%

This study did not require any direct interaction with children. All of the data were from respondent report and collected via the MN-VASS website as part of the participant information and user registration.

Instrumentation

Two instruments were used to collect data in this study. The first is the MN-VASS, the subject of this study, and the second was the Childhood Autism Rating Scale

(CARS) (Schopler, et al., 1980). The CARS was used to establish convergent validity. The psychometric properties of the CARS are discussed in detail in Chapter 2.

The Minnesota Visual Autism Symptom Scale

The Minnesota Visual Autism Symptom Survey (MN-VASS) was developed with three purposes in mind: 1) to help teachers and program supervisors develop a brief, general synthesis of a child's strengths and challenges across 14 domains of behavior, 2) to assist teachers and other professionals in communicating these strengths and challenges between different teachers and therapists, and 3) to communicate with parents about perceived strengths and challenges of their child. The MN-VASS could also be used as a research tool to examine the co-occurrence of particular symptoms, to gain an understanding of which symptoms are more commonly reported than others, and to examine whether any behavioral profile patterns are present.

The MN-VASS is an observer-completed assessment composed of 14 subscales which measure the relative severity of a symptom commonly considered in teaching children with an autism spectrum disorder. The instrument was developed on the basis of the Diagnostic and Statistical Manual IV-TR (APA, 2000) description of autism spectrum disorder and a body of other autism literature (which is explored within the discussion of each subscale). In broad terms, the DSM-IV-TR parses the symptoms of autism into three areas of deficit: social development, language, and restricted interests. Within these three areas, commonly referred to as the autistic triad, the DSM describes behaviors which would provide evidence for applying the diagnostic label.

The DSM-IV-TR criteria form the basis of eight of the 14 MN-VASS subscales. Table 3 shows the DSM-IV-TR criteria for ASD and the number and title of the MN-VASS subscales developed from them.

Table 3

DSM-IV-TR Criteria to MN-VASS Subscale

DSM-IV-TR Criteria	Includes	MN-VASS Subscale
Social Interaction		
	Non Verbal Behavior	Susbscale 4- Imitation Subscale 10 - Facial Referencing
	Peer Relations	Subscale 12 - Sociability
Language		
	Lack of or delay of language	Subscale 8 - Manding Subscale 9 - Communication Complexity
Restricted Patterns of Interests and Behaviors		
	Non functional routines	Subscale 5 - Rigidity
	Stereotypic behavior	Subscale 7 - Stereotypy

In addition to these criteria, seven additional MN-VASS subscales were developed to address commonly reported co-occurring areas of need not explicit in the DSM-IV-TR. Each of the MN-VASS subscales is described, and empirical support is provided for its inclusion in the scale.

The 14 symptoms measured by the MN-VASS are toileting, eating, dressing, imitation, rigidity, sensory responding, stereotypy, manding, communication complexity, facial referencing, sociability, activity level, challenging behavior, and safety. During the development of the MN-VASS, each subscale was analyzed to provide separate measures of internal consistency, and items within each subscale were evaluated using item total correlations. The subscales are grouped into areas of intervention. Each of the subscales derives support for its validity from a body of

evidence suggesting that it is either an overt behavior commonly associated with an ASD or that it is a construct in and of itself associated with ASD.

Subscales 1 – 3: Toileting, Eating, & Dressing: The Self-Help Subscales

Subscales 1 through 3 probe three areas of adaptive functioning: toileting, eating and dressing. These adaptive skills are often lagging in children with ASD (Carter, et al., 1998; Lis et al., 2001; Rodrigue, Morgan & Geffken, 1991). Klin and colleagues investigated the relationship between adaptive skills as measured by the Vineland in 187 males of ages 4 through 18 with autism. Their findings support the instruction of adaptive skills in this population due to skill deficits (Klin, et al., 2007). Additional research found that including a measure of adaptive functioning actually enhanced sensitivity in the ADOS from 75% to 84% (Tomanik, Pearson, Loveland, Lane, & Shaw, 2007).

Issues surrounding toilet training and children with autism are commonly reported. Bartak and Rutter (1976) conducted a study comparing the toileting habits of neurotypical children, children with autism, and children with intellectual disability. These early results indicated that a) children with autism often were well behind the neurologically typical children in acquiring independent toileting. Further they found no significant differences between the children with autism and the children with intellectual disability. One study conducted by Darymple and Ruble (1992) surveyed 100 parents of children with autism, with a mean age of the children of 19.5 years and found that toilet training averaged about 2.1 years later than typical peers and that numerous individuals with autism were still not toilet trained at the time of the study (when they were 19 years old). Needless to say, the attainment of independent toileting

skills is a major factor for inclusion in general education settings and should be an emphasis in a comprehensive program for intervention.

In addition to toileting issues, another area of adaptive functioning in which children with autism are reported to present difficulties relate to feeding issues. Specifically children with autism have been reported to have food sensitivities, self restricting diets, and other deficits related to self-feeding (Schreck, Williams & Smith, 2004). Williams, Dalrymple, and Neal (2000) documented issues surrounding children with autism and eating habits and noted that eating habits were heavily influenced by environmental stimuli and included abnormal behaviors such as smelling, throwing and licking food, as well as ingesting non-food items. Similar to toileting, eating behaviors are teachable skills which can become a part of a comprehensive intervention program.

Another adaptive skill that is important for gaining independence is dressing. Dressing is often accomplished through occupational therapy and the skills are documented to be needed in this population (Klin, et al., 2007; Myers & Johnson, 2007; Watling, Deitz, Kanny, & McLaughlin, 1999). These skills are essential to helping children with autism keep up with their non-disabled peers, reduce parental stress (Lecavalier, Lenone, & Wiltz, 2006) and function independently.

Subscales 4, 5, & 11 – The Social Behavior Subscales

Subscale 4 measures imitation skills. These skills are often found deficient in children with autism (DSM-IV-TR, 2000; Vivanti, Aparna, Ozonoff & Rogers, 2008). While the presence or absence of these skills can be a marker for autism, these skills are teachable, and are included as an essential component in many early intervention programs such as the Assessment of Basic Language and Learning Skills - Revised

(Partington, 2006), The Denver Model (Rogers, Hall, Osaki, Reaven & Herbison, 2000) and Behavioral Interventions for Young Children with Autism (Maurice, et al., 1996).

Subscale 5 measures rigidity. Rigid behavior is most often associated with the "restricted, repetitive behavior" as noted in the DSM-IV-TR (APA, 2000; Lewis, Tanimura, Lee & Bodfish, 2006; South, Ozonoff & Mcmahon, 2007). Rigid behaviors can often lead to challenges when integrating an individual with autism into a classroom or community setting and are therefore important to consider when designing a treatment plan or educational programming for a particular child. Rigid behaviors are often addressed through behavioral approaches (Kuhn, Hardesty & Sweeney, 2009; Myers & Johnson, 2007) and are appropriate targets for intervention. At the same time, many children with autism do not display rigid patterns of behavior, and it is equally important to recognize this as an area of strength for those children. While this is true for almost all of the skills covered by the MN-VASS, rigidity has been documented so pervasively in the literature that a teacher or therapist could be led to assume that all children with an ASD are rigid, without taking the time to evaluate the individual child.

Subscale 11 measures sociability. Many researchers have defined social reciprocity and other social deficits as the core impairment in ASDs (Constantino, et al., 2004; Mundy, Sigman & Kasari, 1990) and it is one of the primary indicators of autism in the DSM-IV-TR (APA, 2000). Both Kanner (1943) and Asperger (1944) noted that the children who were originally identified as autistic were socially remote. The term "autism" implies a withdrawal from others. Wing (1988) described three types of individuals who all qualify for a diagnosis of autism, but seem to fall into rather distinct groups: aloof, passive, and active but odd. These references to social functioning

provide further evidence that a measure of sociability is important in tailoring a child's educational programming. On the other hand, there are a good number of children with ASD who are socially awkward, but not socially aloof (Siegal, 1996). This subscale seeks to differentiate those students, and as a result, provide a basis for where social skill interventions might begin. For example, a small ellipsis (potential strength) on the MN-VASS output would indicate a child who wants to socialize, and therefore might need different instruction or curricula than a child with a large ellipsis (greater need), indicating that a more fundamental approach to socialization would be prudent.

Subscales 6 and 7 – The Sensory Subscales

Rinner (2000) observed that although there is currently no sensory component directly listed in the DSM-IV-TR for ASDs, the inclusion of "odd responses to sensory stimuli," does imply that sensory responding is a factor to be considered with this population. Ayres (1979) proposed that sensory integration is a problem which affects children with autism, and also includes children with learning difficulties. Additional research and anecdotal reporting also support the inclusion of sensory issues in assessment and treatment of autism (Anzalone & Williamson, 2000; Grandin, 1995; Rinner, 2000).

While this research supports the argument that some children with autism often respond to sensory stimuli in atypical ways, it is equally important to document those who do not. Assuming that all children with ASDs have sensory challenges could potentially lead to treatments that are superfluous to the individual child's needs. An awareness of how sensitive a child is to noise, light, and tactile stimuli can help educators provide intervention when needed and focus on other areas of need if it is not.

Subscale 7 measures stereotypy. Stereotypic behavior is associated with ASDs (DSM-IV-TR, 2000; Goldman, et. al, 2009). The scale is written to evaluate the conditions under which a child engages in stereotypical behavior. This scale does not emphasize the form of the stereotypy, but rather the contextual surroundings in which it occurs which could imply the function served by the behavior (Cunningham & Schreibman, 2004). Research conducted by Yianni-Coudurier and associates (2008) found that stereotypical behaviors were a barrier to inclusion of children with disabilities in inclusive classrooms. To the extent that these behaviors interfere with learning and inclusionary opportunities, they can be targets for behavioral intervention.

Subscales 8, 9, & 10 - The Communication Subscales

Subscale 8 measures the child's ability to get his or her needs met through appropriate communication. A mand is generally a request for something such as food, attention, escape, or information. The term "mand" is a term used by B.F. Skinner (1959) in his explanation of Verbal Behavior. According to Skinner's theory of verbal behavior, a mand is the only verbal operant which is generated by an establishing operation. Functional communication training is often a form of mand training. This training has shown good results in eliciting communicative behavior and in decreasing challenging behavior among members of this population (Harding, et al., 2009; Langdon, Carr, & Owen-DeSchryver, 2008). A great deal of positive behavioral support literature attributes a reduction in problem behavior when appropriate requesting behavior is learned (Dunlap, Carr, & Horner, 2008; Horner, et al., 2005).

Subscale 9 is a general measure of the sophistication of the child's communication. Communication difficulties are a hallmark of ASDs (DSM-IV-TR,

2000). The scale ranges from age-appropriate speech to hand leading and challenging behavior as the form of communication used by the child. While many children require intensive teaching in communication, it could be helpful for educational staff to understand the skills that the child already utilizes in this area so as to capitalize on their current abilities and begin instruction from an appropriate level tailored to that child.

Subscale 10, facial referencing, indicates how well a child modulates eye contact. Eye contact is noted to be deficient in children with autism (DSM-IV-TR, 2000). Much nonverbal information is conveyed through facial expressions, and teaching children to attend to another's facial expressions and interpret them can be a helpful basic skill in improving social functioning (Attwood, 2006; Westphal & Volkmar, 2008).

Subscales 12, 13 & 15 - The Behavioral Subscales.

Subscale 12 appraises the child's activity level. It is interesting to note that the International Classification of Diseases-10 (ICD) (World Health Organization, 1990) contains a pervasive developmental disorder known as "Overactive disorder associated with mental retardation and stereotyped movements" which does not have a parallel in the DSM-IV-TR (2000). The diagnostic guidelines site "inappropriate severe overactivity" as one of the criteria. There is extensive documentation of treatment for over-activity in the autism literature (Burack, Ennis, and Johannes, 1997; Farmer, Hollway, Arnold, & Med, 2008; Handen, Johnson, & Lubestsky, 2000;). Instructional strategies for these children should be different than for those children who can sit still for various amounts of time. Knowing that a child has high activity levels could suggest a number of interventions and strategies that would help the child such as

frequent breaks, special seating, and instructional strategies that have been found useful in teaching children with ADHD (Chronis, Chacko, Fabiano, Wymbs & Pelham, 2004; Miranda, Presentacion, & Soriano, 2002; Stein, Szumowksi, Blondis, & Roizen, 2006).

Subscale 13 assesses the degree to which the child engages in non-compliant or challenging behavior. Children with autism can engage in challenging behavior for a number of reasons, such as insufficient language and delayed social development (Buschbacher & Fox, 2003; Donnellan, Mirenda, Mesaros, & Fassbender, 1984; Horner, et al., 2002). Often challenging behavior can be ameliorated through positive behavioral supports and behavioral techniques which reinforce appropriate behavior and extinguish inappropriate behavior (Carr, et al., 1999; Dunlap & Fox, 1999; National Research Council, 2001). Teachers and behavior therapists often must teach children appropriate learning behaviors in order for content instruction to occur.

Subscale 15 of the MN-VASS measures safety behaviors. Disregard for personal safety, elopement, and aggression are documented in this population (Anckarsater, 2006; Debbaudt, 2003; Koegel, Stible, & Koegel, 1998; Perrin, Perrin, Hill & DiNovi, 2008; Matson & Rivet, 2008; McDougle, Stigler & Posey, 2003). Especially when transitioning between teachers, this information can help a teacher to take the extra precautions to keep children from endangering themselves. Safety should also be a priority for teaching.

The subscales focus on skills that are appropriate for instruction in either a classroom or a behavioral day treatment program. Other symptoms associated with an ASD that are not typically ameliorated through education were not included, such as sleep disturbances (Patzold, Richdale & Tonge, 1998; Richdale, 1999; Taira, Taira, &

Sasaki, 1998), gastrological disturbance (Goldberg, 2004; Horvath & Perman, 2002; Molloy & Manning-Courtney, 2003), and anxiety (Bellini, 2004; Gillott, Furniss, & Walter, 2001; Kim, Szatmari, Bryson, Streiner, & Wilson, 2000).

The composition and navigation of the questions of the MN-VASS were written according to the recommendations of Dillman (2000). Responding to the items of the MN-VASS requires that the user select a response from among 5 choices on a Likert-type scale. Each scale has 4 responding points: "Strongly Agree," "Agree," "Strongly Disagree," "Disagree," and an option for "not observed." Each symptom of autism that is measured with the MN-VASS forms a subscale. The subscales fell in two categories: those that were developmentally arranged and those that were not. Developmental subscales are toileting, eating, dressing, imitation, manding, communication complexity, facial referencing, and activity level. Non-developmental subscales are rigidity, sensory responding, self-stimulatory behavior, behavioral challenges and safety.

Responses to the MN-VASS are scored differently depending upon the type of subscale. Subscales that are developmental in nature begin by asking the user to respond to a question which poses the highest level of functioning for that particular subscale. For example, the statement, "The student uses age-appropriate speech," is developmentally the highest response in the subscale measuring communication complexity. Developmental arrangement was determined using the Hawaii Early Learning Profile (Furuno, 2004) and the Preschool Developmental Profile (D'Eugenio, 1998). Items for the MN-VASS were compared to both checklists to determine the

order in which they were presented, i.e., most advanced to least advanced. A copy of the content of the MN-VASS is provided in Appendix A.

The purpose of the developmental arrangement of the items on the subscale is to take advantage of computer technology in streamlining the survey and reducing the amount of time a respondent spends in answering questions. If the user responds with a "Strongly Agree" response, the software assigns the smallest value to the output and moves the respondent to the next scale. The assumption here is that if a student can use age-appropriate speech, he or she could perform all of the other tasks that are contained in the subscale. This is a strategy that is commonly employed in a wide variety of psychometric assessments.

When the response to the first item in each developmental subscale is *not* "Strongly Agree," the respondent is required to answer a series of questions that comprises the remainder of the sub-scale. The mean of the responses to the items in the subscale is used as the score on that subscale. This is the same algorithm used when the subscale is not developmental in nature. Responses of "not observed" are not included in determining the mean for the subscale.

Questions in the MN-VASS are worded either positively or negatively. Positively worded questions are those in which the response "Strongly Agree" indicates more of a symptom of autism than the response "Strongly Disagree." Negatively worded questions are those in which the response "Strongly Disagree" indicates more of a symptom of autism than the response, "Strongly Agree." For example, the statement, "This child becomes distressed if changes are made to his routine," is positively worded because a response of "Strongly Agree" indicates more rigidity, which is a symptom

measured by the MN-VASS. On the other hand, the statement, "This child can ask another person to stop doing something the child finds annoying," is a negative question, since "Strongly Agree" in this case would indicate less inability to request, another symptom being measured by the MN-VASS.

Scoring the MN-VASS is accomplished through the scale's software. The software assigns a numerical value to each Likert response. The mean response of the item is returned in the output as an elliptical shape. The size of the shape is commensurate with the evaluation of the symptom by the respondent. If the respondent endorses selections that indicate that the symptom is not present in the behavioral repertoire of a particular child, the program assigns the smallest value to that subscale, and the output reflects the smallest shape next to that particular symptom. An example of a MN-VASS output is provided in Appendix B.

It is important to note that the MN-VASS output reflects the rater's perception of the child's skills on each of the subscales. The extent to which the rater is familiar with the child, the propensity of the rater to either select extreme Likert responses (i.e., strongly agree) versus middle Likert responses (agree) will heavily influence the appearance of the MN-VASS output. As a note, reliability measures such as test-retest and coefficient alpha, are affected by these variables, since it is assumed that raters who possess these characteristics will do so consistently over time.

Childhood Autism Rating Scale

While the focus of this study is the development of the MN-VASS, the Childhood Autism Rating Scale (CARS) (Schopler, et al., 1980) was used to establish convergent validity for the MN-VASS. The CARS is a diagnostic scale for assessing

the presence and severity of autism. The psychometric properties of the CARS are well established and detailed discussion of the scale is provided in Chapter 2. The CARS is similar to the MN-VASS in a number of ways: both scales are intended for use in the ASD population and both employ subscales measuring specific behaviors. The CARS, however, is used primarily as a diagnostic instrument, is paper-based, and does not produce graphical output. In addition, all of the MN-VASS subscales measure behavior which is directly teachable.

Procedures

Data Collection

The MN-VASS is hosted on a web site with the URL, www.visualautismsurvey.net. All respondents used the website to complete the MN-VASS from a personal computer. The MN-VASS software establishes an account for each respondent. The respondent supplies a user name and password. The software records responses to each question on the MN-VASS which can be exported in a comma delimited format for further analysis. Other than the person who has started the account, only the programmer and the author have access to these data.

Behavior therapists were recruited from two large private providers in the Twin Cities. The first provider coordinated the completion of the scales during regular business hours in a center based program. A designated computer was used and the respondents each logged on to the internet site, established an account and rated from one to five children who were on their caseload. The second provider directed lead behavior therapists to access the site within a given time frame and complete the MN-VASS on at least one of their clients.

Teachers in the study were recruited by e-mail and accessed the website from their own computers. Most of these teachers were contacted directly from e-mail lists for teachers of children with autism by the author. Other teachers were recruited from referrals by local autism experts and autism program coordinators.

A subset of therapists was recruited to complete the CARS. A different (but somewhat overlapping) subset of therapists was asked to re-evaluate their student using the website and adding an "r" to the end of the student's name. The "r" in the database identified those responses as a re-test response.

Data were collected for the validity sub-study via respondent interview with the examiner on site at Partners in Excellence.

Data Analysis

The Standards (AERA, 2004) emphasize that the two most crucial characteristics of an assessment are validity and reliability. Data from the MN-VASS were analyzed according to a number of procedures which are commonly used to provide evidence of reliability and validity.

Reliability

Four measures of reliability of the MN-VASS were used in this study. First, Chronbach's Alpha was calculated as a measure of internal consistency. This measure is reported for the overall scale and each subscale. Test-retest reliability is reported as a Pearson correlation coefficient on a subset of students. Inter-rater agreement is reported as both a weighted and unweighted kappa coefficient from a subsample of respondents.

Item Analysis

Item analysis for the scale was accomplished by analysis of item total correlations.

Validity

The composition of the 14 scales is based on a review of literature which supports each of the subscales as an important domain of behavior associated with educational programming for children with autism. Further evidence supporting the validity of the MN-VASS is garnered through three methods: convergent validity, survey responses to questions that probe end user's opinions of the usefulness of the survey output, and a small study on how well the MN-VASS output characterizes the actual student from whom the profile was drawn.

Convergent validity evidence was gathered by having 12 therapists complete both a CARS and a MN-VASS on a child that they teach. While the MN-VASS and the CARS differ in their intent (the CARS is primarily a diagnostic instrument), both purport to measure a number of behaviors which when codified and summed, yield a numerical score which indicates "more" of the construct of autism. Thus higher scores on the CARS should correspond to high scores on the MN-VASS. A correlation coefficient was calculated on the matched scores of the CARS and MN-VASS.

An additional series of questions was added at the end of the MN-VASS in its online format in order to gain feedback about the usefulness of the MN-VASS as a tool to help guide the planning of instruction for children with ASD. Summary statistics are provided.

In the final validity study, 12 behavior therapists were asked to select from a field of three MN-VASS profiles, the profile of one specific student with whom they

had worked for a minimum of 6 months. As evidence of validity, the hypothesis under study in this project was that respondents would be able to differentiate the MN-VASS profile of a student they served from among the profiles of different students rated by different therapists.

The process employed in this study involved respondents being asked to come into a private room with the author. The author sat across from a table with the field of three profiles placed face down on the table. The profiles were oriented toward the respondent. The position of the target child's profile was altered from the far left to the center to the right for each respondent. The first respondent's target child's profile was placed on the far left in a field of three, the second respondent's target child's profile was placed in the center of the field of three and the third respondent's target child's profile was placed on the far right of the field of three. This rotation was continued for each respondent.

The examiner directed the respondent to turn over all of the profiles and then to select his or her target child from the field, "Please select the profile that most closely matches (child's name) and give it to me." The examiner was sure to look at the respondent and not at the field of profiles. After the respondent gave the profile to the examiner, the examiner thanked the respondent and asked them to leave the room. The examiner recorded the response as either correct or incorrect. The examiner then set up the next three profiles for the next respondent and repeated the procedure.

CHAPTER 4

Results

In this chapter, the psychometric properties of the MN-VASS are described. Results of item analysis, measures of reliability, and measures of validity are reported. Four measures of reliability are reported: internal consistency as coefficient alpha, internal consistency as split-half reliability, test-retest reliability, and inter-rater reliability. Support for the scale's validity is provided through a measure of convergent validity with an existing scale, responses to a survey on the utility and accuracy of the MN-VASS, and the results of a quasi-experimental test of face validity.

The data for this study were collected from July of 2009 through December of 2009. All data were collected via the MN-VASS website, located at www.visualautismsurvey.net. One hundred five entries, each representing a unique child, were included in the final analysis of the data. These assessments were conducted by 63 respondents, thirty one of whom assessed more than one child. The number of assessments per respondent was capped at three, with 12 respondents assessing three children.

Item Analysis

Item analyses were conducted using corrected item total correlations. A preliminary examination of the items was conducted using the same items selected for the analysis for internal consistency. In addition, because the MN-VASS measures 14 separate domains of behavior, each with a subscale, the item total correlations were conducted within the subscales. The main purpose in conducting item analysis is to

determine the best selection of questions which will yield the highest reliability with the fewest questions. The item total correlation coefficient is an indicator of how well an item is performing; however, it is a guide more than a strict rule. By convention, item total correlations exceeding .3 are considered acceptable. However, when there is a conceptual justification for retaining the item, the developer who is familiar with the construct being measured should make the final decision on retaining or eliminating an item.

Items below are coded with the first two characters representing the subscale of the item, the next two characters identifying the question number within the subscale, a designation of "h" for a header question and an underscore and the number of the item in the overall scale. Thus, item s5q1h_23 is the first item in Subscale 5; it is a header item; and it is the twenty-third item in the scale. The output produced by SPSS generates a "what if" calculation of coefficient alpha if the item were to be deleted from the scale. This information helps the scale developer to determine the effect on internal consistency that removing the item from the scale would have. Coefficient alpha was reported at .85 for the overall scale.

Table 4

Item Total Correlations for Subscale Headers and Non-Developmental Items

Item Number	Corrected Item Total Correlations	Alpha if Item Deleted
s1q1h_1	.45	.90
s2q1h_7	.47	.90
s3q1h_13	.47	.90
s4q1h_19	.67	.90
s5q1h_23	.58	.90
s6q1_29	.43	.90

Item Number	Corrected Item Total Correlations	Alpha if Item Deleted
s6q2_30	.25	.90
s6q3_31	.34	.90
s6q4_32	.32	.90
s6q5_33	.33	.90
s7q1h_36	.54	.90
s8q1h_41	.71	.90
s9q1h_48	.70	.90
s10q1_53	.46	.90
s10q2_54	.39	.90
s10q3_55	.23	.90
s10q4_56	.50	.90
s11q1h_58	.61	.90
s12q1_65	.03	.91
s12q2_66	-.08	.91
s12q3_67	.36	.90
s12q4_68	-.01	.91
s12q5_69	.36	.90
s13q1_70	.61	.90
s13q2_71	.69	.90
s13q3_72	.53	.90
s14q1_73	.50	.90
s14q2_74	.50	.90
s14q3_75	.50	.90
s14q4_76	.43	.90
s14q5_77	.40	.90
s15q1_78	.54	.90
s15q2_79	.44	.90
s15q3_80	.48	.90
s15q4_81	.41	.90
s15q5_82	.38	.90
s15q6_83	.41	.90

Using this data, Items 55, “This child will talk to you, but won’t look at you without a prompt,”; 65, “This child is constantly on the go,”; 66, This child will not sit still while eating or other activities that most children at a similar age are able to sit

while doing,"; and 68, "This child seems like he/she never gets tired," are under-performing. While eliminating Scale 12, Activity Level, from which most of these items come, would eliminate the negatively contributing items; their absence would not dramatically increase the MN-VASS reliability. The Alpha if item deleted only increased by one one-hundreth. Scale 12 measures the activity level of a child and gives the teacher or therapist information that could be helpful when planning the type of activities that will help this child learn best. In addition to the conceptual argument for retaining Scale 12, subscale analysis produced item total correlations between .43 for Item 69, "This child is sluggish and will only engage in physical activity when prompted," and .79 for Item 79, "This child is constantly on the go."

The table which follows shows the item total correlations for all of the items on the MN-VASS arranged by subscale. There are two estimates of the item total correlations for each item. The first estimate represents the scale item when all of the possible cases are included. This would also include those respondents who selected a "Strongly Agree" response to a developmentally arranged scale. Item total correlations are also reported for the subgroup of cases who did not respond with 'Strongly Agree." To conduct this analysis, cases were selected if the response to the header item in each of the developmental scales was greater than one. This eliminated those respondents who had 1s filled in their data for the remaining subscale questions. Therefore, the item total correlations that are reported by subgroup are a more conservative estimate of item functioning and internal consistency than those reported for the aggregated group. For the subscales that are not developmentally arranged, there is no need to conduct a subgroup analysis. There are no skipped items in these scales; each item is answered by

the respondent, therefore all cases are included in the analysis. Item analyses were conducted on the 83 items of the MN-VASS broken down by subscale. All correlations were greater than .30, suggesting that all items in the scale are functioning acceptably.

Table 5

Item Total Correlations for the MN-VASS by Subscales

	Subscale	Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
Item Total Correlations for Subscale 1, Toileting					
1	This child uses the toilet independently.	.88	.94	.78	.90
2	This child can adequately wipe him/herself after using the toilet.	.74	.96	.53	.93
3	This child is toilet trained for urination.	.89	.94	.86	.85
4	This child is toilet trained for bowel movements.	.85	.94	.73	.90
5	This child is successful on a toileting schedule.	.84	.94	.82	.89
6	This child is not potty trained.	.91	.94	.87	.88
Item Total Correlations for Subscale 2, - Eating Skills					
7	This child can cut meat into bite sized pieces with a knife and fork.	.63	.76	.88	.95
8	This child can pour liquid into a cup from a larger container	.70	.74	.78	.96
9	This child can spread butter on a piece of toast.	.74	.73	.88	.95
10	This child can feed himself/herself with a fork.	.55	.78	.88	.95
11	This child can get a drink from a water fountain.	.30	.82	.84	.95
12	This child independently drinks from a cup that does not have a lid.	.53	.78	.91	.94
Item Total Correlations for Subscale 3, Dressing Skills					
13	This child can get dressed independently.	.68	.73	.67	.74

Subscale		Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
14	This child can tie her or his own shoes.	.40	.81	.33	.83
15	This child can manage most fasteners.	.64	.76	.55	.76
16	This child can put on his/her own winter coat.	.62	.75	.56	.76
17	This child can put on and take off his/her shoes.	.68	.75	.68	.74
18	This child does not participate in his or her own dressing and requires extensive adult support.	.36	.81	.60	.75
Item Total Correlations for Subscale 4, Imitation Skills					
19	This child will spontaneously imitate the actions of others.	.72	.86	.54	.86
20	This child will imitate an action performed by another child when prompted.	.83	.84	.80	.81
21	This child will imitate an adult when the adult asks the child to perform an action and models it for the child.	.75	.86	.75	.83
22	This child does not imitate.	.39	.90	.58	.85
23	This child can copy simple drawings.	.71	.87	.54	.87
24	This child can imitate a sequence of two or more actions.	.86	.83	.79	.81
Item Total Correlations for Subscale 5, Rigidity					
25	This child is flexible and responds well to changes in his or her routine or schedule.	.74	.77	.62	.69
26	This child adheres to a rigid routine.	.70	.78	.62	.68
27	This child becomes distressed if changes are made to her/his routine.	.78	.76	.70	.66
28	This child becomes distressed if unexpected people show up.	.53	.82	.44	.73

Subscale		Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
29	This child likes to wear the same clothes to school every day.	.40	.84	.31	.76
30	This child engages in nonfunctional routines.	.48	.83	.32	.77
Item Total Correlations for Subscale 6, Sensory Responding					
31	This child squints or covers his/her eyes in florescent lighting…	.46	.62	--	--
32	This child will often cover his/her ears in the presence of a moderately loud ….	.36	.66	--	--
33	This child resists being hugged or patted.	.33	.66	--	--
34	This child smells objects, food, or people more frequently than a typical child.	.44	.62	--	--
35	This child is strongly influenced by the texture of his or her food.	.56	.55	--	--
Item Total Correlations for Subscale 7, Stereotypy					
36	This child does not engage in any noticeable stereotypic behaviors beyond what would be considered normal for his/her age.…	.72	.96	.68	.95
37	When alone, this child engages in hand flapping, tapping, rocking, finger flicking or other stereotypy.	.94	.90	.93	.86
38	In the company of others this child engages in hand flapping, tapping, rocking, finger flicking or other stereotypy.	.94	.90	.93	.87
39	During a demanding task or when bored this child engages in hand flapping, tapping, rocking or other stereotypy.	.85	.93	.81	.91
Item Total Correlations for Subscale 8, Manding					
40	This child can ask for information.	.74	.86	.66	.85
41	This child can ask another person to perform an action.	.80	.85	.75	.84

Subscale		Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
42	This child can ask for help.	.76	.86	.69	.85
43	This child can ask another person to stop doing something the child finds annoying.	.76	.86	.66	.85
44	This child has many things he/she can ask for using words.	.67	.87	.61	.86
45	This child can ask for things he/she needs by using sign or pictures. (Do not select if the student uses a more sophisticated method of manding.)	.34	.90	.51	.87
46	The most common way this child requests is through crying…	.67	.87	.66	.85
Item Total Correlations for Subscale 9, Communication Complexity					
47	This child uses age-appropriate speech to communicate.	.66	.82	.45	.75
48	This child can use two word combinations to communicate.	.75	.79	.73	.64
49	This child uses one word utterances or phrases that function as one word to communicate.	.53	.85	.40	.78
50	This child can echo what you say.	.72	.80	.70	.67
51	This child can ask and answer "wh" questions.	.66	.82	.46	.75
Item Total Correlations for Subscale 10, Facial Referencing					
52	This child looks at you from time to time when he/she is communicating with you.	.69	.69	--	--
53	This child actively avoids making eye contact.	.65	.68	--	--
54	This child will talk to you, but will not look at you without a prompt.	.51	.75	--	--
55	This child will look at your face to see your emotional state…	.55	.75	--	--

	Subscale	Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
Item Total Correlations for Subscale 11, Sociability					
56	This child plays imaginatively with other children.	.70	.90	.53	.87
57	This child plays with other children.	.85	.87	.80	.83
58	This child is interested in other children his own age.	.82	.88	.79	.83
59	This child seems interested in adults in the environment.	.66	.90	.59	.86
60	This child ignores other children.	.82	.88	.77	.83
61	This child will actively avoid other children.	.63	.90	.56	.86
62	This child will ignore adults in his/her environment.	.56	.91	.47	.87
Item Total Correlations for Subscale 12, Activity Level					
63	This child is constantly on the go.	.79	.69	--	--
64	This child will not sit still while eating	.61	.75	--	--
65	This child enjoys climbing and physical play	.47	.79	--	--
66	This child seems like she/he never gets tired.	.62	.75	--	--
67	This child is sluggish and will not engage in physical activity unless prompted.	.43	.80	--	--
Item Total Correlations for Subscale 13, Challenging Behavior					
68	This child follows directions.	.56	.81	--	--
69	This child accepts "no" for an answer	.60	.80	--	--
70	This child will leave a preferred activity with an adult request without engaging in challenging behavior.	.68	.79	--	--
71	This child will allow other children to enter his/her space and use the same materials or sit next to them.	.66	.78	--	--

Subscale		Aggregate Item Total Correlations	Alpha if Item Deleted	Subgroup Item Total Correlations	Alpha if Item Deleted
72	This child will go along with a group of children (such as down to the gym/motor room) without engaging in challenging behavior.	.68	.78	--	--
73	This child will separate from significant adults at an age appropriate level without engaging in challenging behavior.	.38	.84	--	--
Item Total Correlations for Subscale 14, Safety					
74	This child will respond to a strong command from an adult such as "No!" "Stop!" or "Hot!" to prevent the child from engaging in dangerous behavior.	.46	.75	--	--
75	This child will elope from familiar places (like his or her house or school).	.55	.72	--	--
76	This child will bolt in dangerous places such as in malls or parking lots.	.66	.69	--	--
77	This child can tell a stranger from a familiar person and will respond differentially to them.	.50	.74	--	--
78	This child will behave with appropriate caution around animals.	.46	.75	--	--
79	This child will climb on furniture	.45	.75	--	--

Non-Developmentally arranged subscales do not have item total correlations for the non aggregated groups. All cases were included.

Reliability Measures

Coefficient Alpha and the item analyses undertaken for the MN-VASS were adjusted to compensate for the computer-programmed skip patterns in the scale. Nine of the 14 subscales of the MN-VASS are developmentally arranged. If the respondent replies "strongly agree" to the first question on a developmentally arranged subscale, the program automatically fills in 1s in the data, and presents the first question of the next scale. The addition of 1s to the data artificially inflates coefficient alpha. To account for this inflation, this analysis uses only the header questions from the scales which are developmentally arranged (Scales 1, 2,3,4,5,7,8,9, and 10). This reduced the number of items from 83 to 35.

Internal Consistency

Internal consistency, measured as coefficient alpha was computed with all questions on the MN-VASS and with only header questions and nondevelopmental questions. This was done to mitigate artificial inflation due to computerized scoring of skip patterns. Overall Alpha for the full version of the MN-VASS was .96. Alpha for the scales with the subscale responses removed was .89 with 35 items, suggesting good reliability for the scale.

Table 6

Coefficient Alpha by Subscale, Aggregated and Disaggregated

Scale Number	Scale Name	Number of Items	Aggregate Alpha	Cases	Sub group Alpha	Cases
1	Toileting	6	.96	95	.91	68
2	Self Feeding	6	.81	95	.96	68
3	Dressing	6	.81	94	.81	77
4	Imitation	6	.89	90	.87	67
5	Rigidity	6	.83	92	.76	82
6	Sensory Responding	5	.68	91	--	--
7	Self Stimulatory Behavior	4	.94	89	.93	78
8	Manding	7	.89	88	.87	73
9	Communication Complexity	5	.85	88	.77	74
10	Facial Referencing	4	.80	88	--	--
11	Sociability	7	.90	89	.87	79
12	Activity Level	5	.79	89	--	--
13	Challenging Behavior	6	.82	89	--	--
14	Safety	6	.77	89	--	--

Seven of the scales reviewed in this study employed an alpha coefficient as evidence of internal consistency. Table 7 shows the MN-VASS in comparison to the other instruments.

Table 7

Comparison of Coefficient Alpha for Selected Scales

Instrument	Coefficient Range
MN-VASS	.89
Adaptive Subscales	.81- .96
Behavioral Subscales	.76 - .93
Communication Subscales	.77 - .87
CARS	.94
ADI-R	
Communication	.85
Restricted & Repetitive Behaviors	.69
Social Functioning	.95
ADOS	
Social Affect	.87 - .92
Restricted & Repetitive Behaviors	.51 - .66
BPASS	.76 - .95
CSBQ	.76 - .92
Q-CHAT	
ASD Group	.81
Non-ASD Group	.67
M-CHAT	.83

Split Half Reliability

A split half coefficient expressed as a Spearman-Brown corrected correlation was computed for the MN-VASS. The scale was split into two halves by moving entire scales such that the first half of the analysis contained Scales 1 through 7, and the second half contained Scales 8 through 14. The value of the split-half coefficient was .90, indicating satisfactory reliability.

A split half coefficient was also computed for each of the subscales in the MN-VASS. The coefficients are reported for the entire data set and for only those cases in

which the response to the header question was greater than 1, “Strongly Agree.” This split half procedure placed every other item in the subscale into the respective halves for analysis.

Table 8

Split Half Coefficients by Subscale

Scale Number	Scale Name	Correlation Coefficient (all cases)	Correlation Coefficient (cases where header > 1)
1	Toileting	.93	.88
2	Self Feeding	.69	.71
3	Dressing	.74	.77
4	Imitation	.81	.86
5	Rigidity	.75	.63
6	Sensory Responding	.61	--
7	Stereotypy	.94	.93
8	Manding	.78	.79
9	Communication Complexity	.93	.89
10	Facial Referencing	.79	--
11	Sociability	.88	.83
12	Activity Level	.85	--
13	Challenging Behavior	.69	--
14	Safety	.72	--

Test-Retest Reliability

Test-Retest reliability was calculated on a small number of children (n=22). The first test was administered in August of 2009 and the retest administered in December of 2009. A Pearson correlation yielded a coefficient of .90 indicating good test-retest reliability. Table 9 shows test-retest reliability coefficients for each subscale. The lowest correlation was between the Facial Referencing Subscale (.27) and the Challenging Behavior Subscale (.97).

Table 9

Test – Retest Subscale Coefficients

Scale Number	MN-VASS Scale Name	Correlation Coefficient
1	Toileting	.82*
2	Self Feeding	.47
3	Dressing	.87*
4	Imitation	.87*
5	Rigidity	.53
6	Sensory Responding	.66
7	Stereotypy	.93*
8	Manding	.79*
9	Communication Complexity	.90*
10	Facial Ref.	.27
11	Sociability	.49
12	Activity Level	.79*
13	Challenging B.	.97*
14	Safety	.87*

Test-retest coefficients for the MN-VASS subscales have a wide range of values. Seven of the subscales have good test-retest reliability with coefficients at or exceeding .80, and nine of the subscale correlations were significant. However, five of the subscales did not achieve a significant correlation and one-half did not achieve .80. Self feeding, Rigidity, Sensory Responding, Facial Referencing, and Sociability were not significant, with Activity Level, and Manding achieving a significant correlation at .79. Challenging Behavior and Stereotypy had the highest test-retest correlations at .97 and .93 respectively.

Test-retest coefficients were calculated for five of the other scales reviewed in this study. The MN-VASS and the CARS both used a Pearson r, while the ADOS, CSBQ, and Q-Chat employed an interclass correlation coefficient. The ADOS reported ICCs for each of the subscales. The STAT used a kappa statistic of inter rater agreement and then correlated the statistic between the first administration between two raters and then the second.

Table 10

Test-Retest Correlations for Selected Scales

Instrument	Method	Coefficient
MN-VASS	Pearson *r*	.90
Adaptive Subscales		.47 - .87
Behavioral Subscales		.53 - .97
Communication Subscales		.27 - .90
CARS	Correlation Coefficient	.77
ADOS	ICC	.59 - .82
STAT	Kappa*	.90
CSBQ	ICC	.82
Q-CHAT	ICC	.82

Inter-rater Reliability

The MN-VASS requires the respondent to select the degree to which the subject displays or does not display the behavior or characteristic described in each item. To that extent, it is of interest to determine the degree to which a set of ratings from one observer resembles a set of ratings from another observer (Wiggins, 1973). For each pair of raters, a percent of exact agreement was calculated across subscales (Hartmann, 77). Table 11 provides the range and the average of the percent of exact agreement by subscale.

Table 11

Percent Exact Agreement for 11 Raters by Subscale

Scale Number	Scale Name	Mean	Range
1	Toileting	70%	33 – 100%
2	Self Feeding	79%	50 – 100%
3	Dressing	81%	60 – 100%
4	Imitation	80%	17 - 100%
5	Rigidity	78%	67 – 100%
6	Sensory Responding	60%	20 – 100%
7	Stereotypy	81%	0 – 100%
8	Manding	78%	50 – 100%
9	Communication Complexity	82%	60 – 100%
10	Facial Referencing	61%	0 – 100%
11	Sociability	87%	71 – 100%
12	Activity Level	67%	20 – 100%
13	Challenging Behavior	78%	50 – 100%
14	Safety	63%	0 – 100%

Percent exact agreement varied widely between raters. The variability among the rater pairs gives ranges from perfect disagreement on three of the scales to perfect agreement between at least one rater pair on each scale. Inter rater agreement was determined for seven of the scales reviewed for this study. Percent exact agreement was reported for the ADI-R and the ADOS.

Table 12

Inter rater Agreement for Selected Scales

Instrument	Method	Coefficient
MN-VASS	Exact Agreement	60% - 87%
CARS	Correlation Coefficient	.71
ADI-R	Weighted Kappa	.52 - .59
	Exact Agreement	87%
ADOS	Weighted Kappa	.46 – 1.0
	Exact Agreement	81% - 100%
STAT	Kappa	.90
AOSI	Kappa	-.05 – 1.0
BPASS	ICC	.71 - .95
CSBQ	Pearson *r*	.64 - .85

Measures of Validity

The composition of the 14 scales is based on a review of literature which supports each of the subscales as an important domain of behavior associated with educational programming for children with autism. Further evidence supporting the validity of the MN-VASS is garnered through four methods: scale intercorrelations, convergent validity, survey responses to questions that probe end user's opinions of the usefulness of the survey, and a small study on how well the MN-VASS output resembles the actual student from whom the profile was drawn.

Scale Intercorrelations

Scale intercorrelations allow us to examine the relationships between the dimensions being measured. Subscale total scores for each student were computed and then these totals were correlated.

Table 13

Interscale Correlation Coefficients

Scale		1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Toilet	1													
2	Eat	.47	1												
3	Dress	.65	.58	1											
4	Imitate	.49	.39	.53	1										
5	Rigid	.16	.21	.00	.36	1									
6	Sensory	-.01	.26	.08	.22	.47	1								
7	Stereotypy	.21	.27	.28	.52	.37	.29	1							
8	Mand	.53	.43	.51	.76	.47	.26	.51	1						
9	Comm Complexity	.47	.39	.50	.74	.40	.29	.46	.89	1					
10	Facial Referencing	.09	.24	.11	.41	.42	.22	.39	.34	.33	1				
11	Social	.38	.38	.32	.68	.52	.30	.61	.71	.67	.49	1			
12	Activity	-.08	-.02	-.00	.12	.10	.13	.06	.04	.09	.04	.08	1		
13	Behavior	.39	.34	.24	.51	.58	.33	.30	.51	.50	.35	.50	-.01	1	
14	Safety	.26	.43	.43	.51	.37	.38	.37	.52	.52	.32	.50	-.19	.49	1

Using the convention of .5 and above as a strong correlation, .3 to .49 as a moderate correlation and below .3 as a small correlation, the following scales display strong, moderate or weak magnitudes of relation:

Table 14

Strongly Correlated Subscales

1	Toileting	Dressing	Manding			
2	Self Feeding	Dressing				
3	Dressing	Imitation	Manding	Complexity		
4	Imitation	Stereotypy	Manding	Complexity	Social	Behavior
5	Rigidity	Sociability	Behavior			
6	Sensory Responding					
7	Stereotypy	Manding	Sociability			
8	Manding	Complexity	Sociability	Behavior		
9	Complexity	Sociability	Behavior	Safety		
11	Sociability	Behavior	Safety			

Given that all of the subscales measure some specific behavior within the construct of an ASD, it is not surprising that many of the subscales are highly correlated. What does appear somewhat surprising is that imitation skills seem to be strongly correlated with stereotypy. This could possibly suggest that some stereotypic behavior could be associated with seeing the behavior and then imitating it. It is also interesting to note that challenging behavior seems to be correlated with skills generally associated with higher functioning individuals such as communication complexity and sociability.

Convergent Validity

The Childhood Autism Rating Scale (CARS) (Schopler, et al., 1980) is a diagnostic scale for assessing the presence and severity of autism. The psychometric properties of the CARS are well established and a detailed discussion of the scale is provided in Chapter 2. The CARS is similar to the MN-VASS in a number of ways: both scales are intended for use in the ASD population and both employ subscales measuring specific behaviors. The CARS produces a summative score which indicates the severity of autism. While the MN-VASS is not meant as a diagnostic assessment, it is scaled so that a higher score on the subscales indicates a higher degree of the construct being measured.

A group of behavior therapists (n=22) completed a MN-VASS and a CARS in the same session. A Pearson correlation between scores was calculated yielding a correlation coefficient of .89 which indicates good convergent validity between the total scores on the two instruments.

The MN-VASS and the CARS share a number of subscales. A correlation coefficient was computed between the common subscales for the two instruments and is reported in Table 15.

Table 15

MN-VASS CARS Subscale Correlation Coefficients

Scale Number	MN-VASS Scale Name	CARS Scale Name	Correlation Coefficient
1	Toileting	--	--
2	Self Feeding	--	--
3	Dressing	--	--
4	Imitation	Imitative Behavior	.94
5	Rigidity	Adaptation to Change	.78
6	Sensory Responding	Perceptive Response	.34
7	Stereotypy	--	--
8	Manding	--	--
9	Communication Complexity	Verbal Communication	.84
10	Facial Referencing	Non Verbal Behavior	.58
11	Sociability	Relating to People	.89
12	Activity Level	Activity Level	-.20
13	Challenging Behavior	--	
14	Safety	--	

Between the CARS and the MN-VASS, four subscales are directly related: Adaptation to Change, Verbal Communication, Relating to People, and Activity Level. Three other scales are somewhat related and therefore are included in this

analysis: The MN-VASS subscale, Sensory Responding, includes statements about how the child responds to stimuli in each of the sensory systems: visual, tactile, auditory, olfactory and gustatory. The CARS has three separate scales, one for visual response, listening response and perceptive responding. A separate correlation coefficient was computed between the MN-VASs Sensory Responding Scale and the CARS, Perceptive Response, Visual responding and Auditory Responding subscales. There are seven MN-VASS scales which have no direct corollary in the CARS. These are the three self help scales (toileting, eating, dressing), stereotypy, manding, challenging behavior, and safety. The differences are to be expected because the CARS is a diagnostic instrument and the MN-VASS measures behaviors which may accompany a diagnosis of ASD.

Many of the scales reviewed in this study employed measures of validity. The STAT, CSBQ and GARS-II utilized a correlation coefficient with another instrument. Results for the MN-VASS and other scales are contained in Table 16.

Table 16

Validity Evidence of Selected Scales

Instrument	Method	Coefficient
MN-VASS	Convergent Validity with the CARS	Pearson r = .89
CARS	CARS Scores and Expert Opinion	Pearson r = .84
ADI-R	Sensitivity	96%
	Specificity	92%
ADOS	Sensitivity	86% - 100%
	Specificity	68% - 100%
STAT	Convergent with the ADOS	Kappa -
CSBQ	Concurrent validity with the ABC	Pearson r = .35 - ,71
	Convergent with the CBCL	Pearson r = .32 - .63
	Specificity	50%
ABC	Sensitivity	80%
GARS-2	Convergent Validity with the ABC	Correlations = .56 - .78
M-CHAT	Specificity	.99
	Sensitivity	.89
	Positive Predictive Power	.80

User Ratings

After completing the MN-VASS, the respondent was asked seven additional questions about the instrument. Forty nine of the 61 respondents answered these questions. The responses are summarized in Table 17.

Table 17

User Feedback on the MN-VASS

Question		Strongly Agree	Agree	Not Sure	Disagree	Strongly Disagree
1	*Does the profile output seem to match up with what you perceive to be the student's strengths and challenges?*	16%	67%	2%	4%	6%
2	*Would you find the instrument helpful in planning the child's IEP?*	6%	61%	6%	10%	2%
3	*Would you find the instrument helpful in communicating the child's strengths and needs to another teacher?*	4%	71%	8%	6%	2%
4	*Would you find the instrument helpful in communicating the child's strengths and needs to his or her parents?*	16%	67%	0%	8%	0%
5	*Would you use an instrument like this?*	10%	73%	2%	8%	0%
6	*Could you write an IEP objective in each of the areas covered by the scale?*	8%	65%	12%	10%	0%
7	*Do you feel that the scales cover all of the major areas of strengths and challenges for this population of children?*	2%	67%	12%	16%	2%

Overall, the ratings of the MN-VASS were very positive with most respondents reporting in the affirmative to the questions about the potential use and validity of the scale. However, there are some negative responses that are important to consider in future

versions of the MN-VASS. A one sample t-test was conducted on these responses to determine if the mean responses were significantly different from 3. A response of 4 or 5 indicated that the respondent agreed or strongly agreed with the statement. A score of 1 or 2 indicated that the respondent disagreed or strongly disagreed with the statement. A test value of 3 was chosen because a value less than 3 implies a negative view of the utility of the MN-VASS. Table 18 shows t-scores, significance levels (2-tailed) and effect size for each question.

Table 18

Results from Respondent Survey of the MN-VASS
n = 49

Question	*t*	*p*	*d*
1. *Does the profile output seem to match up with what you perceive to be the student's strengths and challenges?*	1.075	.288	.2
2. *Would you find the instrument helpful in planning the child's IEP?*	3.093	.003	.4
3. *Would you find the instrument helpful in communicating the child's strengths and needs to another teacher?*	5.22	.000	.8
4. *Would you find the instrument helpful in communicating the child's strengths and needs to his or her parents?*	5.387	.000	.8
5. *Would you use an instrument like this?*	4.988	.000	.7
6. *Could you write an IEP objective in each of the areas covered by the scale?*	4.965	.000	.7
7. *Do you feel that the scales cover all of the major areas of strengths and challenges for this population of children?*	4.893	.000	.7

All of the responses were significant at the .05 level, with the exception of question 1. Judged against conventional standards, effect sizes are large for questions 3,4,5,6, and 7 and small for questions 1 and 2.

Face Validity

One of the purposes of the MN-VASS was to create a visual output which would summarize a child's strengths and challenges across 14 domains of behavior. The extent to which the visual output matches up with the respondent's input is important. In the final validity study, 12 behavior therapists completed the MN-VASS and were able to view the student's profile on the computer screen when they had completed the assessment. Following a brief period (15 minutes to 1 hour), the respondents were asked to identify the output that was generated from their assessment from a field of three unidentified MN-VASS outputs. Ten of the 12 respondents correctly identified the profile of their student. The two respondents who did not correctly identify the output were asked to explain how they arrived at their decision. This feedback is documented in Chapter 5.

Differences in Responses by Group

All of the measures used to establish reliability and validity for the MN-VASS used a group of respondents who varied in the professional roles that they occupied and in their years of experience working with children on the autism spectrum. In order to determine if there were any significant differences in the patterns of responses to the MN-VASS, a one-way analysis of variance was conducted to evaluate the mean difference scores on the subscales between therapists and teachers. The dependent variables were mean scores on the subscales. The independent variable was respondent role as teacher or therapist. Three comparisons were significant: Subscale 3, Dressing ($F(1,60) = 13.858$. $p = .00$); Subscale 4, Imitation ($F(1,60) = 5.447$. $p = .02$); and Subscale 13, Challenging Behavior (.$F(1,60) = 22.222$,. $p = .00$).

For Subscale 3, Dressing Skills, and Subscale 4, Imitation Skills, therapists rated this need significantly higher than teachers. This could be due to the fact that behavior therapy programs differ in the emphasis that both of these skills might be given. Behavior therapy programs often encompass a high degree of functional instruction in dressing whereas teachers are generally not responsible for teaching dressing skills. Additionally, these skills, when taught in public schools, are often in the domain of the occupational therapist.

For Subscale 13, Challenging Behavior, teachers rated this trait as significantly higher than behavior therapists. This could be due to the fact that the ratio of behavior therapists to children is 1:1 with challenging behavior again a primary emphasis of the programming.

In addition to the role that the respondents occupied (teacher or behavior therapist), another source of variation in the respondents was the number of years of experience working with children on the autism spectrum. To investigate if the number of years of experience had an effect on respondent rating, a one way analysis of variance was conducted. Each of the 62 respondents was categorized into one of five groups by years of experience. Group 1 consisted of individuals with less than one year of experience ($n = 7$); group 2 consisted of individuals with one to three years of experience ($n = 23$), group 3 consisted of individuals with four to five years of experience ($n = 18$), group 4 consisted on individuals with five to ten years of experience ($n = 11$), and group 5 consisted of individuals with over 11 years of experience ($n = 3$). Using years of experience as the independent variable and subscale mean scores as the dependent variable, no significant differences were found between these groups. This suggests that

the perception of the need for instruction in a particular area covered by the subscales is independent of how many years of experience a teacher or therapist has with children in this population.

CHAPTER 5

Discussion

The purpose of this study was to develop a reliable and valid instrument which could help individuals who work with children with autism summarize the unique blend of strengths and challenges of each child. The data generally support the reliability and validity of the MN-VASS. The process for developing the MN-VASS and for establishing its reliability and validity was similar in many respects to many autism assessments that are currently in use.

The psychometric properties of a test are relative to other tests which are used for the same or similar purposes. For example, a math test which has a reliability of .80 would be considered good if all of the other math tests had similar reliability coefficients. However, if there were math tests which had much higher coefficients, then a coefficient of .80 would be considered less that acceptable. In contrast, a reliability coefficient of .65 would be considered by convention to be unacceptable, but if there were no other assessment for that construct or other assessments had poorer reliability, then the .65 coefficient might be viewed as the best available. Because of the relative nature of reliability and validity, the results for the MN-VASS are discussed in terms of the 12 assessments reviewed in Chapter 2. All of the results for the MN-VASS, while promising, should be interpreted with caution due to the nature and size of the sample.

Item Analysis

Item analysis was generally not reported in the literature. This is likely due to the restrictive space allowance for publication and the role that item analysis plays early in the development of a scale. Item analysis is generally conducted during item try-outs or

pilot tests so that items can be deleted or modified before a larger investigation of the scale is undertaken.

Corrected item total correlations were computed for the MN-VASS. These item statistics reflect a correlation between each item on the scale and the total scale score, excluding the item of interest. Corrected item total correlations can help a test developer delete low contributing items and retain those items which most contribute to a scale's reliability. Using corrected item total correlations is an iterative process when one begins to delete items from the scale. Deleting items changes the total score on the scale and thus directly influences all of the item total correlations. Items which are eliminated for underperformance can be reintroduced when other items are deleted with the result that a previously under-performing item can produce an adequate coefficient.

Item total correlations should serve as a guide, rather than a rule. Conceptual considerations outweigh the statistic (Green & Salkind, 2008). For example, Scale 12, Activity Level, contained the greatest number of under-performing items when item total correlations were computed for the whole scale. Three of the five items on the subscale functioned below .30 and two of them even had negative correlations. The two remaining items had coefficients of .36, which is not particularly strong. However, conceptually, Activity Level is an important characteristic to understand when planning the instructional program for a child with an ASD. Co-morbid Attention Deficit Hyperactive Disorder is common in this population (Angelica, Edelson, Asherson & Saudino, 2009; Leyfer, et al., 2006) and should be considered when summarizing a child's strengths and challenges. In addition to the conceptual basis for retaining the scale, which is the most important; when the subscales were analyzed separately, item

total correlations ranged between .43 to .79, with all of the items performing over the conventionally accepted minimum of .30.

Measures of Reliability

Like most of the instruments reviewed in Chapter 2, the MN-VASS employed a measure of internal consistency for both the overall assessment and each of the 14 subscales. The overall assessment was reduced from 83 potential questions to 35 for purposes of calculating internal consistency. The reasoning for this reduction in items is addressed in Chapter 4. Results indicate that overall the MN-VASS has good internal consistency (alpha = .89). An Alpha coefficient was also calculated for each of the subscales, and ranged from a low of .68 for the Sensory Responding Subscale and a high of .96 for the Eating Skills subscale. In general, a coefficient of .80 is considered acceptable (Carmines & Zeller, 1979). Seven of the fourteen subscales have alpha coefficients above .80. Two factors should be taken into account with the interpretation of the coefficients for the subscales. First, none of the subscales should be administered without the entire scale and second, the subscales are very short, ranging from four to seven items. Because Coefficient Alpha is related to the length of the test, these shorter subscales should not be summarily dismissed as unreliable because their coefficients are below .80.

When compared to the other Alpha Coefficients reported, the MN-VASS is well within the range of the other assessments. The CARS and the M-CHAT reported alpha coefficients for the overall scales. These were .94 and .83 respectively. The MN-VASS alpha of .89 falls between the two. It is also very common that the coefficients are reported in ranges. The lowest coefficient reported was the bottom range score of the

ADOS Restricted and Repetitive Behavior Subscale (.51). The lowest coefficient of the MN-VASS subscales was .77 for the Communication Subscales.

The MN-VASS is unique among the 12 reviewed assessments in its use of a Split half reliability coefficient. This is probably due to the fact that Coefficient Alpha was created to compensate for the spurious results that can occur using Split Half procedures (Thorndike, 2005). Split Half coefficients can vary widely depending on how one splits the test. In the conventional application, every other test item is separated into two forms of the test. However, this split would not be appropriate for the MN-VASS because of the developmental arrangement of the subscales and the imbalance in the number of questions that apply to each scale. To compensate for this problem, The MN-VASS was divided by placing entire subscales between the two halves. Scales 1 through 7 comprised the first half and Scales 8 through 14 comprised the second half. A split half coefficient expressed as a Spearman-Brown corrected correlation coefficient was computed for the MN-VASS *(r* = .90) The split half measure was used to gather as much evidence as possible to support the reliability of the scale, especially given the limitations from the small number of individuals who participated in the test-retest.

In addition to an overall split half procedure, each subscale was also analyzed using this method. The coefficients for the subscales ranged from a low of .61 for Sensory Responding and a high of .93 for stereotypy.

Test-Retest Reliability

A test-retest reliability coefficient was also employed with the MN-VASS. Using a subsample of 22 therapists, test retest was calculated using a Pearson correlation coefficient and is reported as .90. The MN-VASS test retest period was 4 months. Test-

retest reliability is subject to a number of threats to its integrity. Among these is that there is actual change in the subjects between testing at time one and time two. This does not seem to be the case with the MN-VASS. While one might despair that a high correlation would indicate that there was no real change in the subject, an equally viable interpretation could be that all of the subjects have made progress at about the same rate and their rank order remains the same. Another issue with test retest reliability is the influence of memory between the test and the retest. The extensive time period for the MN-VASS (four months) probably precludes this threat.

Test-retest reliability was also computed for each of the MN-VASS subscales. When broken out by subscale, a wide range of coefficients was evident. Subscale 10, Facial Referencing, had a coefficient of .27 while challenging behavior had a coefficient of .97. To the extent that test-retest coefficients can be a measure of stability, this could suggest that challenging behavior could be a more stable trait than that of facial referencing.

Test-retest reliability was reported for five instruments reviewed in Chapter 2. Coefficients were reported as Interclass correlation coefficients, Pearson's r, and one instrument (the STAT) used a kappa statistic. Ranges for the coefficients were from a lower bound ICC of .59 reported for the ADOS and a high of .90 reported for the MN-VASS and the STAT. This places the MN-VASS at the top of the other scales in regard to test retest reliability. However, as with all of the statistics generated for the MN-VASS, the sample size is small. In addition, the STAT's somewhat unorthodox use of the kappa statistic makes it difficult to interpret the result in the context of the other scales.

Inter Rater Agreement

Inter-rater reliability was reported for the MN-VASS as percent of exact agreement for each subscale. Eleven children were rated by a separate pair of raters. The scores on the MN-VASS items were grouped into two categories. Scores of one or two indicated that the characteristic did not represent an instructional need for that student, these scores between two raters were counted as an agreement. The same scoring was applied to students who received a rating of four or five on an item. Percent exact agreement was reported as both a range of agreements as well as an overall mean of all pairs of raters. At least one rater pair for each subscale recorded a perfect agreement. However, for many scales there was a wide range of agreement. For example, the subscales Challenging Behavior and Stereotypy, there were a pair of raters who had perfect disagreement! While one might hypothesize that this variability could be due to the instructional control of the rater with that child (i.e., one rater has excellent behavioral control over the student and one does not), there could also be a fundamental problem with how the behavior is operationalized. To compensate for the variability in how each rater interprets the behavior described in the question stem, each question on the MN-VASS has a help screen which contains a description and examples of how one might rate a particular behavior. It is impossible, however, at this point to ascertain whether any of the respondents used the help text. In any case, further investigation of the inter-rater agreement with the MN-VASS is warranted.

Of the 12 assessments reviewed in Chapter 2, inter-rater agreement was reported for nine of the studies. Methods for computing inter-rater reliability varied between kappa, weighted kappa, intra class correlation coefficients and percent of exact

agreement. Some scales employed more than one. These cases were the ADOS and the ADI-R which offered the percent exact agreement to perhaps compensate for their kappa coefficient, which only suggested moderate agreement.

The surprising negative kappa statistic arose from disagreement between raters in regard to the socialization items on the AOSI. This scale is used for infants and the statistic illustrates the need for further operationalizing the definition of social responding in children under 6 months of age. Surprisingly, the lower bound of the range provided for kappa for the ADOS (.46 – 1.0) and the ADI-R (.52 - .59) fell only in the moderate range. However, an additional measure (Percent Exact Agreement) was also provided. These percentages ranged between 81% – 100% for the ADOS and 87% for the ADI-R.

The CARS was the only scale reporting a percent exact agreement for the overall scale. This was reported as a mean of item-specific agreements. The overall agreement was 71% with item agreement ranging from a low of .55 to a high of .93. The MN-VASS range of .60 to .87 is somewhat comparable to the range of the CARS, and the average agreement of .75 is higher. Again, the small sample size for the MN-VASS must be considered in comparing the MN-VASS to other scales.

Validity

Validity is determined, not by a score, but by the interpretation of that score in a meaningful context (Thorndike, 2005). The Standards (AERA, 1999) identify validity as "the most fundamental consideration in developing and evaluating tests" (p. 9). The validity of autism assessments depends a great deal on the autism literature. All of the assessments relied on the literature to formulate the basis of the assessments. Most of the assessments considered the DSM/ICD formulations of the disorder which are composed

of a triad of deficits in communication, social reciprocity, and restricted and repetitive areas of interest. Like the other assessments, the MNVASS is constructed from the literature surrounding the presentation of symptoms of individuals with autism, which focuses on, but is not limited to the autism triad presented in the DSM-IV-TR (APA 2000).

Convergent validity was used to develop validity evidence for six of the scales. In some cases, such as the GARS-2 and the STAT, convergent validity was established through comparing scores on other assessments. The other instruments compared the results of the assessments with expert clinician opinions. Convergent validity was assessed on the MN-VASS using the Childhood Autism Rating Scale CARS. The MN-VASS and the CARS show very good convergence, $r = .89$. The CARS is scaled so that a higher score indicates greater severity of an ASD. While the MN-VASS is not a diagnostic instrument, and total scores are not necessarily interpretable, the MN-VASS does produce scores which are keyed so that higher scores indicate more of the trait. This is similar to how the CARS is scored. The correlation suggests that the constructs of autism are measured similarly in the two instruments and provides evidence that suggests that higher scores on the MN-VASS overall indicate more of the presence of autism than lower scores, much like the CARS.

In addition to an overall correlation between the two scales, subscales were also compared. A correlation coefficient was calculated for each of the subscales of the CARS and the MN-VASS that measure somewhat of the same construct. The correlations ranged from a low of -.20 between the Activity Level subscales on the CARS and the MN-VASS and a high of .94 between the Imitative Behavior Subscale on the CARS and

the Imitation Subscale of the MN-VASS. Perceptive Response on the CARS was negligibly correlated to the Sensory Responding Subscale of the MN-VASS (.34). The remaining four subscales that were paired had correlations from .78 (CARS – Adaptations to Change, MN-VASS Rigidity) to .89 (CARS – Relating to People, MN-VASS – Sociability). Seven of the subscales of the CARS were not comparable to the MN-VASS. The three adaptive behavior subscales do not have a CARS corollary, as well as Stereotypy, Manding, Challenging Behavior, and Safety. While the CARS and the MN-VASS measure some similar underlying constructs, their purposes as well as the scope of the measurement is different. The negative correlation between the only two scales with exactly the same name, "Activity Level" illustrates how even a similar concept can be interpreted and measured differently.

Convergent validity was reported for three of the scales reviewed in this study. The STAT used the ADOS as a comparison, and the CSBQ and the GARS-II both used convergence with the ABC. The CSBQ also used a correlation between its subscales and subscales of the Childhood Behavior Checklist (CBCL) (Achenbach, 1991). The correlations for the CSBQ ranged from .35 to .71 with the ABC and from .32 to .63 with the CBCL. It should be noted that correlations below .30 were not reported. The GARS-II reports correlations of between .56 and .78 with the ABC. The MN-VASS has a wider range of convergence on both the bottom and the top of the range of all of the instruments; however, in the case of the CSBQ, the lowest correlations were not reported.

User Ratings and Face Validity

The MN-VASS employed two additional techniques for establishing validity which are somewhat unique from the other assessments. MN-VASS produces a one page

visual summary of the child's strengths and challenges across 14 domains of behavior. Twelve therapists who completed the instrument were asked to identify the results of the assessment from a field of three profiles. One of the profiles was the one produced by that therapist's responses to the MN-VASS; the other two were profiles of other children in the study. Ten of the 12 therapists were able to correctly select the output that matched their student. This finding suggests that MN-VASS visual output is somewhat descriptive of and sensitive to the unique presentation of symptoms of each child.

Therapists who did not correctly identify their student were asked to explain their decision making process. This explanation yielded some interesting points which should be taken into consideration when refining the scale. First, the sociability subscale is intended to measure the level of social responsiveness that a child displays. The questions were written to elicit the degree to which the child enjoys or at least is tolerant of others. Questions range from describing very gregarious behaviors to those which are more aloof. The therapist in question had interpreted the sociability output, as "social skills" and identified social skills as a great need for this particular child. The MN-VASS does not include a social skills subscale for two reasons. First, even children with the mildest forms of PDD-NOS display a need for social skill instruction. The inclusion of a social skills subscale was hypothesized to only produce an area of need, therefore not truly contributing to any meaningful differentiation among children. The second reason was that social skills are so exceedingly complex that the author felt she could not construct a meaningful set of questions which would elicit distinguishable levels of social skills.

Another point brought up by a therapist who did not correctly identify her student was in the area of Subscale 13, Challenging Behavior. The subject of her assessment engaged in low frequency, but high intensity behaviors. The MN-VASS does not measure this profile of challenging behavior well, with most of the questions constructed to measure day-to-day responses to demands and requests, rather than more infrequent and dramatic displays of behavior. A possible correction would be a statement about what the subscale measures or an opportunity for the user to register the occurrence of low frequency, but high intensity behaviors.

The last measure of validity was in the form of a survey at the end of the MN-VASS. After the respondent had completed the subscales, but before the output was produced, respondents were surveyed to find out how they felt about various aspects of the scale. Of the 63 unique respondents, 49 completed this portion of the scale.

Overall, the ratings of the MN-VASS were very positive with most respondents reporting in the affirmative to the questions about the potential use and validity of the scale. However, there are some negative responses which are important to consider in future versions of the MN-VASS.

Many of the respondents felt that the profile produced by the program did indeed accurately reflect the strengths and challenges of the children that they were assessing. However, 10% of the respondents did not feel that the output represented their child, and an additional 2% were not sure. Part of the problem with this item was that it required the respondent to complete all of the questions about the scale, and then click on the next button to produce the output. After viewing the output, the respondent was then required to navigate back to the last subscale in order to complete the question. The other choice

would have been to direct the user to another web site such as Survey Monkey to complete their evaluations of the scale. Ideally, the output would have been produced and then the questions posed; however, this was not part of the original design of the program, and therefore, would have required more resources than were available to rewrite the program. Given the risk of losing all of this feedback data by redirecting respondents to another website, the decision was made to ask the respondents to circle back to the questions rather than completely redirect them.

Construct Representation

How we choose to measure an ASD will directly impact the information that we gain from any assessment like the MN-VASS. The traits that we select to include on measurement scales often represent an underlying theoretical notion about the disorders. An example of this discussed earlier is Constantino's Social Responsiveness Scale (2000), which contends that autism is a unidimensional disorder of social responsiveness. Purposes of the assessment also drive what we choose to include in the assessment. Diagnostic assessments will select items which discriminate best between neurotypical children and children on the autism spectrum. Most of the assessments reviewed in Chapter 2 have some questions relating to communication and social reciprocity, and many also include questions which address unusual sensory responding. Beyond these common questions, each assessment is somewhat unique in the selection of behaviors to evaluate. The MN-VASS is specifically designed to elicit information which will contribute to the formation of an instructional program. The inclusion of questions pertaining to adaptive skills is derived from the inclusion of ASDs in the Pervasive Developmental Disorder framework, which implies that the disability affects all areas of

development. The other scales were constructed specifically to focus on the types of behaviors and skills which can be taught in a classroom or therapeutic environment.

Samples

The samples that were used to establish the psychometric properties varied from an n of 26 infants for the 18 month old portion of the AOSI to 912 reported for the Q-CHAT. Larger samples generally included a number of children who were not affected by autism. Some studies included children who were only affected by the disorder, such as the AOSI which was only administered to infants in multiplex families of autism. Other studies such as the Q-Chat included vast numbers of children who were screened during routine physical examinations. The MN-VASS samples included only children on the autism spectrum. There were 61 unique responders to the MN-VASS, most of whom were full time behavior therapists. School teachers comprised only 23% of the respondents.

The MN-VASS sample sizes are not unreasonable for a small scale test of psychometric properties. The smaller subsamples of individuals who responded to the user survey, who participated in test, retest, convergent validity or face validity test are small, but not out of the realm of what other scales have used for the same purposes.

Sources of Data

The MN-VASS would fall more directly into the parent/caregiver completed category of assessment. All of the MN-VASS data collected for this study were from behavior therapists or teachers who were all familiar with children with ASD in general and with their student specifically. The assessment relies on report; there is no direct observation of behavior required to complete the assessment.

Scaling

All of the scales reviewed for this study used a form of Likert Scaling. Scaling for the MN-VASS was rather typical using a 4 point Likert Scale with the anchors, Strongly Agree, Agree, Disagree and Strongly Disagree. There was a neutral point designated as Not Observed. One of the primary highlights of the MN-VASS is its visual output. The output consists of 14 ellipses, one for each subscale. If the ellipses are small, the profile would indicate a small need or a strength for that child. If the ellipses are big, that is meant to suggest a bigger need. The output is generated in two conditions. In the first condition, the scale is a developmentally arranged scale with a header question. If the respondent replies "Strongly Agree" to the header question, the MN-VASS software is programmed to assign the smallest ellipsis to the subscale output and automatically skip the respondent to the first question on the next scale. In cases where the respondent does not select "Strongly Agree" to the header question, the MN-VASS software calculates a mean response. In order to ensure that the ellipses visually discriminated strengths from challenges, the program use the scores 5, 4, 2, and 1. The absence of a 3 prevents the program from producing mid-range ellipses, which would not be informative to the end user, i.e., all of the ellipses are the same size because of the tendency for the mean to center around 3.

Likert Scaling can pose problems with how respondents interpret scale anchors. For example, for Item 73, "This child accepts "no" for an answer," how does one decide between "Strongly Agree" and "Agree"? The MN-VASS has on-line contextual help screens for each question. The help text explains to the respondent examples of responses and behaviors that would constitute a "Strongly Agree" and differentiate it

from an "Agree" response. With the current configuration, there is no way to assess how many respondents used the help text. An interesting future study would be to measure the difference in reliability between a group of respondents who read the help texts and a group that did not.

All of the scales reviewed in Chapter 2 employed classical test theory methodologies for determining reliability. Most of the scales employed a method of adding up total scores and determining scale cutoffs. (The ABC applied weights to different score items.) While two scales (ADOS and BPASS) used ROC estimates to determine scale cut offs, virtually no IRT methods were used in scaling, scoring, or development of any of the measures. This is probably due to insufficient sample sizes to generate acceptable item parameters.

Relationship to Instruction

Of the instruments reviewed here, only two, the GARS-2 (Gilliam, 2006) and the SRS (Constantino, 2000) have instructional material provided to remediate identified deficits. For the most part, while some of the instruments can provide data to be used in instructional planning, that is not their primary purpose. In terms of Fillipek et al.'s (1999) observation that assessment should in some way have a bearing on instruction, it would seem that as of yet, that connection has not been adequately addressed.

Limitations of the Instrument

The data suggest that MN-VASS output correctly matched the respondent's impression of the subjects strengths and challenges across the 14 domains of behavior measured by the MN-VASS. However, even though the relative sizes of the output

correctly matched a child's perceived strengths or challenges, the output does not recommend a prioritizing of skills among the 14 behavioral domains.

Often before one undertakes a course of instruction for a child on the spectrum who is exhibiting challenging behavior, the behavior must be brought under control to the point where the child will cooperate with his therapists of teachers (Horner, et al., 2002). Closely related to behavior is the issue of functional communication and language. Often focusing on teaching the child appropriate ways to communicate his or her needs will mitigate challenging behavior. This type of language instruction, therefore, could possibly warrant priority over some of the other behaviors measured by the scale, such as self help skills or sociability.

There are both benefits and drawbacks to the electronic format of the MN-VASS. Among the benefits are ease of access, speed of completion, and data collection. The MN-VASS is web-hosted at www.visualautismsurvey.net. This domain is annually renewable. The MN-VASS software has generally operated without glitches, although the system did crash intermittently when adding a new student. Updates to the software in December of 2009 appear to have eliminated this glitch. Users are able to use the instrument from any computer with internet access and multiple users may log on simultaneously. The time to complete the survey is also recorded in the database. Completion times ranged from 3 minutes to15 minutes, with the overall average of 7 minutes 22 seconds. The speed with which the scale can be completed is partially attributable to the skip patterns that are programmed into the scale. While the MN-VASS is comprised of 83 questions, a user could conceivable complete the MN-VASS by answering only 35. This would occur if the respondent replied "Strongly Agree" to each

header question on the nine developmental subscales. The MN-VASS site collects and maintains a great deal of useful data. Data reports from the MN-VASS include not only user responses, but also the date and time when users accessed the scale and an e-mail address. Thus, users of the survey could be tracked and contacted if necessary. Data collected on the website is transferrable to a comma delimited format which can be pasted directly into an Excel spreadsheet, notepad, or SPSS, eliminating the need to enter data by hand. It should be noted that data extracted from the server still require extensive manipulation in preparation for analysis.

Among the disadvantages of using a computer-based platform are user suspicion of divulging private data on the internet, loss of control over incomplete surveys, limited ability to make major changes to the scale, and amount of time required to respond to end user access problems.

It is certainly prudent of teachers not to risk divulging private student data over the internet. While the data is very secure (being subject to three different levels of password protected security) and there is very little data collected that could possible identify students, there is still the risk that any internet user takes when entering data into a web based instrument, that is not knowing the constraints of the security and how the data could be used.

In addition to the barrier of user trust in regard to soliciting use of the scale, the computer based platform allowed users in the study to quit the program before completing the scale. Also limiting in this study was the availability of programming resources to modify the instrument after the programming was complete. The program is capable of allowing the system administrator to add, delete, or modify new scales and add

delete, or modify questions. The program also allows the administrator to modify help texts, designate a question as being scored negatively or positively, and assign a question as a header question and invoke the programmed skip patterns. What the program did not allow for was modification to the Likert Scale labels. While the scale was designed to collect Likert Scale data, the addition of questions at the end of the survey were limited to questions which could reasonably be answered through Likert Scale responses. Scales which require computer programming also require solid conceptualization at the beginning of the programming task, which might precede some unforeseen complication or functionality which would enhance the scales usefulness. An example of the latter is the limitation of the software to allow end users to print out a copy of their responses to each of the questions in the survey. This ability would enhance the abilities of teachers and therapists to review their responses to each specific question in the assessment. This could be very helpful when using the MN-VASS to compare observations and perceptions between parents and teachers by allowing the parties to compare the specific questions that could potentially yield disagreement.

Further, the use of a web site that allows each user to establish and protect data in his or her personal accounts requires that the user remember his user name and password. A number of individuals in the test-retest study could not remember their user name or password. These data are retrievable, but requires some time to access and an e-mail exchange with the site administrator.

While new and better assessments are available to professionals, it is difficult to recruit individuals who work with children with autism to take the time to complete the scales. Some of the reluctance on the part of the teachers has to do with the requirement

of providing some identifying information about students into a computer database. While there is not enough data to identify a student out of the context of the survey, one cannot help but attribute the reluctance of teachers to guard their student's privacy.

Limitations of the Study

While many of the psychometric properties of the MN-VASS appear to fall into an acceptable range compared with other instruments used for similar purposes, a number of cautions must be used in making any determination of the reliability and validity of the MN-VASS.

First, the size of the group on whom the study is based in small in comparison with most published assessments. Some scales such as the CHAT were tested with over 1,000 respondents, and most of the scales had populations of at least 300 upon which to conduct their data analysis.

The majority of respondents to the MN-VASS were behavior therapists working with pre-school aged children with autism at a private provider of therapy services. This poses a number of limitations on the results of the study. While most of the behavior therapists held at least bachelor's degrees, the therapists were not certified as special education teachers. There were not enough responses from certified special educators to perform any between groups analysis, therefore, conclusions about the scale's reliability and validity should be limited to behavior therapists, and not assumed to be applicable to certified special educators.

Most of the children receiving behavior therapy services were from homes where private insurance was available to pay for the services. This suggests a socio-economic status that precludes children in poverty. Thus the sample of students on whom the

results are based could be limited in their economic diversity, and results of the study might not be applicable to a group with wider variability in their socioeconomic status. Additionally, the sample of students was largely drawn from suburban neighborhoods, thus further limiting how robust the reliability might be if the sample included more children from urban and rural settings.

In terms of validity, the purpose of the MN-VASS is primarily to drive instruction. This use of the instrument was not assessed in this study. Although, many respondents replied as to their opinions as to whether the MN-VASS would be useful, there is no way to judge from this study whether the results of the assessment had any impact on instruction of any of the students.

Suggestions for Further Research & Development

The MN-VASS shows some promise as a tool which summarizes the strengths and challenges of individual children on the spectrum. However, there remains a number of further studies which would provide more evidence of the scale's reliability and validity. First, the scale should be tested in a school setting. This would provide much needed evidence of how well the scale captures the instructional needs of students in a school setting as opposed to a therapeutic one.

Students from more diverse backgrounds should be included in the sample. Currently the group from whom most of the MN-VASS data was collected is largely white, suburban, middle class children. The MN-VASS needs to be tested on a group of students with characteristics more in keeping with the current diverse populations that attend public schools.

Another area of study which should be undertaken is whether the MN-VASS has covered all of the important areas where a child might receive instruction. The current configuration of the software and the data gathered for the study assumes that most of the important areas of instruction are included in the MN-VASS, but there could be areas of instructional need that are not currently addressed.

Another question to be answered is whether the results of the assessment have any influence on either the planning or execution of classroom instruction. For example, do students' IEPs reflect the same emphasis on skills that the MN-VASS would suggest were important. Also, do classroom activities (which may or may not have a direct relationship with IEP goals) show evidence of effort in skill areas commensurate to the areas of need identified by the MN-VASS.

In addition to these questions, a number of improvements to the instrument could be undertaken. Currently, the MN-VASS provides a summary of strengths and challenges, but does not offer any educational plan for addressing the challenges. A feasible addendum to the current software would be to generate some appropriate goals and objectives based on a student's profile.

Another are for expanded development would be with generating different scales based on learner characteristics. For example, if a student was at a pre-symbolic level of communication, the scale could be adjusted to provide different levels of questions.

Summary

There are many, many scales that are being developed for children with autism. Most of the recent emphasis is on scales that can detect autism spectrum disorders at very early ages. The ADOS is the gold standard of diagnostic instruments currently in the

field. However, the ADOS requires extensive training to administer, direct interaction with a child, and takes a great deal of time. This suggests that even in the realm of diagnostic instruments, there is still the need for a reliable instrument which would require fewer resources to administer. In terms of scales which are specifically meant to guide instruction, the GARS-II is the only assessment which is accompanied by materials with which to develop an instructional program. All assessments are not intended for the same uses. Considering the paucity of assessments that are specifically designed to drive instruction, the MN-VASS could potentially fill a need and not duplicate the numerous, current assessments for children with autism.

The purpose of the MN-VASS is to help educators plan an individualized program of instruction for children with autism. The promises of early intervention may be realized through effective programming. That programming should not be based on a broad and general view of the disorder, but rather on the strengths and challenges of each individual child.

References

Allen, R.A., Robins, D., & Decker, S. (2008). Autism spectrum disorders: neurobiology and current assessment practices. *Psychology in the Schools. 45(10),* 905-917.

Allison, C.., Baron-Cohen, S., Wheelwright, S., Charman, T., Richler, J., Pasco, G & Brayne, C. (2008). The Q-Chat (Quantitative Checklist for Autism in Toddlers): a normally distributed quantitative measure of autistic traits at 18-24 months of age: preliminary report. *Journal of Autism and Developmental Disabilities, 38,*1414-1425.

Altman, D.G. & Bland, M. (1994). Diagnostic tests 2: Predictive Values. *British Medical Journal, 309,* 102.

American Educational Research Association, American Psychological Association & National Council on Measurement in Education (2004). Standards for educational and psychological testing: Washington, D.C.: American Educational Research Association.

American Psychiatric Association. (1980). Diagnostic and statistical manual of mental disorders (Vol. III). Washington, D.C.: Author.

American Psychiatric Association. (2004c). The diagnostic and statistical manual of mental disorders (Vol. IV-TR). Washington, D.C.: Author.

Baird, G., Charman, T., Baron-Cohen, S., Cox,.A., Swettenham, J., Wheelwright, S., et al. (2000). A screening instrument for autism at 18 months of age. A 6-year follow up study. *Journal of the American Academy of Child and Adolescent Psychiatry, 39,* 694-702.

Baron-Cohen, S. (2002). The extreme male brain theory of autism. *Trends in Cognitive Science,* 6(6), 248-254.

Baron-Cohen, S., Allen, J., & Gillberg, C. (1992). Can autism be detected at 18 months? The needle, the haystack, and the CHAT. *The British Journal of Psychiatry, 161,* 839-843

Baron-Cohen, et. al. (2000). Early identification by the Checklist for Autism in Toddlers (CHAT). *Journal of the Royal Society of Medicine, 93,* 521-525.

Baron-Cohen, S., Wheelwright, S., Skinner, R., Martin, J., & Chubley, E. (2001). The Autism Spectrum Quotient (AQ): evidence from Asperger Syndrome/high-functioning autism, males and females, scientists and mathematicians. *Journal of Autism and Developmental Disabilities, 31,* 5-17.

Bartak, L. & Rutter, M. (1976). Differences between mentally retarded and normally intelligent autistic children. *Journal of Autism and Childhood Schizophrenia,* 6(2), 109-128.

Bellini, S. (2004). Social skill deficits and anxiety in high-functioning adolescents with autism spectrum disorders. *Focus on Autism and Other Developmental Disorders, 19(2)* 78-86.

Bryson, S.E., Zwaigenbaum, L. McDermott, C. Rombough, V. & Brian J. (2008). The Autism Observation Scale for Infants: scale development and reliability data. *Journal of Autism and Developmental Disorders, 38,* 731-738.

Campbell, J. M. (2003). Efficacy of behavioral interventions for reducing problem behavior in persons with autism: a quantitative synthesis of single-subject research. *Research in Developmental Disabilities,* 24, 120-138.

Campbell, Jonathan. (2005). Diagnostic assessment of Asperger's disorder: A review of five third party rating scales. *Journal of Autism and Developmental Disabilities.* 35(1) 25-35.

Carmines, E.G. & Zeller, R.A. (1979). Reliability and validity. [Monograph] *Series: Quantitative Applications in the Social Sciences.* Number 07-017.

Carter, A.S., Volkmar, F.A., Sparrow, S.S., Wang, J., Lord, C., Dawson, G., et al., (1998), The vineland adaptive behavior scales: supplementary norms for individuals with autism. *Journal of Autism and Developmental Disabilities, 28*(4), 287-302.

Carvill, S. (2001). Sensory impairments, intellectual disability and psychiatry. *Journal of Intellectual Disability Research*, 45(6), 467-483.

Centers for Disease Control. (2007). Prevalence of the Autism Spectrum Disorders in Multiple Areas of the United States, Surveillance Years 2000 and 2002: Center for Disease Control.

Charak, D. & Stella, J.L. (2002). Screening and diagnostic instruments for identification of autism spectrum disorders in children, adolescents and young adults: A selective review. *Assessment for Effective Intervention; 27*(5), 5-17.

Charman, T., Baron-Cohen, S., Baird, G., Cox, A., Wheelwright, S., Swettenham, J. & Drew, A. (2001). Commentary: The Modified Checklist for Autism in Toddlers. *Journal of Autism and Developmental Disorders, 31*(2). 145-148.

Charwarska, K., Klin, A., Paul, R., & Volkmar, F. (2007). Autism spectrum disorder in the second year: stability and change in syndrome expression. *Journal of Child Psychology and Psychiatry, 48(2),* 128-138.

Clark, L.A. & Watson, D. (1995). Constructing validity: Basic issues in objective scale development. *Psychological Assessment,* 7(3), 309-319.

Chronbach, L.J., & Meehl, P.E. (1955). Construct validity in psychological test. *Psychological Bulletin*, 52, 281-302.

Constantino, J. (2004). *The Social Responsiveness Scale,* Western Psychological Services, Los Angeles, California.

Constantino, J.N. & Todd, R.D. (2000). Genetic structure of reciprocal social behavior. *American Journal of Psychiatry, 157,* 2043-2045.

Constantino, J., Davis, S.A., Todd, R.D., Schindler, M.K., Gross, M.M., Brophy, S.L., et al. (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 Disabilities,* 33(4) 427-433.

Constantino, J.N., Gruber, C.P., Davis, S., Hayes, S., Passanante, M., & Przybeck, R. (2004). The factor structure of autistic traits. *Journal of Child Psychology and Psychiatry, 45*(4),719-726.

Dalrymple, N.J. & Ruble, L.A. (1992). Toilet training and behaviors of people with autism: Parent views. *Journal of Autism and Developmental Disorders, 22*(2) 265-275).

Dawson, G., Estes, A., Munson, J., Schellenberg, G. Bernier, R. & Abbott, R. (2007). Quantitative assessment of autism symptom-related traits in probands and parents: Broader Phenotype Autism Symptoms Scale. *Journal of Autism and Developmental Disorders, 37,* 523-536.

De Bildt, A., Sytema, S., Ketelaars, C., Kraijer, D., Volkmar, F., & Minderaa, R. (2004). Measuring pervasive developmental disorders in children and adolescents with mental retardation: A comparison of two screening instruments used in a study of the total mentally retarded population from a designated area. *Journal of Autism and Developmental Disorders, 33*(6), 595-605.

D'Eugenio, D. B., et. al. (1998). The Preschool Developmental Checklist. Ann Arbor: University of Michigan Press.

DiLalla, D. & Rogers, S. (1994). Domains of the Childhood Autism Rating Scale: Relevance for diagnosis and treatment. *Journal of Autism and Developmental Disorders, 24*(2), 115-128.

DiLavore, P.C., Lord, C., & Rutter, M. (1995). The pre-linguistic autism diagnostic observation schedule. *Journal of Autism and Developmental Disorders, 25,* 355-379.

Eaves, R.C. & Williams, T.O. Jr. (2006). The reliability and construct validity of ratings for the Autism Behavior Checklist, *Psychology in the Schools, 43(2),* 129 - .

Filipek, P. et al. (1999). The screening and diagnosis of autism spectrum disorders. *Journal of Autism and Developmental Disorders, 29*(6), 439-484.

Furuno, S. (2004). The Hawaii Early Learning Profile (HELP). Palo Alto: VORT Corporation.

Gall, M.D., Gall, J.P. & Borg, W.R. (2003).Educational Research, An Introduction, Allyn and Bacon, Boston, 7th Edition.

Garfin, D.G. & McCallon, D. (1988). The validity and reliability of the Childhood Autism Rating Scale with autistic adolescents. *Journal of Autism and Developmental Disorders. 18*(3), 367-378.

Gilliam, J. (1998). Gilliam Autism Rating Scale (2 ed.). Autism: Pro-ed.

Gillott, A., Furniss, F. & Walter, A. (2001). Anxiety in high-functioning children with autism. *Autism,* 5(3), 277-286.

Goldberg, E.A., (2004). The link between gastroenterology and autism. *Gastroentereology Nursing,* 27(1), 16-19.

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

Gotham, K., Risi, S., Dawson,G., Tager-Flushberg, H., Johseph, R., Carter, A., et al. (2008). A replication of the Autism Diagnostic Observation Schedule (ADOS) Revised Algorithms. *Journal of the American Academy of Child and Adolescent Psychiatry, 47(6),* 642-651.

Gray, K.M., Tonge, B.J., Sweeney, D. J. (2008). Using the Autism Diagnostic Interview-Revised and the Autism Diagnostic Observation Schedule with young children with developmental delay: evaluating diagnostic validity. *Journal of Autism and Developmental Disorders, 38, 657-667.*

Harrison, J. & Hare, D.J. (2004). Brief Report: Assessment of sensory abnormalities in people with autistic spectrum disorders. *Journal of Autism and Developmental Disorders,* 34(6), 727-730.

Hartmann, D. P. (1977). Considerations in the choice of interobserver reliability estimates. *Journal of Applied Behavior Analysis, (10),* 103-116.

Heflin, L.J., & Simpson, R. (1998). Intervention for children and youth with autism: Prudent choices in a world of exaggerated claims and empty promise. Part I:

Intervention and treatment option review. *Focus on Autism and Developmental Disabilities,13,* 212-220.

Horner, R.H., Carr, E.G., Strain, P.S., Todd, A. W., & Reed, H.K. (2002). Problem behavior interventions for young children with autism: a research synthesis. *Journal of Autism and Developmental Disabilities* 32(5), 423-446.

Horvath, K. & Perman, J.A. (2002). Autism and gastrointestinal symptoms. *Current Gastroenterology Reports,* 4(3), 251-258.

Hurth, J., Shaw, E., Izeman, S.G., Whaley,K, & Rogers, S.J. (1999). Areas of agreement about effective practices among programs serving young children with autism spectrum disorders. *Infants and Young Children, 12(2),* 17-26.

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

Kim, J.A., Szatmari, P., Bryson, S.E., Streiner, D.L. & Wilson , F.J. (2000). The prevalence of anxiety and mood problems among children with autism and Aspergers Syndrome. *Autism,* 4(2), 117-132.

Kleinman, J.M., Robins, D.L., Ventola, P.E, Pandey, J., Boorstein, H.C.,Esser, E.L. (2008). The Modified Checklist for Autism in Toddlers: A follow-up study investigating the early detection of autism spectrum disorders. *Journal of Autism and Developmental Disorders, 38,* 827-839.

Klin, A., Saulnier, C.A., Sparrow, S., Cicchetti, D.V., Volkmar, F.R. & Lord, C. (2007). Social and communication abilities and disabilities in higher functioning individuals with autism spectrum disorders: The Vineland and the ADOS. *Journal of Autism and Developmental Disorders, 37,* 748-759.

Krug, D.A., Aric, J. R., & Almond, P.G. (1980). Behavior checklist for identifying severely handicapped individuals with high levels of autistic behavior. *Journal of Child Psychology and Psychiatry, 21,* 221-229.

Lecavalier, L., Aman, M.G., Scahill, L., McDougle, C.J., McCracken, J.T., Vitiello, B., et al. (2006). Validity of the Autism Diagnostic Interview-Revised. *American Journal on Mental Retardation, 111(3),*199-215.

Lecavalier, Luc (2005). An evaluation of the Gilliam Autism Rating Scale. *Journal of Autism and Developmental Disabilities* 35(6) 795-805.

LeCouteur, A.L., Haden, G., Hammal, D., McConachie, H. (2008). Diagnosing autism spectrum disorders in pre-school children using two standardized assessment instruments: the ADI-R and the ADOS. *Journal of Autism and Developmental Disabilities, 38,* 362-372.

Le Couteur, A., Rutter, M., Lord, C., Rios, Pl., Robertson, S., Holdgrafer, M. & McLennan, J.D. (1989). Autism Diagnostic Interview: A semi-structured interview for parents and caregivers of autistic persons. *Journal of Autism and Developmental Disorders, 19,* 363-387.

Leyfer, O.T., Folstein, S.E., Bacalman, S., Davis, N.O., Dinh, E., Morgan, J., et al. (2006). Comorbid psychiatric disorders in children with autism: Interview development and rates of disorders. *Journal of Autism and Developmental Disorders, 36,* 849-861.

Lord, C., Risi, S., Lambrecht, L., Cook, E.H., Leventhal, B.L.,DiLavore, P.C., Pickles, A., & 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., Goode, S., Heemsbergen, J., Jordan, H. Mawhood, L. & Schopler, E. (1989). Autism Diagnostic Observation Schedule: A standardized observation of communicative and social behavior. *Journal of Autism and Developmental Disorders*, 19, 185-212.

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,* 659-685.

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

Lovaas, I.O. (1987). Behavioral treatment and normal education and intellectual functioning in young children. *Journal of Consulting and Clinical Psychology, 55,* 3-9.

Love, J.R., Carr, J. & LeBlanc, L.A. (2009). Functional assessment of problem behavior in children with autism spectrum disorders: A summary of 32 outpatient cases. *Journal of Autism and Developmental Disabilities* 39, 363-372.

Luteijn, E., Luteijn, F., Jackson, S., Volkmar, F., & Minderaa, R. (2000). The Children's Social Behavior Questionnaire for milder variants of PDD problems: Evaluation of psychometric characteristics. *Journal of Autism and Developmental Disabilities, 30*(4), 317-330.

Mahoney, W.J., Szatmari, P., Maclean, J., Bryson,S.e., Bartolucci, G., Walter, S.E., et al., (1998). Reliability and accuracy of differentiating pervasive developmental

disorder subtypes. *Journal of the American Academy of Child & Adolescent Psychiatry, 37*(3), 278-285.

Matson, J. ed. (2008). *Clinical Assessment and Intervention for Autism Spectrum Disorders,* Elsevier, Inc., Oxford.

Mazefsky, C.A. & Oswald, D.P. (2006). The discriminative ability and diagnostic utility of the ADOS-G, ADI-R, and GARS for children in a clinical setting. *Autism, 10*(6), 533-549.

Mirenda-Linne, F.M. & Melin, L. (2002). A factor analytic study of the Autism Behavior Checklist. *Journal of Autism and Developmental Disabilities 32(3),* 181-.

Molloy, C.A., & Manning-Courtney, P. (2003). Prevalence of Chronic Gastrointestinal Symptoms in children with autism and autism spectrum disorders. *Autism,* 7(2), 165-171.

Montgomery, J.M., Newton, B. & Smith, C. (2006) GARS-2: Test Review: Gilliam Autism Rating Scale Second Edition. Austin TX: PRO-ED.

Mun, E.Y., & Von Eye, A. (2004) Analyzing Rater Agreement: Manifest Variable Methods. Laurence Earlbaum Associates.

Mundy, P., Sigman, M. & Kasari, C. (1990). A longitudinal study of joint attention and language development in autistic children. *Journal of Autism and Developmental Disabilities, 20,* 115-128.

Murphy, G.H., Beadle-Brown, J., Wing, L., Gould, J. Shah, A. & Holmes, N. (2005). Chronicity of challenging behavior in people with intellectual disabilities and/or autism: a total population sample. *Journal of Autism and Developmental Disorders.* 35(4), 405-418.

National Institutes of Health, National Institute of Mental Health (2007). Autism spectrum disorders pervasive developmental disorders with addendum, January 2007. Retrieved November 17, 2007, from http://www.nimh.nih.gov/health/publications/ autism/complete-publication.shtml#pub3.

Oosterling, I.J., Swinkels, S.H., van der Gaag, R.J., Visser, J.C., Dietz, C. & Buitelaar, J.K. (2009). Comparative analysis of three screening instruments for autism spectrum disorder in toddlers at high risk. *Journal of Autism and Developmental Disorders,* 39, 897-909.

Patzold, L.M., Richdale, A.L., & Tonge, B.J. (1998). An investigation into sleep characteristics of children with autism and asperger's disorder. *Journal of Pediatrics & Child Health,* 34(6), 528-533.

Perry, A., Condillac, R.A., Freeman, N.L., Dunn-Greier, J., & Belair, J. (2005). Multisite study of the Childhood Autism Rating Scale (CARS) in five clinical groups of young children. *Journal of Autism and Developmental Disorders, 35*(5), 625-634.

Pine, E., Luby, J., Abbacchi, A., & Constantino, J. N. (2006). *Quantitative assessment of autistic symptomology in preschoolers.* Autism, 10(4), 344-352.

Posserud, M-B., Lundervold, A.J. & Gillberg, C. (2009). Validation of the Autism Spectrum Screening Questionnaire in a total population sample. *Journal of Autism and Developmental Disorders,* 39:126-134.

Prevention, U. S. C. f. D. C. a. (n.d.). Prevalence of the autism spectrum disorders in multiple areas of the United States, surveillance years 2000 and 2002. Retrieved, March 2007 from, from http://www.cdc.gov/ncbddd/dd/addmprevalence.htm

Rellini, E., Tortolani, D., Trillo, S., Carbone, S., & Montecchi, F. (2004). Childhood Autism Rating Scale (CARS) and Autism Behavior Checklist (ABC) correspondence and conflicts with the DSM-IV criteria in diagnosis of autism. *Journal of Autism and Developmental Delay,* 34 (6), 703-708.

Richdale, A.L. (1999). Sleep problems in autism: prevalence cause and intervention. *Developmental Medicine & Child Neurology,* 41, 60-66.

Ring, H., Woodbury-Smith, M., Watson, P., Wheelwright, S., & Baron-Cohen, S. (2008). Clinical heterogeneity among people with high functioning autism spectrum conditions: evidence favouring a continuous severity gradient. *Behavioral and Brain Functions, 4(11),*

Robins, D. (2008). Screening for autism spectrum disorders in primary care settings. *Autism, 12(5),* 537-556.

Robins, D.L., Fein, D., Barton, M.L. & Green, J.A. (2001). The Modified Checklist for Autism in Toddlers: An initial study investigating the early detection of autism and pervasive developmental disorders. *Journal of Autism and Developmental Disorders, 31*(2), 131-144.

Rodrigue, J.R., Morgan, S.B. & Geflken, G.R. (1991). A comparative evaluation of adaptive behavior in children and adolescents with autism, down syndrome, and normal development. *Journal of Autism and Developmental Disorders. 23,* 123-140.

Ronald, A., Happe, F., Bolton, P, Butcher, L.M., Price, T.S., Wheelwright, S. et al. (2006). Genetic heterogeneity between the three components of the autism spectrum: a twin study. *Journal of the American Academy of Child and Adolescent Psychiatry*, 45(6), 691- 699

Rutter, M. & Schopler, E. (1987). Autism and pervasive developmental disorders: Concepts and diagnostic issues. *Journal of Autism and Developmental Disorders, 17(2).* 159-186.

Saemundsen, E., Magnusson, P., Smari J., & Sigurdardottir, S. (2003). Autism Diagnostic Interview-Revised and the Childhood Autism Rating Scale: convergence and discrepancy in diagnosing autism. *Journal of Autism and Developmental Delay,* 33 (3), 319-328.

Schloper, E. & Reichler, J. (1980). Toward objective classification of childhood autism: Childhood Autism Rating Scale (C.A.R.S.). *Journal of Autism and Developmental Disorders,* 10, 91-103.

Schopler, E., Reichler, R. J., & Renner, B. R. (1986). *The childhood autism rating scale (CARS): for diagnostic screening and classification of autism.* Western Psychological Services.

Schreck, K.A., Williams, K., & Smith, A. (2004). A comparison of eating behaviors between children with and without autism. *Journal of Autism and Developmental Disorders.* 34(4). 433-438.

Sevin, J.A., Matson, J.L., Coe, D.A. Fee, V.E. & Sevin, B. (1991). A comparison and evaluation of three commonly used autism scales. *Journal of Autism and Developmental Disabilities, 21(4),* 417 - 429.

Shrout, P.E. & Fleiss, J.L. (1979). Intraclass correlations: uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Siegel, B. (1998). *The World of the Autistic Child.* Oxford University Press, New York.

South, M., Williams, B.J., McMahon, W.M., Owley, T., Filipek, P.A., Shernoff, E., et al., (2002). Utility of the Gilliam Autism Rating Scale in research and clinical populations. *Journal of Autism and Developmental Disorders,32*(6), 509-619.

Sim, J., & Wright, C.C. (2005). The kappa statistic in reliability studies: use, interpretation and sample size requirements. *Physical Therapy, 85*(3), 257-268.

Singh, N.N., Lancioni, G.E., Sinton, A.S.W., Fsher, B.C., Wahler, R.G., McAleavey, K., Singh, J. & Sabaawi, M. (2006). Mindful parenting decreases aggression, noncompliance, and self-injury in children with autism. *Journal of Emotional and Behavioral Disorders,* 14(3), 169-177.

Spiker, D., Lotspeich, L.J., Dimiceli, S., Myers, R.M., & Risch, N. (2002). Behavioral phenotypic variation in autism multiplex families: evidence for a continuous severity gradient. *American Journal of Medical Genetics Part B: Neuropsychiatric Genetics. 114*(2), 129-136.

Spondheim, E. (1996). Changing criteria of autistic disorders: A comparison of the ICD-10 research criteria and DSM-IV with DSM-III, CARS, and ABC. *Journal of Autism and Developmental Disorders. 26(*5), 513-525.

Stone, W.L., Coonrod, E.E., & Ousley, O.Y. (2000). Brief report: Screening Tool for Autism in Two Year Olds (STAT): development and preliminary data. *Journal of Autism and Developmental Disorders, 30(6),* 607-612.

Stone, W.L., Coonrod, E.E., Pozdol, S.L., & Turner, L.M. (2003). The parent interview for autism-clinical version (PIA-CV): A measure of behavioral change for young children with autism. *Autism. ,* 9-13.

Stone, W.L. & Hogan, K.L. (1993). A structured interview for identifying young children with autism. *Journal of Autism and Developmental Disorders. 23*, 639-652.

Stone, W.L., Coonrod, E.E., Turner, L.M. & Pozdol, S.L. (2004). Psychometric properties of the STAT for Early Autism Screening. *Journal of Autism and Developmental Disabilities.* 34(6) 691-701.

Tachimori, H., Osada, H., & Kurita, H. (2003). Childhood autism rating scale – Tokyo version for screening pervasive developmental disorders. *Psychiatry and Clinical Neurosciences, 57*(1), 113-118.

Taira, M., Takase, M., & Sasaki, M.D., (2008). Sleep disorder in children with autism. *Psychiatry and Clinical Neurosciences, 52*(2), 182-183.

Thorndike, R.M. (2005) *Measurement and Evaluation in Psychology and Education*. Pearson Education Inc., Upper Saddle River, New Jersey.

Tomanik, S.S., Pearson, D.A., Loveland K.A., Lane, D.M., & Shaw, J.B. (2007). Improving the reliability of autism diagnoses: examining the utility of adaptive behavior. *Journal of Autism and Developmental Disability, 37,*921-928.

Ventola, P.E., Kleinman, J., Pandey, J, Barton, M., Allen, S., Green, J., Robins, D & Fein, D. (2006) Agreement among four diagnostic instruments for autism spectrum disorders in toddlers. *Journal of Autism and Developmental Disorders 36*. 839-847.

Ventola, P.E., Kleinman, J., Pandey, J, Wilson, L., Esser, E, Boorstein, H., et al. (2007). Differentiating between autism spectrum disorders and other developmental disabilities in children who failed a screening instrument for ASD. *Journal of Autism and Developmental Disorders 37*. 425-436.

Volkmar, F., Charwarska, K., & Klin, A. (2005). Autism in infancy and early childhood. *Annual Review of Psychology, 56,*315-336.

Volkmar, F.R., Cicchetti, D.V., Dykens, E., Sparrow, S., Leckman, J.F. & Cohen, D.J. (1988). An evaluation of the Autism Behavior Checklist. *Journal of Autism and Developmental Disorders.* 18(1) 81-97.

Wadden, N.P.K, Bryson, S.E., & Rodger, R.S. (1991). A closer look at the Autism Behavior Checklist: Discriminant validity and factor structure. *Journal of Autism and Developmental Disorders, 21(4)* 529- 541.

Wiggins, L.D., & Robins, D. (2008). Excluding the ADI-R behavioral domain improves diagnostic agreement in toddlers. *Journal of Autism and Developmental Disorders. 38*(5), 972-976.

Williams, G.P., Dalrymple, N. & Neal, J. (2000). Eating habits of children with autism. *Pediatric Nursing*, 26(3) 259-264.

Williams, J.G., et al. (2008). The Childhood Autism Spectrum Test (CAST): Sex differences. *Journal of Autism and Developmental Disorders,* 38:1731-1739.

William, J.G. & Brayne, C.(2006). Screening for autism spectrum disorders: what is the evidence? *Autism, 10*(1), 11-35.

Williams, J., Scott, F., Stott, C., Allison, K. C., Bolton, P. Baron-Cohen, S., & Brayne, C. (2005). The CAST (Childhood Asperger Syndrome Test (CAST): Test-retest reliability. *Autism.10,* 415-427.

Williams, J., Allison, C., Scott, F., Stott, C., Bolton, P. Baron-Cohen, S., & Brayne, C. (2006). The Childhood Asperger Syndrome Test (CAST): Test-retest reliability. Autism, 10, 415-427.

Wing, L. (1996). Autism spectrum disorders. *British Medical Journal, 312: 327-328.*

Wing, L. & Gould, J. (1997). Severe impairment of social interaction and associated abnormalities in children: Epidemiology and classification. *Journal of Autism and Developmental Disorders. 23,* 639-652.

APPENDIX A

The Questions on the Minnesota Visual Autism Symptom Scale (MN-VASS)

1. This child uses the toilet independently.
2. This child can adequately wipe him/herself after using the toilet.
3. This child is toilet trained for urination.
4. This child is toilet trained for bowel movements.
5. This child is successful on a toileting schedule.
6. This child is not potty trained.
7. This child can cut meat into bite sized pieces with a knife and fork.
8. This child can pour liquid into a cup from a larger container
9. This child can spread butter on a piece of toast.
10. This child can feed himself/herself with a fork.
11. This child can get a drink from a water fountain.
12. This child independently drinks from a cup that does not have a lid.
13. This child can get dressed independently
14. This child can tie her or his own shoes.
15. This child can manage most fasteners
16. This child can put on his/her own winter coat.
17. This child can put on and take off his/her shoes.
18. This child does not participate in his or her own dressing and requires extensive adult support.
19. This child will spontaneously imitate the actions of others.
20. This child will imitate an action performed by another child when prompted.
21. This child will imitate an adult when the adult asks the child to perform an action and models it for the child.
22. This child does not imitate.
23. This child can copy simple drawings
24. This child can imitate a sequence of two or more actions.
25. This child is flexible and responds well to changes in his or her routine or schedule.
26. This child adheres to a rigid routine.
27. This child becomes distressed if changes are made to her/his routine.
28. This child becomes distressed if unexpected people show up.
29. This child likes to wear the same clothes to school every day.
30. This child engages in nonfunctional routines.
31. This child squints or covers his/her eyes in florescent lighting…
32. This child will often cover his/her ears in the presence of a moderately loud ….
33. This child resists being hugged or patted.

34. This child smells objects, food, or people more frequently than a typical child.
35. This child is strongly influenced by the texture of his or her food.
36. This child does not engage in any noticeable stereotypic behaviors beyond what would be considered normal for his/her age.…
37. When alone, this child engages in hand flapping, tapping, rocking, finger flicking or other stereotypy
38. In the company of others this child engages in hand flapping, tapping, rocking, finger flicking or other stereotypy
39. During a demanding task or when bored this child engages in hand flapping, tapping, rocking or other stereotypy
40. This child can ask for information.
41. This child can ask another person to perform an action.
42. This child can ask for help.
43. This child can ask another person to stop doing something the child finds annoying.
44. This child has many things he/she can ask for using words.
45. This child can ask for things he/she needs by using sign or pictures. If the student uses a more sophisticated method of manding
46. The most common way this child requests is through crying
47. This child uses age-appropriate speech to communicate.
48. This child can use two word combinations to communicate.
49. This child uses one word utterances or phrases that function as one word to communicate.
50. This child can echo what you say.
51. This child can ask and answer "wh" questions.
52. This child looks at you from time to time when he/she is communicating with you.
53. This child actively avoids making eye contact.
54. This child will talk to you, but will not look at you without a prompt.
55. This child will look at your face to see your emotional state: smiling
56. This child plays imaginatively with other children.
57. This child plays with other children.
58. This child is interested in other children his own age.
59. This child seems interested in adults in the environment.
60. This child ignores other children.
61. This child will actively avoid other children.
62. This child will ignore adults in his/her environment.
63. This child is constantly on the go.
64. This child will not sit still while eating
65. This child enjoys climbing and physical play

66. This child seems like she/he never gets tired.
67. This child is sluggish and will not engage in physical activity unless prompted.
68. This child follows directions.
69. This child accepts "no" for an answer
70. This child will leave a preferred activity with an adult request without engaging in challenging behavior.
71. This child will allow other children to enter his/her space and use the same materials or sit next to them.
72. This child will go along with a group of children (such as down to the gym/motor room) without engaging in challenging behavior.
73. This child will separate from significant adults at an age appropriate level without engaging in challenging behavior.
74. This child will respond to a strong command from an adult such as "No!" "Stop!" or "Hot!" to prevent the child from engaging in dangerous behavior.
75. This child will elope from familiar places (like his or her house or school).
76. This child will bolt in dangerous places such as in malls or parking lots.
77. This child can tell a stranger from a familiar person and will respond differentially to them.
78. This child will behave with appropriate caution around animals.
79. This child will climb on furniture to a degree which is dangerous.

APPENDIX B

Sample Visual Output from a MN-VASS