AN INVESTIGATION OF THE RELATIONSHIP BETWEEN THE ITEM-CONTENT AND DIAGNOSTIC VALIDITY OF AN AUTISM SCREENING INSTRUMENT

By

JUSTIN HAMPTON

A dissertation submitted in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

WASHINGTON STATE UNIVERSITY
Department of Psychology

JULY 2016

© Copyright by JUSTIN HAMPTON, 2016 All Rights Reserved
To the Faculty of Washington State University:

The members of the faculty Committee appointed to examine the dissertation of JUSTIN HAMPTON find it satisfactory and recommend that it be accepted.

___________________________________
Paul Strand, Ph.D., Chair

___________________________________
David Marcus, Ph.D.

___________________________________
Maria Gartstein, Ph.D.
ACKNOWLEDGEMENTS

I would like to thank all the people who helped and supported me with completing this research project.

First, I would like to express my gratitude to my advisor and mentor, Dr. Paul Strand, for his guidance, patience and support both on this project and throughout my graduate training.

Second, I would like to thank Dr. David Marcus and Dr. Masha Gartstein for being on my committee and for their encouragement.

Third, I would like to thank all the members of the comprehensive autism disorder evaluation team (CADET) at the Children’s Developmental Center in Kennewick, WA, led by Dr. Scott Grewe, who helped in completing in-depth assessments that served as the method of data collection for this study.

Last, but not least, I would like to thank my wife, Tegan, and my family for their understanding, constant support and encouragement toward the successful completion of this dissertation.
AN INVESTIGATION OF THE RELATIONSHIP BETWEEN THE ITEM-CONTENT AND DIAGNOSTIC VALIDITY OF AN AUTISM SCREENING INSTRUMENT

Abstract

by Justin Hampton, Ph.D.
Washington State University
July 2016

Chair: Paul Strand

The present study investigated whether the diagnostic validity of a poorly-performing instrument (the Gilliam Autism Rating Scale – Second Edition, or the GARS-2) designed to screen for autism spectrum disorders could be improved through item content manipulation (via altering the weighting of items and including new items). In previous studies, the GARS-2 has been shown to have poor sensitivity and specificity. Item content manipulations were theory-based, and involved efforts to align item content with existing diagnostic criteria and with the item content of frequently utilized criterion instruments [e.g., the Autism Diagnostic Observation Schedule (ADOS)]. Consistent with expectations, results revealed that by reweighting items to mirror the ADOS and diagnostic criteria distributions, and by including items that assess executive functioning deficits, the sensitivity and specificity could be significantly improved. These results have implications for (a) improving the validity of a currently available instrument, and (b) developing instruments with high diagnostic validity in the future. These findings are timely, given that the recent advent of the DSM-5 (and the new autism diagnostic criteria it contains) necessitates the development of new screening instruments or the modification of existing ones.
TABLE OF CONTENTS

ACKNOWLEDGEMENTS .......................................................................................... iii

ABSTRACT ............................................................................................................ iv-vii

LIST OF TABLES ................................................................................................... viii

LIST OF FIGURES ............................................................................................... ix

CHAPTER

1. INTRODUCTION ............................................................................................... 1
   Importance of Assessing Children For Autism ..................................................... 1
   Screening Instrument Characteristics ............................................................... 2
   Age Considerations ........................................................................................... 3
   Setting Considerations ....................................................................................... 3
   Breadth of Application ..................................................................................... 4

2. LITERATURE REVIEW ....................................................................................... 5
   Currently Available Screening Instruments ....................................................... 5
   Social Responsiveness Scale ............................................................................ 5
   Child Behavior Checklist/1.5-5 ......................................................................... 6
   Withdrawn ......................................................................................................... 7
   Pervasive Developmental Problems ................................................................. 7
   Childhood Autism Rating Scale ....................................................................... 7
   Social Communication Questionnaire ............................................................. 8
   Gilliam Autism Rating Scale .......................................................................... 9
   Diagnostic Validity ........................................................................................... 10
Sensitivity and Specificity……………………………………………………..10
Summary Statistics……………………………………………………..10
Positive and Negative Predictive Value. ........................................11
Receiver Operating Characteristic (ROC) .................................12
Diagnostic Validity of Currently Available Screening Instruments..............13
Social Responsiveness Scale..........................................................13
Child Behavior Checklist/1.5-5.......................................................14
Childhood Autism Rating Scale....................................................16
Social Communication Questionnaire...........................................17
Gilliam Autism Rating Scale........................................................19
Summary Statistics of Currently Available Instruments.............................20
Item Content Differences and Validity............................................21
Item Distribution of Currently Available Instruments..............................23
Executive Functioning in ASD Screening........................................26
Purpose of Current Research.........................................................27
Hypotheses..................................................................................28
3. METHODOLOGY........................................................................30
Participants..................................................................................30
Measures.....................................................................................31
Autism Diagnostic Observation Schedule – Generic...............................31
Gilliam Autism Rating Scale – Second Edition.................................32
Behavior Rating Inventory of Executive Functioning – Preschool Version....32
Procedure....................................................................................33
4. RESULTS

Participant Characteristics

Diagnostic Validity of the Baseline and Theoretical Models

Sensitivity and Specificity Results

Bivariate Results

Univariate Results

ROC Results

5. DISCUSSION

ROC Interpretation

Limitations

Recommendations and Conclusions

REFERENCES

APPENDIX
LIST OF TABLES

Table 1. SRS, CBCL, CARS, SCQ, & GARS Diagnostic Validity........................................56
Table 2. Average Sensitivity And Specificity Scores For Each Screening Instrument........58
Table 3. A Breakdown Of Items On Screening Instruments According To DSM-IV-TR Diagnostic Domains..................................................................................................................59
Table 4. Participant Characteristics.........................................................................................60
Table 5. A Breakdown Of Item Weights Per Domain For Models...........................................61
Table 6. Descriptive Statistics Of The Baseline And Theoretical Models..............................62
Table 7. Diagnostic Validity Of Models....................................................................................63
Table 8. McNemar Chi-Square Test Between Sensitivity Rates Of The Baseline And Theoretical Models For Which The Specificity Was Set To .70.................................................64
Table 9. Statistical Significance Using 95% Confidence Intervals Of Differences Between The Baseline And Theoretical Models...............................................................65
**LIST OF FIGURES**

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Summary statistics and confidence intervals of the true (sensitivity) and false (1 – specificity) positive rates of evaluated ASD screening instruments.</td>
<td>66</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Summary statistics and confidence intervals of the true (sensitivity) and false (1 – specificity) positive rates of the Baseline and theoretical models.</td>
<td>67</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Summary statistics and confidence intervals for the sensitivity of the Baseline and theoretical models when the specificity is .70.</td>
<td>68</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Receiver operating curve (ROC) for the Baseline and theoretical models.</td>
<td>69</td>
</tr>
</tbody>
</table>
CHAPTER 1 – INTRODUCTION

Importance of Assessing Children For Autism

According to the Centers for Disease Control and Prevention (CDC), 1 in 68 American children was diagnosed with one of three autism spectrum disorders [(ASDs); Autistic disorder, Asperger’s Syndrome and Pervasive Developmental Disorder-Not Otherwise Specified (PDD-NOS)] in 2010, which is a 30% increase from 2008 (CDC, 2014). The outcomes for children with one of these three diagnoses vary greatly, as a function of the quality and timing of the treatments they receive (Granpeesheh, Dixon, Tarbox, Kaplan, & Wilke, 2009; Makrygianni & Reed, 2010; National Research Council [NRC], 2001; Sallows & Graupner, 2005; Smith et al., 2010). Early intervention appears to be the best treatment option (Fein et al., 2013; Harris & Handleman, 2000; Luiselli, Cannon, Ellis, & Sisson, 2000; Matson, Mahan, & LoVullo, 2009; NRC, 2001; Smith et al., 2010), making the early identification of ASDs urgent.

Identifying ASDs in children is made difficult due to considerable symptom variability and varying levels of severity (Volkmar, Paul, Klin, & Cohen, 2005), overlapping symptomatology with other disorders (Mayes, Calhoun, Mayes, & Molitoris, 2012), and the occasional late onset of symptoms (National Institute of Mental Health [NIMH], 2011). As a result of these difficulties, the procedure for obtaining a formal diagnosis of an ASD often involves a two-step process, in which standardized assessment instruments are used to screen for the possibility of ASDs and then, if concerns persist, further assessment is undertaken to confirm the diagnosis using criterion instruments. The ability of screening instruments to identify individuals with a disorder can be determined using measures of diagnostic validity, in which the number of individuals identified as either having, or not having, a disorder by a screening instrument is compared to the number of individuals with, and without, a disorder diagnosed.
using a criterion instrument. In this way, performance can be compared and contrasted between optimal and sub-optimal screening instruments.

Previous investigations into comparing screening instruments have relied on interpreting differences in reliability estimates (e.g., internal consistency, interrater reliability, test-rest reliability) between instruments. However, an additional element that varies between instruments is item-content distributions. Measuring the impact of variations in item-content distribution between instruments is appropriate, as it may help explain why some screening instruments demonstrate good performance, and others demonstrate poor performance. Therefore, the current study seeks to answer whether manipulations of item content improve the performance of a poorly performing instrument, in a sample of young children who underwent screening due to a concern about the presence of ASDs. Such a study is timely in light of the recent advent of new diagnostic criteria outlined in the DSM-5, owing to the fact that the criteria change will require the development of new screening instruments. That is, the present study evaluates whether a particular strategy for generating item content for screening instruments results in improved performance, as indicated by higher diagnostic sensitivity and specificity rates.

**Screening Instrument Characteristics**

To evaluate and improve instruments used for diagnosing ASDs, it is important to identify the similarities and differences between instruments that are currently in use. As noted above, a distinction exists between instruments used to screen for ASDs and those used to confirm diagnoses of ASDs. The former are referred to as “screening instruments” and they are the subject of the current paper. Within this class of instruments, important distinctions exist with respect to the age-groups for which they have been developed, the settings in which they are most frequently used, the breadth of disorders they are used to screen for, and the distribution of
items used to assess deficits across diagnostic, and non-diagnostic, domains. Each of these variables is considered, in order to establish a basis for identifying elements of instruments that might be explored to improve upon their diagnostic properties.

**Age Considerations**

Valid screening instruments have been developed and are available for use with children of almost any age (CDC, 2014). In fact, the American Academy of Pediatrics recommends that all children be evaluated for an ASD using screening instruments during well-child checkups at 9, 18, and 24 months (AAP, 2012). Many symptoms of ASDs can be identified at or before the age of 3 and last throughout a person’s life; however, symptoms often do not appear until 24 months or later, and, in some cases, a child begins to lose skills around 24 months (CDC, 2014). For this reason, many ASD screening instruments have been developed for use with individuals aged 3 through adulthood (CDC, 2014), and are the focus of the present study.

**Setting Considerations**

Screening for ASDs can occur in a variety of settings including schools, psychological and psychiatric clinics, and pediatric clinics during well-child checkups (NIMH, 2011). Researchers have made distinctions between screening instruments based on the settings in which they are commonly used. Level 1 screening instruments are those that detect children at risk for ASDs in the general population. As such, they are generally used by primary care physicians in the context of well-child checkups (Robins & Dumont-Mathieu, 2006). One such instrument is the Modified Checklist for Autism in Toddlers (M-CHAT; Robins, Fein, Barton, & Green, 2001). Level 2 screening instruments are those used with children for whom previous interactions suggest they may be at risk, and are most frequently administered in clinical settings by professionals trained to evaluate children for developmental problems (Nah, Young, Brewer,
Level 2 instruments are typically more time-consuming to administer and score. An example of a Level 2 instrument is the Gilliam Autism Rating Scale (GARS; Gillam, 1995). This, and other Level 2 instruments, are the focus of this paper.

**Breadth of Application**

Another factor differentiating screening instruments is the number and nature of the disorders they seek to address. Because of this variability in the number and nature of disorders of interest between instruments, researchers often refer to these as *broad band* and *narrow band* instruments (Glascoe, 2005). Broad band instruments are those that allow for the evaluation of a wide range of behavioral and emotional problems related to several psychiatric and psychological domains. Conversely, *narrow band* instruments are those developed for screening a single psychiatric or psychological domain, such as autism (Glascoe, 2005). Often, broad band instruments include subscales that may function as narrow band instruments. For example, the Child Behavior Checklist/1½-5 (CBCL/1½-5; Achenbach & Rescorla, 2000) qualifies as a broad band instrument because its item content allows for evaluating behaviors and abilities specific to multiple possible diagnoses; however, the pervasive developmental problems (PDP) subscale of the CBCL can be employed as a narrow band checklist because the information assessed is specific to ASD diagnostic criteria. Both broad and narrow band instruments have been utilized and evaluated with respect to making ASD diagnoses, and are the focus of the present study.
CHAPTER 2 – LITERATURE REVIEW

Currently Available Screening Instruments

Screening instruments also differ with regard to their construction and psychometric characteristics. To get a sense of the effectiveness of currently available instruments, six instrument scales were selected and reviewed, based on criteria requiring that instruments assess children ages 1.5 – 5, take the form of a Level 2 instrument, and have been previously evaluated by at least two independent studies. The six scales, listed in Table 1, are reviewed below, with respect to their construction characteristics, such as age range, setting (Level 1 or 2), and breadth (broad or narrow band) considerations, and psychometric characteristics, such as reliability estimates (e.g., internal consistency, test-rest reliability, and interrater reliability), number of instrument items, subscales, and summary scores. Once again, the goal of this review is to identify common factors between high and low performing instruments, which may aid in exploring sources of diagnostic validity differences between instruments.

Social Responsiveness Scale

The Social Responsiveness Scale (SRS; Constantino & Gruber, 2005) is a narrow band instrument designed to identify social impairments and quantify severity of symptoms for persons aged 2.5 through adulthood. It consists of 65 items on which parents rate behaviors within the past six months on a 4-point Likert scale, ranging from “not true” to “almost always true.” Materials are divided into two forms: the Preschool Form, for use with children ages 2.5 – 4.5, and the School-Age Form, for use with individuals aged 4 – 18. Scores corresponding to specific ASD-related domains (social awareness, social cognition, social communication, social motivation and autistic mannerisms) are summed and converted to $T$-scores to provide an index of severity of social deficits. $T$-scores of 60 – 75 are considered typical for children with “high-
functioning” ASDs, and scores 76 or higher are strongly associated with autism (Constantino & Gruber, 2005).

Psychometric information is provided in the clinical manual for the Preschool Form. Results of analyses utilizing data from the normative sample (n = 247) identify high levels of internal consistency, with Cronbach’s alpha values ranging from .92 – .97 across genders, ages, and clinical groups (Constantino & Gruber, 2005). Additionally, the test-retest reliability was identified as high, with a correlation of .88, in a study comparing the results of 30 raters over a period of 137 days (Constantino, Przybeck, Friesen, & Todd, 2000). Finally, interrater reliability was calculated based on responses by parents and teachers (n = 227), and identified as adequate, with a correlation of .77 (Constantino & Gruber, 2005).

**Child Behavior Checklist/1.5-5**

The Child Behavior Checklist/1.5-5 (CBCL/1.5-5; Achenbach & Rescorla, 2000) is a broad band instrument designed to identify behavioral problems and social competencies in children 1.5 – 5 years old. It consists of 99 items on which behaviors are rated on a three-point Likert scale (0 = not true, 1 = somewhat or sometimes true, 2 = very true or often true). Items correspond to seven syndrome scales (emotionally reactive, anxious/depressed, somatic complaints, withdrawn, sleep problems, attention problems, and aggressive behavior) and five DSM-oriented scales (affective problems, anxiety problems, pervasive developmental problems, attention deficit/hyperactivity problems, and oppositional defiant problems). Two global scales provide information about children’s behavioral and emotional problems: the Internalizing scale (emotionally reactive, anxious/depressed, somatic complaints, and withdrawn), and the Externalizing scale (aggressive behavior and attention problems; Achenbach & Rescorla, 2000). All items on the CBCL/1.5-5 are summed to produce a Total Problems scale. Scores ≥70 are
considered clinically significant. Alternate versions of the CBCL are available for screening children older than five, however, only the CBCL/1.5-5 contains both subscales withdrawn and PDP (Achenbach & Rescorla, 2000).

**Withdrawn.** The withdrawn scale is a syndrome scale measure based on responses to 16 items that are oriented to whether a child would rather be alone, is secretive, shows shyness, sulks, or is sad and withdrawn. Based on analyses of the normative sample (n = 700), the internal consistency of the withdrawn scale is identified as moderately high, with a Cronbach’s alpha of .75 (Achenbach & Rescorla, 2000). Additionally, based on ratings by 68 mothers completing the instrument one week apart, the test-retest reliability was also identified as moderately high, with a correlation of .80 (Achenbach & Rescorla, 2000). Finally, the interrater reliability, calculated based on ratings by 72 sets of mothers and fathers, is identified as low, with a correlation of .57 (Achenbach & Rescorla, 2000).

**Pervasive developmental problems.** The pervasive developmental problems (PDP) subscale is a DSM-oriented scale, such that all 13 questions about the child’s behavior that are included in the subscale pertain directly to DSM-IV-TR diagnostic criteria for autistic disorder. Utilizing identical methods employed for the withdrawn subscale, the PDP is identified as having a moderately high internal consistency, with a Cronbach’s alpha of .80, high test-retest reliability, with a correlation of .86, and moderate interrater reliability, with a correlation of .67 (Achenbach & Rescorla, 2000).

**Childhood Autism Rating Scale**

The Childhood Autism Rating Scale (CARS; Schopler, Reichler, & Renner, 1986) is a narrow band instrument designed to assess individuals 2 and older for impairments associated with ASDs, as well as to quantify the severity of these impairments. It consists of 15 subscales
that assess behaviors specific to *DSM-IV-TR* criteria, as well as frequently associated features of ASDs. These subscales consist of: relating to people; imitation; emotional response; body use; object use; adaptation to change; visual response; listening response; taste, smell and touch response and use; fear or nervousness; verbal communication; nonverbal communication; activity level; level and consistency of intellectual response; and general impressions. Items are scored on a 7-point Likert scale ranging from 1 (no impairment) to 4 (severe impairment), with intermediate values of 0.5 points. A combined score is used to classify mild to moderate or severe autism, with cutoffs of 30 and 37, respectively.

Based on relevant data of the development sample (*n* = 1,034), the internal consistency for the total summary score was high, with a Cronbach’s alpha of .93 (Schopler et al., 1986). In addition, test-retest reliability was identified as relatively high, with a correlation of .88, after completion of the instrument for 91 cases, one year apart. Finally, in a sample of 280 sets of independent, trained raters, interrater reliability was relatively high, with a correlation of .84 (Schopler et al., 1986).

**Social Communication Questionnaire**

The Social Communication Questionnaire (SCQ; Rutter, Bailey, & Lord, 2003) is a narrow band instrument that includes 40 yes/no items derived from the Autism Diagnostic Interview – Revised (ADI-R; Lord et al., 1994). The caregiver-completed instrument is intended for screening individuals having a chronological age of over 4 years and mental age of at least 2 years, and consists of four subscales: social interaction, communication, abnormal language, and stereotyped behaviors. Items are summed, and established cutoffs of 22 and 15 are used to place individuals into autism or ASD groups, respectively. Two forms are available for screening: a
lifetime form, focusing on the entire developmental history, and current form, referencing the
previous three-month period.

Results utilizing normative sample data (n = 214) found in the clinical manual indicate
moderate-to-high levels of internal consistency across ages 2 – 18, with Cronbach’s alpha values
ranging .84 – .93 (Rutter, Bailey, & Lord, 2003). Test-retest and interrater reliability estimates
are not available for the SCQ.

**Gilliam Autism Rating Scale**

The Gilliam Autism Rating Scale (GARS; Gilliam, 1995) is a narrow band instrument
designed to screen for ASDs among individuals 3 – 22 years of age. It consists of 42 items on
which parents and teachers rate behaviors on a four-point Likert scale, ranging from “never
observed” to “frequently observed.” Items are grouped into three subscales: communication,
social interaction, and stereotyped behaviors. A primary summary score, called the Autism
Quotient (AQ; \( M = 100, \ SD = 15 \)) is used to predict diagnostic group placement, with a cutoff
score of 90 indicating that child is “probably autistic” (Gilliam, 1995).

Psychometric characteristics for the GARS are provided in the clinical manual. Utilizing
data from the normative sample (n = 1,092), the internal consistency for the GARS is high, with
Cronbach’s alpha values for subscales and summary scores ranging from .88 – .96 (Gilliam,
1995). Additionally, based on results of raters (n = 11) completing the instrument twice, one
week apart, the test-retest reliability of the GARS subscales and summary scores are identified as
moderately high, with correlations ranging from .81 – .88 (Gilliam, 1995). Finally, interrater
reliability between parents and teachers (n = 57) was moderately high for subscales and the AQ,
with correlations ranging from .73 – .88 (Gilliam, 1995).
Diagnostic Validity

Having reviewed important instrument characteristics, it is necessary to examine the criteria used for determining the effectiveness of instruments for distinguishing between autistic and non-autistic individuals (diagnostic validity). The various metrics for assessing diagnostic validity include the following.

**Sensitivity and Specificity**

The two most common metrics for judging the diagnostic validity of instruments are sensitivity and specificity. Sensitivity refers to the proportion of children with a disorder correctly identified as such. Specificity refers to the proportion of children without a disorder correctly identified as such. Although open to some debate, current consensus is that clinically acceptable sensitivity rates are between .70 – .80, and clinically acceptable specificity rates are closer to .80 (Coonrod & Stone, 2005; Glascoe, 2005; Norris & Lecavalier, 2010). However, as there is commonly a tradeoff between sensitivity and specificity, it is most important for screening instruments to err on the side of sensitivity rather than specificity, as incorrectly identifying a child as not having a disorder (a false negative) can result in missed opportunities for early intervention (South et al., 2002). Given the current standard that involves the utilization of screening and confirmatory instruments, eliminating false positives within the confirmatory phase of the diagnostic process is an important assessment outcome.

**Summary statistics.** Instruments assessed via reports of sensitivity and specificity are useful, in that they give reliable estimates of diagnostic validity. However, due to variations between results, interpreting the general performance of instruments across multiple sensitivity and specificity reports can be difficult. Utilizing a meta-analytic procedure described by Deeks (2001), a single point on a graph can aid in conducting easy comparisons of diagnostic validity.
between instruments. This procedure for calculating average sensitivity and specificity values first involves calculating the true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) of individual studies. The true positives of each study are calculated by multiplying the sample size of each ASD group (identified by criterion instruments) by the sensitivity reported within the study. False negatives for each study are calculated by subtracting the newly calculated TP value from the ASD sample size. The true and false negatives for each study are obtained through similar means. Specifically, multiplying the sample size of each non-ASD group by the specificity reported within studies yields true negatives, and the difference between the non-ASD sample size and the newly calculated true negative value yields false negatives. Fourth, the true and false positives and negatives of each study are summed to form total true and false positives and negatives. Using the basic equations for calculating sensitivity \([TP/(TP + FN)]\) and specificity \([TN/(TN + FP)]\), the total true and false positives and negatives are used to produce summary statistics for instrument sensitivity and specificity scores.

**Positive and Negative Predictive Value**

The diagnostic validity of an instrument can be further described in terms of *positive* and *negative predictive value* (Glascoe, 2005; Norris & Lecavalier, 2010). Positive predictive value (PPV) is the proportion of individuals identified by a screening instrument as having a disorder who are later diagnosed with the disorder using criterion instruments. Negative predictive value (NPV) is the proportion of individuals identified by a screening instrument as not having a disorder that are later confirmed to not have the disorder using criterion instruments. There are no universally agreed upon standards for PPV or NPV; however, each of these indicators is useful for consideration when evaluating the diagnostic validity of a test, as the values provide information about the prevalence of a disorder in a given setting.
Diagnostic validity estimates (sensitivity, specificity, PPV and NPV) also differ at the equation level. Calculating these equations once again involves the use of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN). Thus, calculation of sensitivity \[
\frac{TP}{TP + FN}\]
involves dividing the number of individuals correctly identified as having a disorder (identified by criterion instruments) by the total number of individuals classified by the criterion instrument as having the disorder, regardless of the findings by the screening instrument. In contrast, PPV \[
\frac{TP}{TP + FP}\]
involves dividing the number of individuals correctly classified with the screening instrument as having a disorder by the total number of individuals identified as having the disorder by the screening instrument, regardless of the diagnosis made by the criterion instrument. Similarly, the specificity \[
\frac{TN}{TN + FP}\]
and NPV \[
\frac{TN}{TN + FN}\]
differs such that specificity involves dividing the number of individuals correctly identified as not having a disorder by the number of individuals diagnosed as not having the disorder by the criterion instrument, and NPV involves dividing the number of individuals correctly classified as not having a disorder by the number of individuals identified as not having the disorder by the screening instrument. One critical element differentiates sensitivity and specificity from PPV and NPV: whereas sensitivity and specificity are independent from the prevalence of a disorder in a given population, PPV and NPV are highly dependent on the prevalence of a disorder in the population. Therefore, all four values should be reported whenever possible to fully describe the a given sample.

**Receiver Operating Characteristic (ROC)**

A third method of estimating diagnostic validity, ROC, involves plotting the performance of an instrument with respect to accurately identifying a given disorder (Cicchetti, Volkmar, Klin, & Showalter, 1995). This graphical representation of the accuracy of a test in differentiating
groups with and without a disorder involves placing sensitivity on a y-axis and (1-specificity) on an x-axis. Using this graph, the performance of a test can then be mapped by placing a single sensitivity by (1-specificity) point on the graph for each cutoff score for the instrument. Connecting the points creates the ROC curve, from which an Area Under the Curve (AUC) can be determined that allows for determining the presence of a disorder. Based on pre-established standards, AUC values are deemed poor (< .70), fair (.70 – .79), good (.80 – .89), or excellent (.90 – 1.00; Cicchetti, Volkmar, Klin, & Showalter, 1995).

**Diagnostic Validity of Currently Available Screening Instruments**

The goal of the current review was to identify common factors between high and low performing instruments that may help explain differences in the diagnostic validity of instruments. Following the review of psychometric and construction characteristics of instruments, the diagnostic validity of instruments was evaluated, in order to determine which are high, and which are low, performing instruments. Therefore, findings regarding the sensitivity, specificity, PPV, NPV and AUC of each instrument are described below and listed in Table 1.

**Social Responsiveness Scale**

Three independent studies evaluated the diagnostic validity of the SRS. Constantino et al. (2007) evaluated the SRS in a sample of 577 children and adolescents aged 4 – 18 receiving outpatient treatment for diagnosed psychiatric disorders (autism prevalence = .61). Using the ADI-R and ADOS to assess individuals involved in the study, the researchers found high levels of sensitivity (.75), specificity (.96), PPV (.97), and NPV (.71), as well as an AUC within the excellent range (.95). Schanding et al. (2011) evaluated the SRS in a sample of 3,375 children and adolescents aged 3 – 17, who participated in a multi-site genetics research study.
investigating families with only one child with an ASD (autism prevalence = .52). Individuals were assessed using both the ADI-R and ADOS. Schanding et al. (2011) identified high sensitivity (.95), specificity (.96), PPV (.96) and NPV (.96). Aldridge, Gibbs, Schmidhofer, and Williams (2012) investigated the SRS in a sample of 48 children aged 4 – 15 years, referred for evaluation due to social concerns. Aldridge et al. (2012) identified a high sensitivity (.91) and PPV (.71), but a low specificity (.08) and NPV (.25).

In summary, all three studies found sensitivity rates (Aldridge et al., 2012; Constantino et al., 2007; Schanding et al., 2011) within the 70 – 80% “acceptable” range, and two of the three studies found specificity rates (Constantino et al., 2007; Schanding et al., 2011) within the “acceptable” range. Only one study included an ROC analysis, which obtained an “excellent” estimate of AUC.

**Child Behavior Checklist/1.5-5**

Four independent studies have identified the diagnostic validity of the CBCL/1.5-5 in screening for children with ASDs. Three of the four studies investigated both the withdrawn and PDP subscales with respect to their usefulness for screening for ASDs (Muratori et al., 2011; Narzisi et al., 2013; Sikora et al., 2008), and one study investigated only the PDP subscale (Predescu, Sipos, Dobrean, & Miclutia, 2013). Results of these studies generally conclude that these two scales are valid for autism screening. For instance, Narzisi et al. (2013) compared the results of the CBCL and the Autism Diagnostic Observation Schedule – Generic (ADOS-G; Lord et al., 2000), a highly regarded and empirically validated confirmatory assessment, in a sample of 94 children aged 1.5 – 3 years who were referred over a two-month period for autism evaluation (autism prevalence = .50). They found the sensitivity rate of both the withdrawn (.92) and PDP (.98) subscales to be above the .70 – .80% “acceptable” range for sensitivity (Coonrod
Similarly, the specificity (.97 & .91), PPV (.98 & .92), NPV (.92 & .98), and AUC (.99 for both) of each subscale was high. Sikora et al. (2008) also compared the results of the CBCL subscales to those of the ADOS-G, in a sample of 147 children aged 16 – 71 months who were referred for evaluation due to suspicion of autism (autism prevalence = .73). They found that the sensitivity rate of the PDP (.78) and PPVs of each subscale (.82 & .78) were high. However, the sensitivity of the withdrawn (.65) subscale, and specificity (.62 & .42) and NPVs (.40 & .41) of both subscales were low. Muratori and colleagues (2011) also investigated the sensitivity and specificity rates for the PDP and withdrawn scales for a sample of 313 children aged 24-60 months hospitalized for problems related to autism or other psychiatric disorders. Similar to Sikora et al. (2008), the ADOS was used to place children into diagnostic groups (autism prevalence = .46). Results revealed high sensitivity and specificity scores for both the withdrawn (.89 & .92) and PDP (.85 & .90) subscales. Muratori et al. (2011) further obtained high PPVs (.90 & .88), NPVs (.90 & .87), and AUCs (.95 for both) for each subscale. Predescu et al. (2013) evaluated the CBCL PDP subscale in a sample of 233 children aged 1.5 – 5 years, who received outpatient services for diagnoses based on clinical judgment of autism, Attention Deficit/Hyperactivity Disorder (ADHD), or other developmental delays (autism prevalence = .60). Predescu et al. (2013) identified high sensitivity (.96), specificity (.89), PPV (.93), NPV (.94), and AUC (.97) rates for the PDP subscale.

In summary, two of the three studies that examined the withdrawn subscale found sensitivity and specificity rates within the “acceptable range”. Additionally, for the PDP subscale, all four studies found sensitivity rates, and three of the four studies found specificity rates, within or beyond the .70 – .80 “acceptable” range (Coonrod & Stone, 2005; Glascoe, 2005; Norris &
Lecavalier, 2010). All of the studies that conducted ROC analyses identified high AUC scores. These results suggest that both scales, particularly the PDP, are good measures for screening autism.

**Childhood Autism Rating Scale**

Four independent studies have evaluated the diagnostic validity of the CARS. L. C. Eaves and Milner (1993) evaluated the CARS in a sample of 77 children and adults aged 2 – 21 years, who were either previously diagnosed and being treated for ASDs, or classified into categories other than ASDs (autism prevalence = .62). Diagnostic group membership was based on previous evaluations giving official diagnoses or suggestion of autism. L. C. Eaves and Milner (1993) found that 98% of individuals with ASDs were correctly identified. Perry, Condillac, Freeman, Dunn-Geier and Belair (2005) screened 274 children aged 2 – 6 who were referred for diagnostic assessment due to suspicion of autism with the CARS (autism prevalence = .35). Children were placed into diagnostic groups based on the clinical judgment of an experienced clinician. Perry et al. (2005) found “acceptable” sensitivity and specificity scores (.94 and .85, respectively), an excellent AUC score (.95), and a PPV and NPV of .77 and .96, respectively. Ventola et al. (2006) investigated the CARS groupings against clinical judgment diagnoses in a sample of 45 children aged 16 – 30 months who were referred for having failed the M-CHAT (Robins et al., 2001; autism prevalence = .60). Using the recommended cutoff score, Ventola et al. (2006) found a high sensitivity (.96), specificity below the “acceptable” 70 – 80% range (.67), a PPV of .81 and a NPV of .69. Finally, Chlebowsky, Green, Barton, and Fein (2010) investigated the CARS in samples of 243 2-year-olds and 248 4-year-olds, who were referred for evaluation due to having failed the M-CHAT (autism prevalence = .58 and .70, respectively). Group membership was based on results of the ADOS and ADI-R. Using the recommended
cutoff score in the 2-year-old sample, Chlebowsky et al. (2010) identified a high sensitivity rate (.93), and specificity rate below the “acceptable” range (.49). In the 4-year-old sample, the recommended cutoff score performed optimally, and both the sensitivity (.86) and specificity (.80) were found to be within the “acceptable” range. The PPVs ranged from .72 to .91, and the NPVs ranged from .71 to .83, across age groups.

In summary, all four studies found sensitivity rates for the CARS within the .70 – .80 “acceptable” range. Additionally, all three studies that investigated the specificity of the CARS found rates within the acceptable range (Chlebowsky et al., 2010; Perry et al., 2005; Ventola et al., 2006).

**Social Communication Questionnaire**

Nine independent studies have investigated the diagnostic validity of the SCQ. Witwer and Lecavalier (2007) assessed 98 children aged 4 – 14, who were identified by their school as having a diagnosis of PDD (autism prevalence = .74); no method of evaluation was used to confirm PDD diagnoses. Witwer and Lecavalier (2007) identified a high sensitivity (.92), PPV (.87), NPV (.73) and AUC value (.89), but a low specificity (.62). Corsello et al. (2007) investigated the utility of the SCQ in a sample of 590 individuals, aged 2 – 16, who were referred to autism centers for testing and research (autism prevalence = .74). Using the ADI-R and ADOS to assess children, the researchers identified a high sensitivity (.71), specificity (.71), and PPV (.88), a fair AUC value (.77) and low NPV (.45). L. C. Eaves, Wingert and Ho (2006a) analyzed the SCQ in a sample of 94 children aged 4 – 6 who were referred for assessment due to suspicion of autism (autism prevalence = .54). Children were assessed for ASDs based on the clinical judgment of a multidisciplinary team utilizing data from the Childhood Autism Rating Scale (CARS). L. C. Eaves et al. (2006a) found a high sensitivity (.74) and NPV (.78) and a low
specificity (.54) and PPV (.65). L. C. Eaves, Wingert, Ho, and Mickelson (2006b) further evaluated the SCQ in a sample of 151 children ranging 36 – 82 months, who were referred to either an autism or preschool clinic due to suspicion of autism (autism prevalence = .35). Children were placed into groups utilizing the same method used by L. C. Eaves et al. (2006a); the researchers identified high sensitivity (.71), specificity (.79), and NPV (.84) and low PPV (.65). Allen, Silove, Williams and Hutchins (2007) investigated the utility of the SCQ in a sample of 81 2 – 6 year olds who were referred for assessment due to suspicion of autism (autism prevalence = .35). Children were placed into diagnostic groups based on the clinical judgment of a multidisciplinary team. Allen et al. (2007) found a high specificity (.70) and NPV (.76), and a low sensitivity (.60) and PPV (.52). Lee, David, Rusyniak, Landa, and Newschaffer (2007) evaluated 268 children aged 3 – 5, who were receiving services for preschool special education (autism prevalence = .20). Diagnostic group membership was based on reports by schools and/or scores on the ADOS and ADI-R. Lee et al. (2007) found a high specificity (.92) and NPV (.89), good AUC (.88) and low sensitivity (.54) and PPV (.62). Wiggins, Bakeman, Adamson, and Robins (2007) compared SCQ and ADOS results in a sample of 37 children aged 17 – 45 months, who were participating in a statewide, early intervention program (autism prevalence = .51). Wiggins et al. (2007) found a high specificity (.89) and PPV (.82), and low sensitivity (.47) and NPV (.62). Snow and Lecavalier (2008) screened 65 children aged 30 – 70 months, who were referred to a hospital for possible ASDs (autism prevalence = .62). Each child was assessed with the ADOS by a multidisciplinary team. Snow and Lecavalier (2005) found a high sensitivity (.70) and PPV (.70), and low specificity (.52), NPV (.52) and AUC (.63). Finally, Oosterling et al. (2010) compared the results of the SCQ with those of the ADI-R and ADOS in 208 toddlers aged 20 – 40 months, who were referred for assessment due to suspicion of autism (autism
prevalence = .69). The researchers found a high specificity (.94) and PPV (.96), and low sensitivity (.63), NPV (.54) and AUC (.64).

In summary, five of the nine studies found sensitivity rates for the SCQ within the .70 – .80 “acceptable” range (Corsello et al., 2007; Eaves et al., 2006a; Eaves et al., 2006b; Snow & Lecavalier, 2008; Witwer & Lecavalier, 2007). Similarly, five of the nine studies investigating the diagnostic validity of the SCQ identified specificity scores within the “acceptable” range (Allen et al., 2007; Corsello et al., 2007; Eaves et al., 2006b; Lee et al., 2007; Wiggins et al., 2007). Finally, of the five studies that included ROC analyses, two identified “poor” levels of AUC (Oosterling et al., 2010; Snow & Lecavalier, 2008), one a “fair” AUC estimate (Corsello et al., 2007), and two “good” (Lee et al., 2007; Witwer & Lecavalier, 2007).

**Gilliam Autism Rating Scale**

Five studies investigated the diagnostic validity of the GARS. Three of the five studies examined both the sensitivity and specificity rates of the GARS (R. C. Eaves et al., 2006; Mazefsky & Oswald, 2006; Sikora et al., 2008), whereas two studies examined only sensitivity (Lecavalier, 2005; South et al., 2002). South et al. (2002) investigated only the sensitivity of the GARS by comparing the results of the instrument to findings by the ADOS-G and/or ADI-R in a sample of 119 children, aged 3 – 10.5, who were part of research or treatment associated with autism from four universities (autism prevalence = 1.00); only 48% of children were correctly classified. Lecavalier (2005) evaluated the diagnostic validity of the GARS in sample of 284 students from 29 different school districts, aged 3 – 21, who were receiving educational services for ASDs (autism prevalence = 1.00). Comparing the results to previous classifications by physicians or as listed in Individual Education Plans (IEPs), the author reported a low sensitivity (.38) and theorized that the GARS overemphasized evaluating restricted, repetitive and
stereotyped patterns of behaviors, interests and activities (RRSBIA), while failing to fully evaluate social and communicative areas (Lecavalier, 2005). R. C. Eaves et al. (2006) evaluated the GARS in a sample of 134 children and adults aged 3 – 26 previously diagnosed and receiving services for Pervasive Developmental Disorders (PDDs) or other disorders commonly confused with PDDs (autism prevalence = .83). Using the recommended cutoff score of 90, R. C. Eaves et al. (2006) identified a sensitivity of .79, a specificity of .68, a PPV of .92 and a NPV of .40.

Mazefsky and Oswald (2006) evaluated the GARS in a sample of 78 children aged 22 months to 8 years who were referred for evaluation due to suspicion of PDDs in a university medical setting (autism prevalence = .76). Comparing the results to findings by a diagnostic team, the authors found low sensitivity and specificity values of .39 and .61, respectively, and obtained a PPV and NPV of .76 and .24, respectively. Sikora et al. (2008) assessed the results of the GARS against findings by the ADOS-G in a sample of 147 children aged 3 – 5 years, who were referred for assessment due to suspicion of autism (autism prevalence = .73). Sikora et al. (2008) found both sensitivity and specificity values to be low (.53 and .54, respectively), and obtained a PPV of .76 and a NPV of .29.

In summary, four of the five studies found the diagnostic validity of the GARS to be lacking, as sensitivity and specificity values were below the .70 – .80 “acceptable” range (Lecavalier, 2005; Mazefsky & Oswald, 2006; Sikora et al., 2008; South et al., 2002). Only R. C. Eaves et al. (2006) found a sensitivity value within the “acceptable” range; however, the specificity was low.

Summary Statistics of Currently Available Instruments

Using the methodology identified above, based on procedures described by Deeks (2001), summary statistics for the sensitivity and specificity of each of the six Level 2 screening scales
that qualified for review were created. Results of this analysis are included in Table 2.

Additionally, to aid in the visual inspection of diagnostic validity, each of the summary statistics was plotted, such that the true positive rates, or sensitivity values, were placed on a y-axis and the false positive rates, or 1-specificity values, were put on an x-axis, shown in Figure 1. Each plot point also includes a confidence interval, which illustrates the array of sensitivity and specificity results identified for each instrument.

As can be gleaned from Table 2 and Figure 1, five of the six scales obtained average sensitivity and specificity scores within the “acceptable” range (.70 – .80). The only instrument that did not obtain acceptable sensitivity or specificity scores was the GARS. Comparing across the remaining instruments, the CARS, SRS, and PDP subscale of the CBCL achieved both sensitivity and specificity scores of .80 or higher, which qualifies them as meeting high clinical standards for screening instruments (Coonrod & Stone, 2005; Glascoe, 2005; Norris & Lecavalier, 2010).¹

**Item Content Differences and Validity**

There are a variety of factors that can contribute to the differences in diagnostic validity that exist between instruments. To the extent that instruments illustrate differential validity, it makes sense to identify contributors to those differences. One potential contributor to validity differences across instruments is item content differences. An aspect of item content relevant to the diagnostic process concerns how the item content of the instruments relates to the distribution of symptoms listed in diagnostic manuals (e.g., the *DSM-IV-TR*, *DSM-5*, ICD-10, etc.). For example, the *DSM-IV-TR* organizes autism-relevant symptoms and characteristics into three

¹ Results were further analyzed by including only studies that utilized the ADOS or ADI-R as the criterion measure. This analysis, however, excluded only four studies that evaluated the GARS (1 study), SCQ (2 studies), and CARS (1 study). Summary statistic calculations did not significantly differ when including or excluding those studies.
diagnosis-relevant domains, and a formal diagnosis of autism requires that a certain number of symptoms be present within each of these domains. These domains include social interaction symptoms, communication symptoms, and RRSBIA symptoms. A formal diagnosis requires six related symptoms, with at least two social interaction symptoms and one, each, of communication and RRSBIA symptoms. In addition, the *DSM-IV-TR* identifies non-diagnostic, associated features of autism (such as play and creativity), the presence of which should be noted, but that are not counted when making a formal diagnosis. The diagnostic criteria listed for autism spectrum disorder (ASD) in *DSM-5* are similarly organized, although the symptoms differ slightly and the social problems and communication problem symptoms are collapsed into a single category. However, similar to the diagnostic layout identified in *DSM-IV-TR*, a diagnosis of ASD in accordance with *DSM-5* requires specific numbers of symptoms within each domain (i.e., 3 symptoms in the communication and social interaction domain, and at least 2 symptoms in the RRSBIA domain).

It is also possible that the best performing instruments will show a high degree of calibration not with the DSM, but rather with whatever instrument serves as the criterion measure within the study. For 14 of the 21 studies across the six scales, the criterion measure that was in use was the ADOS. Across these studies, each child assessed was evaluated using one of four ADOS modules, depending on the child’s verbal ability. For the population of interest in the current study (children ages 1.5 – 5), ADOS administrations consisted of Modules 1 and 2. Interestingly, all four modules of the ADOS include not only items that align with the three symptom domains of the *DSM-IV-TR*, but also take into account non-diagnostic, associated features. Additionally, similar to *DSM-IV-TR*, all four modules of the ADOS, but especially Modules 1 and 2, oversample social interaction symptoms relative to communication and
RRSBIA symptoms. Therefore, it may be the case that instruments that oversample social interaction-related items and include at least some items that tap non-diagnostic, associated features may demonstrate better diagnostic validity than those that do not (Lloyd, MacDonald, & Lord, 2013). Consistent with these speculations, I expect that instruments with higher diagnostic validity will have item distributions that more closely resemble the symptom distributions of the *DSM-IV-TR* and the ADOS than will instruments with lower diagnostic validity.

**Item Distribution of Currently Available Instruments**

Qualitative analyses were performed to explore whether the item content distributions of better performing instruments are more closely related to *DSM-IV-TR* and ADOS symptom distributions, compared to poorly performing instruments. In order to determine how the item content of instruments related to the content domains of the *DSM-IV-TR*, the item content of each instrument was categorized with respect to how it was distributed across the *DSM-IV-TR* domains (social interaction, communication, RRSBIA, and non-diagnostic, associated features). The percentage of items for each instrument that align with these domains are displayed in Table 3. For all instruments except the CARS, information about how the items align with the *DSM-IV-TR* domains were obtained from the user manual of the instrument. In the case of the CARS, that information was obtained from Goldfischer (2001). Also included in the table is the distribution of symptoms identified by the *DSM-IV-TR* and ADOS items (Modules 1 and 2) as they relate to the four *DSM-IV-TR* domains. With respect to the *DSM-IV-TR*, relevant symptoms to each domain are identified, for which specific requirements are necessary for an ASD diagnosis to be appropriate. As noted previously, these specific symptom requirements include having two social interaction symptoms, and at least one symptom from each of the communication and RRSBIA domains. Therefore, despite there being a similar number of symptoms across the three domains,
the social interaction domain is effectively more heavily weighted than are the other domains with respect to obtaining a formal diagnosis.

Investigating whether the alignment of item content of screening instruments with DSM symptom content is associated with diagnostic validity involved two steps. In the first step, content differences were investigated using as a starting point the worst performing instrument (GARS). In the second step, content differences were investigated using the best performing instruments (CARS, SRS, and PDP). The goal of these observational analyses was to identify if item content differences could be identified that distinguish between low and high performing instruments, respectively.

With regard to the first comparison, it is clear from Table 3 that the GARS is unique among instruments in that its items are evenly divided across the DSM-IV-TR domains. In contrast, a characteristic of all the other instruments is a greater proportion of items devoted to evaluating the social interaction domain compared to those devoted to evaluating the communication or RRSBIA domains (except for the CBCL PDP subscale, which has an equivalent proportion of RRSBIA items). These observations suggest that the inclusion of an equal number of items across the four domains may not provide the most effective method for screening for autism, and that relative to the other instruments the GARS under-samples social interaction symptoms.

The second step focused on how the item content of the best performing instruments (CARS, SRS, and CBCL PDP subscale) differed from the other instruments. Although there were no easily identifiable patterns with respect to both of these instruments, the CARS is unique with respect to a greater focus on non-diagnostic, associated features (36% of items). That is, a plurality of items for the CARS assess aspects of functioning that are not diagnostic symptoms of
ASDs. This observation suggests that there may be a role in autism screening for items that do not align with the formal diagnostic criteria for the disorder but, rather, are non-diagnostic, associated features.

In sum, the results of the above analyses support the notion that instruments which include items that align well with prevailing diagnostic criteria and/or the criterion measure against which the instruments are assessed have higher validity than those that do not. That is, neither the *DSM-IV-TR*, nor the ADOS sample evenly across the symptom domains. Instead, a formal autism diagnosis based on *DSM-IV-TR* criteria requires the presence of two of four symptoms in the social interaction domain, and only one of four in the communication and RRSBIA domains. Therefore, it makes sense that those instruments that oversample social interaction skills may outperform those that do not, to the extent that DSM criteria play some role in diagnostic decision-making. Similarly, it is also true that the ADOS – the instrument that serves as the criterion measure for most validity studies reviewed – over-samples social interaction symptoms compared to communication and RRSBIA symptoms. Specifically, an item analysis of the ADOS reveals that the plurality (41% [12 items]) of items are concerned with evaluating social interaction abilities. Therefore, the more extensive the sampling of social interaction deficits, the higher the likelihood that the weighting of symptoms will be well-aligned with the *DSM-IV-TR* and ADOS.

Another finding of the qualitative analysis of item content revealed that an instrument illustrating strong validity was comprised of a plurality of items (36%) concerned with non-diagnostic, associated features of ASDs. The inclusion of non-diagnostic, associated features is consistent with the structural characteristics of the ADOS, for which 5 of 29 ADOS items (17%) are concerned with non-diagnostic, associated features of autism. These ADOS items include a
focus on play and creativity, over-activity, anxiety, tantrums, and aggressive behaviors. It is noteworthy, then, that the non-diagnostic, associated features included in the ADOS are relevant to emotional self-regulation abilities. Such self-regulation abilities fall under a larger umbrella of mental processes that aid in achieving future goals, known as executive functions (Luria, 1969).

**Executive Functioning in ASD Screening**

Executive functions typically develop in early childhood (Buckner, Mezzacappa, & Beardslee, 2003) and are theorized as fundamental for academic, social and overall development (Barkley, 2001). Impairments in executive functioning often have wide-reaching effects (Luria, 1969), but typically include having difficulties in the following abilities: (a) controlling impulses and behavior; (b) making transitions from one activity or aspect of a problem to another as the situation demands; (c) modulating emotional responses appropriately to situational demands or contexts; (d) holding information in mind for the purpose of completing a task or making the appropriate response; and (e) anticipating future events or consequences, using goals or instructions to guide behavior (Gioia, Espy & Isquith, 2003).

Because of the clear links that exist between autistic behaviors and executive functioning impairments, including items that assess executive functioning deficits may improve the diagnostic validity of ASD instruments. The basis for this prediction that the inclusion of executive functioning items will improve diagnostic validity is that *DSM-IV-TR* (APA, 2000) and *DSM-5* (APA, 2013) identify several non-diagnostic, associated features of ASDs that are related to deficits in executive functioning, such as difficulties with attention and hyperactivity. Additionally, of the many features of ASDs commonly observed that fall outside the diagnostic domains, executive functioning impairments are frequently implicated as the source of many autistic behaviors (Liss, et al., 2001; Ozonoff, Pennington, & Rogers, 1991; Pennington et al.,
Impairments in the executive functioning ability to shift from one activity to another, for example, are closely related to the insistence on routine and sameness commonly observed in children with ASDs.

The role of executive functioning sub-components in ASDs have been extensively researched in adolescents and adults (Bennetto et al., 1996; Griffith et al., 1999; Liss et al., 2001; McEvoy et al., 1993; Ozonoff et al., 1991; Ozonoff et al., 1994; Ozonoff & McEvoy, 1994; Pennington et al., 1997; Prior & Hoffman, 1990; Russell, 1997; Russell et al., 1996; Schneider & Asarnow, 1987). These studies have found a split between sub-components, such that individuals with ASDs perform poorly on tasks relating to cognitive flexibility (Griffith et al., 1999; McEvoy et al., 1993; Ozonoff et al., 1994; Schneider & Asarnow, 1987), planning (Ozonoff et al., 1991; Ozonoff & McEvoy, 1994; Prior & Hoffman, 1990) and self-regulation (Gomez & Baird, 2005), but do not demonstrate significantly impaired abilities to inhibit responses, as compared to typically developing children (Ozonoff et al., 1994). Additionally, working memory skills have been found as deficient in some studies (Bennetto et al., 1996), but not in others (Russell et al., 1996). Due to these varied results, evaluation of executive functioning deficits would be best achieved by evaluating executive functioning sub-components. Therefore, I expect that ASD-specific screening instruments will be improved by looking closely at the sub-components of executive functioning, rather than the global effects of executive functioning as a whole.

**Purpose of Current Research**

The proposition that improved diagnostic validity of ASD screening instruments may be obtained from manipulating the item content of those instruments has, to my knowledge, not been empirically evaluated. However, this item content evaluation may provide insights into an
effective strategy for developing new instruments with high validity, and for improving existing instruments that have less than adequate validity. This idea will be tested using a poorly performing instrument, the Gilliam Autism Rating Scale – Second Edition (GARS-2; Gilliam, 2006). It is the aim of the current study to investigate whether the item content of screening instruments should (1) mirror the item content organization of “gold standard” instruments and diagnostic structure found in DSM-IV, and (2) include items that assess executive functioning deficits. If the hypotheses are confirmed, the evaluated methodology would be identified as useful for the creation of future instruments, and for improving, via an altered scoring framework, a popular and commercially available instrument (GARS-2).

**Hypotheses**

In the current study, three hypotheses will be evaluated via creation and examination of the validity of theoretical instrument models. First, it is predicted that amplifying the emphasis on social interaction skills by increasing the weighting of existing items that assess those skills will significantly improve diagnostic validity relative to diagnostic validity found for the unmodified GARS-2 instrument (improve sensitivity with a minimal decrease in specificity). Second, it is predicted that adding executive functioning subscale items to the item profile of the GARS-2 will further improve diagnostic validity. Specifically, I hypothesize that by adding items from the broad index of the Behavior Rating Inventory of Executive Functioning – Preschool Version (BRIEF-P; Gioia, Espy & Isquith, 2003) that assess shift and emotional control abilities, the sensitivity of the GARS-2 will be significantly improved. Third, it is predicted that a blending of these two transformations just described will provide improvements to diagnostic validity that are greater than those obtained for either of the two previously-mentioned transformations, separately. Therefore, I expect that by blending together the two
methods described above to mirror the item-content found for the criterion measure used in this study (the ADOS), the sensitivity of the GARS-2 will be significantly improved.
CHAPTER 3 – METHODOLOGY

Participants

The current study examined de-identified, archived data for 136 children between the ages of 3 – 10 years, who were referred to a not-for-profit clinic that provided services to families of children with developmental disabilities in a community of 230,000 people in the Pacific Northwest. Each of these children was referred between the years 2008 and 2014 for suspicion of autism and evaluated by a comprehensive autism disorder evaluation team (CADET) using the ADOS-G, and a battery of caregiver-completed screening instruments that included the BRIEF-P and GARS-2. Of the 136 children evaluated with the GARS-2, 114 were evaluated with the BRIEF-P. Pair-wise procedures were utilized to deal with missing data, meaning that participants were included in any analysis for which the necessary data was available.

Participant characteristics are identified in Table 4. The final sample included 110 boys and 26 girls, with an average age of 4 years, 6 months (SD = 1.28). Additionally, 79 children (58% of the total sample) were diagnosed with, and 57 (42% of the total sample) without, ASD based on ADOS-G criteria. Of the 57 children placed into the non-spectrum group, 14 (25% of the non-spectrum group) were diagnosed with ADHD, 10 (17% of the non-spectrum group) with disruptive behavior disorder (DBD), 14 (25% of the non-spectrum group) with mixed receptive-expressive language disorder (Language) and 19 (33% of the non-spectrum group) with an anxiety disorder not otherwise specified (Anxiety). Finally, in addition to the primary diagnoses given, 79 children met criteria for comorbid diagnoses, including ADHD (12 children; 15% of the sample), DBD (16 children; 20% of the sample), mixed receptive-expressive language disorder (Language; 18 children; 23% of the sample), anxiety disorder not otherwise specified
(Anxiety; 10 children; 13% of the sample), intellectual disability (ID; 16 children; 20% of the sample), developmental coordination disorder (DCD; 4 children; 5% of the sample), and an “Other” category that encompasses seizure (2 children) and adjustment (1 child) disorders (4% of the sample).

Measures

Autism Diagnostic Observation Scale – Generic

The Autism Diagnostic Observation Schedule – Generic (ADOS-G; Lord et al., 2000) is a standardized, semi-structured observation protocol of social and communication abilities. It is administered in a one-to-one format that includes an assessor and the child of interest and takes approximately 30 – 45 minutes to complete. Four distinct modules can be used, depending on a child’s verbal ability. Examiners take notes during administration about the child’s behavior in relation to four domains: communication, social interaction, play/creativity, and restricted/repetitive behaviors or interests. The test is immediately scored after administration on a 0 – 3 scale (0 = No evidence of abnormal behavior, 3 = Markedly abnormal behavior). Cut-off scores in the areas of communication, social interaction, and combined (communication + social interaction) allow the child to be placed into a category of autism, ASD, or non-spectrum, based on normative data.

The authors of the ADOS-G report excellent diagnostic validity for the instrument, such that, in a sample of 213 children clinically evaluated for ASDs, the ADOS-G was found to have .90 – .97 sensitivity rates and .87 – .94 specificity rates across modules (Lord et al., 2000). Reliability estimates have been reported as similarly high for the communication-social interaction totals. Specifically, the internal consistencies across modules (Cronbach’s alphas = .91 – .94), test-retest reliability in a sample of 27 children tested twice over an average period
of nine months (correlation = .82) and interrater reliability in a sample of 97 raters (correlation = .92) are high (Lord et al., 2000).

**Gilliam Autism Rating Scale – Second Edition**

The Gilliam Autism Rating Scale – Second Edition (GARS-2; Gilliam, 2006) is a parent-report instrument designed specifically to screen for ASDs. It is nearly identical to the GARS (described above), with a few exceptions. First, minor changes were made to the wording of the first 42 items. Second, a parent interview form is included in the GARS-2 that consists of 25 yes/no items that allow for reviewing a child’s behavior during the first 3 years of life. Third, the total summary score was changed from the Autism Quotient, with a clinical cutoff score of ≥90, to the Autism Index, with a cutoff of ≥85. Finally, guidelines are included in the GARS-2 manual that aid in interpreting the subscale scores and the Autism Index.

The internal consistency for the GARS-2 was identified as high for all subscales and summary scores (Cronbach’s alpha = .84 – .94), based on the original standardization sample (Gilliam, 2006). Additionally, based on the results of 37 raters who completed the instrument twice, one week apart, the test-retest reliability of the GARS-2 subscales and summary scores were identified as moderately high (correlation = .70 – .90; Gilliam, 2006). Interrater reliability estimates are not available for the GARS-2.

**Behavior Rating Inventory of Executive Functioning – Preschool Version**

The Behavior Rating Inventory of Executive Functioning – Preschool Version (BRIEF-P; Gioia et al., 2003) is a parent- and teacher-completed checklist used to evaluate the executive functioning of children aged 2 years, 0 months through 5 years, 11 months. It consists of 63 items, on which behavioral characteristics are rated on a 3-point Likert scale (0 = never observed, 1 = sometimes observed, 2 = often observed). Executive functioning is measured with scores
from five clinical scales (shift, inhibit, emotional control, working memory, and plan/organize) that form three indices [Inhibitory Self-Control (ISCI) = inhibit + emotional control; Flexibility (FI) = shift + emotional control; Emergent Metacognition (EMI) = working memory + plan/organize], and an overall Global Executive Composite (GEC; Gioia et al., 2003). Scores on each are expressed as T-scores, with scores 65 or higher considered clinically significant.

Reported reliability estimates for the BRIEF-P are varied. The internal consistency of the broad indices and composite score of the BRIEF-P was deemed high after completion by both parents (Cronbach’s alpha = .89 – .95) and teachers (Cronbach’s alpha = .93 – .97) with a sample of 302 children (Gioia et al., 2003). Additionally, in a sample of 52 parents and 67 teachers completing the test twice over an approximate period of 4.3 weeks, the test-retest reliability of the broad indices and composite score were identified as high, ranging from .87 – .90 for parents and .77 – .92 for teachers (Gioia et al., 2003). Interrater reliability correlations between parent and teacher ratings for 302 children were low for indexes and for the composite score (correlations = .11 – .26). The distribution of correlations within this range was wide, however, such that the ISCI and FI correlations ranged from .24 – .26. In contrast, the EMI and GEC correlations ranged from .11 – .17 (Gioia et al., 2003).

Procedure

Four “instrument models” were identified for the present study. These models and their item content breakdowns are presented in Table 5. They include a Baseline Model and the following three theoretical models: a Weighted Social Interaction Model, an Executive Functioning Inclusion Model, and a Blended Model. Consistent with the hypotheses of the study, the instrument models differ from one another with respect to the weighting of the diagnostic domains specified by the DSM-IV-TR and the ADOS. These domains include communication,
social interaction, RRSBIA, and non-diagnostic, associated features (e.g., executive functioning). The method for constructing each of the models was as follows:

The *Baseline Model* is the existing configuration of the GARS-2, and includes 42 items split evenly between the three diagnostic domains listed within the *DSM-IV-TR* diagnostic criteria. Because the *Baseline Model* served as the existing GARS-2, which includes no items used to assess non-diagnostic, associated features, each domain had a 33% weight, and the model required no alterations.

The *Weighted Social Interaction Model* is the first of the three theoretical models described, and allowed for an evaluation of the first hypothesis, which predicted that increasing the weight of social interaction items would improve diagnostic validity. The model was created by increasing the weight of the social interaction domain from 33% to 50% (thereby decreasing the weighting of the remaining domains to 25%). Generating this model entailed, first, transforming the three subscale scores (that correspond to the three diagnostic domains) of the GARS-2 into *z*-scores. *Z*-scores allowed variables from different normal distributions to be placed onto the same standard scale. Second, the social interaction *z*-score for each participant was multiplied by 1.5. In that way, the weight of the social interaction domain was increased to represent a 50% share of the total summary score (up from 33%), and thereby reduced the weighted share of the communication and RRSBIA domains (to 25%, each). Third, a new summary score was calculated and then transformed using a reconfiguration of the *z*-score equation \[ X = (z \ast SD) + \mu \]. This final step was conducted so that that the sensitivity and specificity of the model could be established using the original GARS-2 cutoff score of 85, as described in the “Diagnostic Validity of the *Baseline* and Theoretical Models” section below.
The Executive Functioning Inclusion Model is the second of the three theoretical models and allowed for testing the second hypothesis, which predicted that a model that includes executive functioning items will outperform a similar model that does not include such items (i.e., the Baseline Model). This model was generated by creating a non-diagnostic, associated features domain that was tailored towards assessing executive functioning, in addition to the existing domains assessed by the GARS-2. The process of adding this domain was accomplished utilizing items from the BRIEF-P. The model was created using the following steps. First, \( z \)-scores were calculated for the BRIEF-P FI and the three domains of the GARS-2 (communication, social interaction and RRSBIA). Second, each of these four resulting \( z \)-scores was summed to form a summed derivative \( z \)-score for which each of the four scales (the BRIEF-P FI and the three domains of the GARS-2) contributed equally to the final score. Third, the summed derivative \( z \)-score was converted back into an instrument score using a reconfiguration of the \( z \)-score equation \( X = (z \ast SD) + \mu \). The purpose of this final step was to return to a metric for which a cutoff score could be established to calculate the sensitivity and specificity of the instrument, as described in the “Diagnostic Validity of the Baseline and Theoretical Models” section below.

The Blended Model is the third and final theoretical model, created to test the third hypothesis, which predicted that improved validity would result from an instrument that (a) more heavily weighted the social interaction domain (in a manner similar to the first theoretical model) and (b) included an executive functioning domain (in a manner similar to the second theoretical model). The weighting of the four domains was designed to align specifically with the criterion measure used in many studies that examine diagnostic validity—the ADOS (with the following

---

2 In addition to examining the theoretically derived model described here, we also evaluated models that were comprised of the other BRIEF-P indices, as an exploratory exercise. The results of this analysis are presented and described in Appendix A. As expected, the model based on the FI items was equal to or superior to models generated using items from other executive function domains.
weightings for each of the domains: communication, 30%; social interaction, 40%; RRSBIA, 15%, and non-diagnostic, associated features, 15%). In that way, the Blended Model allowed for assessing the hypothesis that aligning the item content characteristics of a screening instrument with the item content characteristics of a criterion measure is the best way to maximize screening instrument validity.

Creating this model involved the following steps. First, similar to the method used to create the Executive Functioning Inclusion Model, z-scores for the BRIEF-P FI and for the three domains of the GARS-2 (communication, social interaction and RRSBIA) were calculated. Second, the diagnostic domains of the GARS-2 (communication, social interaction, and RRSBIA) were reweighted to 30%, 40%, and 15% respectively, and items that assess executive functioning were added from the BRIEF-P FI, and weighted to 15% of the total. Third, these four z-scores were summed. Finally, the summed derivative z-score was converted back into an instrument score using the z-score equation reconfiguration [X = (z * SD) + µ]. This last step was necessary in order that a cutoff score could be established to identify the diagnostic validity of the instrument, described in the “Diagnostic Validity of the Baseline and Theoretical Models” section below.
CHAPTER 4 – RESULTS

Participant Characteristics

Distribution characteristics of the Baseline and theoretical models are presented in Table 6. Additionally, preliminary analyses were performed in order to assess the comparability of the groups (ASD and non-spectrum) with regard to gender and age. Chi-square analyses revealed no significant differences in gender, $X^2 = 1.88, p = .17$. A between-samples t-test revealed a significant difference in age across the groups, $t(134) = 3.31, p = .001$, such that children in the autism group ($M = 4$ years, 2 months, $SD = 1.16$) were significantly younger than children in the non-spectrum group ($M = 4$ years, 10 months, $SD = 1.34$).

Diagnostic Validity of the Baseline and Theoretical Models

To test the hypotheses of the present study, ROC analyses were used to generate estimates of diagnostic validity (sensitivity, specificity, PPV, NPV, and AUC) for each of the four models. Interpretation of these results will be conducted in two steps. In the first step, differences in sensitivity and specificity distributions utilizing a single cutoff score for each model will be evaluated for statistical significance. The procedure for investigating these differences will consist of, first, running McNemar chi-square analyses to determine whether sensitivity and specificity distributions between models are significantly different, and, second, conducting a more stringent method outlined in Braitman (1991) in which confidence intervals are calculated between sensitivity and specificity results of models to determine statistical significance. The second step in interpreting the results will involve investigating the diagnostic validity of each model across all cutoff scores (AUC), discussed below in the “ROC Results” section.
Sensitivity and Specificity Results

Bivariate Results. The sensitivity, specificity, PPV, and NPV, and their respective 95% confidence intervals, of each model utilizing a single cutoff score are presented in Panel A of Table 7 and are graphed in Figure 2. With regard to determining confidence intervals, values were calculated using a binomial proportion confidence interval method \[p \pm 1.96 \times \sqrt{p(1-p)/n}\], as outlined by Hayes (2013). Differences between sensitivity and specificity distributions between the models were evaluated and are presented in Tables 8 and 9, such that sensitivity distributions are listed in Panels A and specificity distributions are listed in Panels B. I refer to these as bivariate results because the methodology from which they derive allows for comparisons (and variability) across two variables: sensitivity and specificity.

The first hypothesis predicted that a model that more heavily weights social interaction items would have significantly better diagnostic validity than the Baseline Model. Consistent with predictions, the sensitivity distributions between the Baseline and Weighted Social Interaction Models were significantly different \((X^2 = 15.00, p = .001)\). However, the specificity distributions between the Baseline and Weighted Social Interaction Models did not significantly differ \((X^2 = 5.00, p = .063)\).

The second and third hypotheses were evaluated using a similar approach to those of the first hypothesis, although these methods differed with respect to the establishment of cutoff scores for the relevant theoretical models. This is necessitated by the addition of items from the BRIEF-P, which makes a cutoff of 85 non-equivalent across the models. Therefore, based on the ROC analysis conducted for each model, cutoff scores were identified for a pegged specificity value of .70 (in keeping with the recommended .70 – .80 “acceptable” range for sensitivity and specificity; Coonrod & Stone, 2005; Glascoe, 2005; Norris & Lecavalier, 2010). For the
Executive Functioning Inclusion Model, this sensitivity value was obtained with a cutoff score of 72. As noted below, a specificity equal to .70 constitutes a high bar for improving sensitivity given that (1) sensitivity and specificity oftentimes vary inversely with one another, and that (2) the specificity of the Baseline Model (cutoff score = 85) was only .53. Results did not support the second hypothesis, as the sensitivity distributions between the Baseline and Executive Functioning Models (cutoff score = 72) were not significantly different ($X^2 = 0.067, p = 1.00$). However, despite the fact that the sensitivity did not improve, the specificity distributions found for the Baseline and Executive Functioning Inclusion Models were significantly different ($X^2 = 4.50, p = .047$). Therefore, despite showing no sensitivity improvement, this model appears to be superior to the Baseline Model, owing to the improved specificity.

The third hypothesis predicted that a model for which an item-content distribution mirrored that of the ADOS would achieve improved levels of diagnostic validity. Consistent with predictions, significant differences were found between the sensitivity ($X^2 = 3.56, p = .046$) and specificity ($X^2 = 5.44, p = .039$) distributions of the Baseline and Blended (cutoff score = 40) Models. In fact, the Blended Model qualified as an acceptable screening instrument based on stringent criteria of obtaining both sensitivity and specificity values greater than or equal to .70.

Univariate Results. A potential criticism of the method just described concerns the fact that the methods of identifying cutoff scores varied between models. That is, the Baseline and Weighted Social Interaction Models were evaluated on the basis of assessing bivariate estimates of performance (sensitivity and specificity), whereas the Executive Functioning Inclusion and Blended Models were evaluated utilizing a univariate approach (sensitivity, when the specificities were set to .70). To improve comparability across models, the method whereby specificity is pegged to .70 was applied to all four models. The sensitivity, specificity, PPV, NPV,
and their respective confidence intervals, for each of the models utilizing this univariate method appear in Panel B of Table 7. Please note that the diagnostic validity estimates that appear in Panels A and B for the Executive Functioning Inclusion and Blended Models are identical, owing to the fact that, for these models, the bivariate analysis method also involved pegging specificity at .70.

Differences between sensitivity and specificity distributions between the models identified using the univariate approach were evaluated and are presented in Panels C of Tables 8 and 9, and are graphed in Figure 3. Results did not support the first hypothesis, which predicted that more heavily weighting social interaction items would result in improved diagnostic validity, as the sensitivity distributions of the Baseline (cutoff score = 90) and Weighted Social Interaction (cutoff score = 92) Models did not significantly differ ($\chi^2 = 2.27, p = .227$). The results of this univariate method also did not support the second hypothesis, which predicted that including items that assess executive functioning deficits would improve diagnostic validity, as the sensitivity distributions of the Baseline and Executive Functioning Inclusion (cutoff score = 72) Models were not significantly different ($\chi^2 = 0.33, p = .774$). Finally, results did support the third hypothesis, which predicted that altering the item-content distribution of a model to mirror the item distribution found for the ADOS would improve diagnostic validity, as a significant difference was found between the sensitivity distributions of the Baseline and Blended (cutoff score = 40) Models ($\chi^2 = 8.90, p = .004$).

**ROC Results**

As described above, the second step in interpreting the results of the ROC analyses involved evaluating the diagnostic validity of the Baseline and theoretical models across all cutoff scores (AUC). These scores are depicted in Table 7 (Panel A) and graphed in Figure 4.
Results of the ROC analyses revealed that the AUC values of all four models fell within the “poor” range.
CHAPTER 5 – DISCUSSION

The purpose of this study was to evaluate whether systematic manipulations of domain weightings, as they relate to the diagnostic domains depicted in *DSM-IV-TR* and the ADOS, would result in improvements to the diagnostic validity of a poorly performing instrument developed to screen children for ASDs—the GARS-2. Previous research reveals the GARS-2 to have poor sensitivity and specificity (for a review and meta-analysis see Hampton & Strand, 2015). Therefore, the item content of that instrument was altered to (1) mirror the ASD-diagnosis structure identified in *DSM-IV-TR* and item content distribution of “gold standard” instruments, and (2) to include items that assess deficits in executive functioning.

Three hypotheses were tested. First, it was hypothesized that increasing the weighting of items that assess social interaction difficulties would significantly improve the sensitivity of that model relative to the original instrument, without a substantial decrease in specificity. Second, it was hypothesized that including items that assess executive functioning impairments would, likewise, result in significant improvements in sensitivity without a substantial decrease in specificity. Third, it was hypothesized that a model that modified the *Baseline Model* such that both of the above mentioned modifications were made would not only result in improvements to diagnostic validity, but that this instrument would be the best performer of the four models.

The results of the present study supported two of the three study hypotheses. The first hypothesis was supported, such that when the weight of the social interaction domain of the GARS-2 was increased, the sensitivity was significantly improved relative to the sensitivity of the GARS-2, with a minimal decrease in specificity. These findings are in keeping with concerns about the GARS reported by Lecavalier (2005), that the poor sensitivity of the GARS reflects an overemphasis on evaluating RRSBIA, and an under-emphasis on evaluating social interaction
and communication deficits. Therefore, a viable method of improving the GARS-2 may lie in increasing the number of items dedicated to screening social interaction deficits, or scoring that instrument so as to more heavily weight items that assess social interaction deficits.

With respect to the second hypothesis, the results of the present study reveal that when specificity was held constant at .70, the sensitivity of the Executive Functioning Inclusion Model was not significantly higher than the sensitivity of the original Baseline Model; despite this, the “pegged” specificity of .70 for the Executive Functioning Inclusion Model was significantly higher than the specificity found for the Baseline Model. Therefore, although in a technical sense the second hypothesis was not supported, it is the case that the Executive Functioning Inclusion Model is an improvement over the currently available GARS-2 instrument.

This finding is in keeping with previous literature that theorized that a primary cause of autistic behaviors are deficits in executive functioning (Ozonoff et al., 1991; Pennington et al., 1997; Russell, 1997). Many of these behaviors are driven by difficulties in forming a flexible plan of action, focusing and sustaining attention, and inhibiting impulsive responses (Liss et al., 2001). These “autistic behaviors”, then, are related to general disruptions in functioning, and are likely to be noticeable to parents. Therefore, including questions about these elements in parent-report instruments will likely improve disorder identification. Future research should explore this area more thoroughly, however, especially in regard to optimally balancing the executive functioning domain with the ASD diagnostic domains.

The current results supported the third hypothesis that reweighting the item content distribution of the Blended Model to mirror that of the ADOS, and including items used to assess executive functioning deficits, resulted in significantly improved sensitivity, relative to the sensitivity found for the GARS-2. In fact, when the specificity was held at .70, the sensitivity
improved to the .70 – .80 “acceptable” range. These findings are in keeping with literature cited previously suggesting that (a) the GARS could be improved by amplifying the weight given to social interaction items (Lecavalier, 2005), and that (b) individuals with ASDs often demonstrate difficulties with cognitive flexibility (Griffith et al., 1999; McEvoy et al., 1993; Ozonoff et al., 1994; Schneider & Asarnow, 1987), planning (Ozonoff et al., 1991; Ozonoff & McEvoy, 1994; Prior & Hoffman, 1990) and self-regulation (Gomez & Baird, 2005). Therefore, including items that assess these areas are likely to improve identification. The fact that this model was the best of the four suggests that each of the two modifications (increasing the weight of the social interaction domain and adding items that assess executive functioning) contribute uniquely to improving the validity of the GARS-2.

**ROC Interpretation**

Despite the success of achieving “acceptable” levels of sensitivity and specificity using established cutoffs, the evaluated models are limited by having poor AUC levels. These low AUC values indicate that the average accuracy across most cutoff values for each instrument was low. This can be seen in Figure 3, as most of the ROC points are close to the reference line, with high “peaks” at the cutoffs utilized in the current study. Thus, the theoretical models demonstrate improved diagnostic validity using optimal cutoff values, but they are limited to using only those fixed cutoff values.

**Limitations**

The present study had several limitations. First, it included a moderately low sample size of 136 participants. Although larger than the sample size of several other studies exploring diagnostic validity, an even larger sample would provide greater confidence in the results. A second limitation of the present study concerned the fact that the ADOS group sizes were not
equal. Unequal group sizes are common in studies using similar methodologies (Sikora et al., 2008; South et al., 2002), but this can be problematic if the difference between groups is large enough to allow confounding variables to effect the results. A third limitation of the present study concerns the fact that there were age differences across the groups, with younger children somewhat overrepresented in the ASDs group. A final limitation of the current study concerns the inclusion of executive functioning items, only, within the “non-diagnostic, associated features” domain. This is a limitation because there are areas of functioning other than executive functioning that are associated with ASDs, such as fine and gross motor impairments (Lloyd, MacDonald, & Lord, 2013). As such, it must be left to subsequent research to determine the impact of such items on the diagnostic validity of ASDs screening instruments.

Clinical Recommendations

An important clinical implication of the present study concerns improvements that can be made to the diagnostic validity of an existing ASD screening instrument. The results reveal that an alternative method for scoring the GARS-2, that involves re-weighting the value of items, yields a distribution of scores that is more accurate with respect to diagnostic sensitivity. Such improvements may be achieved by simply multiplying by 1.5 the social interaction domain items, prior to summing all item scores. The resulting score, when a cut-off score of 85 is applied, yields greater accuracy with respect to distinguishing between children who do and do not meet criteria for ASDs, as identified by a commonly used criterion instrument, the ADOS. It should be noted, however, that despite being an improvement over the original scoring protocol, this alternative method is still considered to be deficient in terms of meeting criteria for adequate screening instruments (Coonrod & Stone, 2005; Glascoe, 2005; Norris & Lecavalier, 2010).
Generating an instrument that meets minimal criteria for screening instruments, according to the results of the present study, requires not simply increasing the weighting of social interaction domain items, but also the inclusion of items that assess non-diagnostic, associated features of ASDs. This can be accomplished for cases for which both the GARS-2 and the BRIEF-P have been administered, by generating scores that reflect the item-content weightings of the Blended Model. The steps are as follows: First, the domain scores of the GARS-2 and BRIEF-P FI score should be converted into $z$-scores, in order that all scores are placed onto the same standard scale. $Z$-scores can be calculated using GARS-2 and BRIEF-P normative samples, as provided in their respective clinical manuals, to determine the population mean ($\mu$) and standard deviation (SD) for a given age group. Using the $z$-score equation $[z = (x - \mu)/SD]$, scores can be calculated by subtracting the GARS-2 and BRIEF-P population means from each of the instrument scores, and dividing the resultant scores by the standard deviations of each population mean. Second, each of the newly calculated GARS-2 domain $z$-scores (communication, social interaction, RRB) should be weighted by 30%, 40% and 15%, respectively, and the BRIEF-P FI $z$-score should be weighted by 15%. This domain weighting process ensures that each of the domains will have appropriate representation in the total summary score, in keeping with the item distribution found for the ADOS. Third, $z$-scores should be summed, and transformed back into an instrument score using a reconfiguration of the $z$-score equation $[X = (z \times SD) + \mu]$. Fourth, based on a cutoff score of 40, the child is identified as likely (score $\geq 40$) or unlikely (score < 40) to have an ASD.

**Conclusions**

The current study evaluated whether manipulating the item content of an ASD screening instrument would improve the diagnostic validity of that instrument. Specifically, the results
supported constructing the item content of screening instruments in a way that (1) mirrors the item content distribution of “gold standard” instruments and diagnostic structure of ASD identified in *DSM-IV-TR*, and (2) includes items that assess executive functioning deficits. These findings have both clinical and theoretical implications for ASD screening. Clinically, having identified more optimal item content and item weightings for an ASD screening instrument, procedures have been provided for generating scores that allow for more accurate screening. Theoretically, a method has been provided for devising more effective screening instruments, such that the item content of instruments should be altered to align with the item content of “gold standard” criterion instruments. Utilizing these procedures, the current results predict improvements in accurately identifying children with ASDs by current, and future, screening instruments.
References


Robins, D., Fein, D., Barton, M., & Green, J. (2001). The Modified Checklist for Autism in


## Table 1

<table>
<thead>
<tr>
<th>Checklist</th>
<th>Source</th>
<th>Age</th>
<th>Sample Group</th>
<th>Sens</th>
<th>Spec</th>
<th>BR</th>
<th>PPV</th>
<th>NPV</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>SRS</strong></td>
<td>Aldridge et al. (2011)</td>
<td>4-15 years</td>
<td>ASD (35) vs. Non-ASD (13)</td>
<td>.91</td>
<td>.08</td>
<td>.73</td>
<td>.71</td>
<td>.25</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>Schanding et al. (2011)</td>
<td>3-17 years</td>
<td>ASD (1,659) vs. Non-ASD (1,546)</td>
<td>.95</td>
<td>.96</td>
<td>.52</td>
<td>.96</td>
<td>.96</td>
<td>NR</td>
</tr>
<tr>
<td>*Constantino et al. (2007)</td>
<td>4-18 years</td>
<td>ASD (271) vs. Non-ASD (171)</td>
<td>.75</td>
<td>.96</td>
<td>.61</td>
<td>.97</td>
<td>.71</td>
<td>.95</td>
<td></td>
</tr>
<tr>
<td><strong>CARS</strong></td>
<td>Chlebowsky et al. (2010)</td>
<td>2 years</td>
<td>Autism (142) vs. PDD-NOS (101)</td>
<td>.93</td>
<td>.49</td>
<td>.58</td>
<td>.72</td>
<td>.83</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>Ventola et al. (2006)</td>
<td>16-30 mo</td>
<td>ASD (27) vs. Non-ASD (18)</td>
<td>.96</td>
<td>.67</td>
<td>.60</td>
<td>.81</td>
<td>.69</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>Perry et al. (2005)</td>
<td>2-6 years</td>
<td>Autism (104) vs. PDD-NOS (44)</td>
<td>.86</td>
<td>.80</td>
<td>.70</td>
<td>.91</td>
<td>.71</td>
<td>NR</td>
</tr>
<tr>
<td></td>
<td>L. C. Eaves &amp; Milner (1993)</td>
<td>2-21 years</td>
<td>ASD (48) vs. Non-ASD (29)</td>
<td>.98</td>
<td>NR</td>
<td>.62</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
</tr>
<tr>
<td><strong>CBCL Wd</strong></td>
<td>Narzisi et al. (2013)</td>
<td>1.5-3 years</td>
<td>ASD (47) vs. Non-ASD (47)</td>
<td>.92</td>
<td>.97</td>
<td>.50</td>
<td>.98</td>
<td>.92</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>Muratori et al. (2011)</td>
<td>2-5 years</td>
<td>ASD (101) vs. Non-ASD (117)</td>
<td>.89</td>
<td>.92</td>
<td>.46</td>
<td>.90</td>
<td>.90</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>Sikora et al. (2008)</td>
<td>16-71 mo</td>
<td>Autism (79) vs. Non-ASD (50)</td>
<td>.65</td>
<td>.62</td>
<td>.73</td>
<td>.82</td>
<td>.40</td>
<td>NR</td>
</tr>
<tr>
<td><strong>CBCL PDP</strong></td>
<td>Narzisi et al. (2013)</td>
<td>1.5-3 years</td>
<td>ASD (47) vs. Non-ASD (47)</td>
<td>.98</td>
<td>.91</td>
<td>.50</td>
<td>.92</td>
<td>.98</td>
<td>.99</td>
</tr>
<tr>
<td></td>
<td>Predescu et al. (2013)</td>
<td>1.5-5 years</td>
<td>ASD (103) vs. Non-ASD (70)</td>
<td>.96</td>
<td>.89</td>
<td>.60</td>
<td>.93</td>
<td>.94</td>
<td>.97</td>
</tr>
<tr>
<td></td>
<td>Muratori et al. (2011)</td>
<td>2-5 years</td>
<td>ASD (101) vs. Non-ASD (117)</td>
<td>.85</td>
<td>.90</td>
<td>.46</td>
<td>.88</td>
<td>.87</td>
<td>.95</td>
</tr>
<tr>
<td></td>
<td>Sikora et al. (2008)</td>
<td>16-71 mo</td>
<td>Autism (79) vs. Non-ASD (50)</td>
<td>.78</td>
<td>.42</td>
<td>.73</td>
<td>.78</td>
<td>.41</td>
<td>NR</td>
</tr>
<tr>
<td><strong>SCQ</strong></td>
<td>Oosterling et al. (2010)</td>
<td>20-40 mo</td>
<td>ASD (143) vs. Non-ASD (65)</td>
<td>.63</td>
<td>.94</td>
<td>.69</td>
<td>.96</td>
<td>.54</td>
<td>.64</td>
</tr>
<tr>
<td>*Snow &amp; Lecavalier (2008)</td>
<td>18-70 mo</td>
<td>PDD (40) vs. Non-PDD (25)</td>
<td>.70</td>
<td>.52</td>
<td>.62</td>
<td>.70</td>
<td>.52</td>
<td>.63</td>
<td></td>
</tr>
<tr>
<td>*Allen et al. (2007)</td>
<td>24-84 mo</td>
<td>ASD (28) vs. Non-ASD (53)</td>
<td>.60</td>
<td>.70</td>
<td>.35</td>
<td>.52</td>
<td>.76</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>*Corsello et al. (2007)</td>
<td>2-16 yrs</td>
<td>ASD (438) vs. Non-ASD (151)</td>
<td>.71</td>
<td>.71</td>
<td>.74</td>
<td>.88</td>
<td>.45</td>
<td>.77</td>
<td></td>
</tr>
<tr>
<td>*Lee et al. (2007)</td>
<td>3-5 yrs</td>
<td>ASD (54) vs. Non-ASD (214)</td>
<td>.54</td>
<td>.92</td>
<td>.20</td>
<td>.62</td>
<td>.89</td>
<td>.88</td>
<td></td>
</tr>
<tr>
<td>*Wiggins et al. (2007)</td>
<td>17-45 mo</td>
<td>ASD (19) vs. DD (18)</td>
<td>.47</td>
<td>.89</td>
<td>.51</td>
<td>.82</td>
<td>.62</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>*Witwer &amp; Lecavalier (2007)</td>
<td>4-14 yrs</td>
<td>ASD (36) vs. Non-ASD (13)</td>
<td>.92</td>
<td>.62</td>
<td>.74</td>
<td>.87</td>
<td>.73</td>
<td>.89</td>
<td></td>
</tr>
<tr>
<td>*L. C. Eaves et al. (2006a)</td>
<td>4-6 yrs</td>
<td>ASD (35) vs. Non-ASD (59)</td>
<td>.74</td>
<td>.54</td>
<td>.54</td>
<td>.65</td>
<td>.78</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>Study</td>
<td>Age Range</td>
<td>Group Comparison</td>
<td>Sensitivity</td>
<td>Specificity</td>
<td>PPV</td>
<td>NPV</td>
<td>AUC</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>-------------------------------</td>
<td>------------</td>
<td>-------------------------------------------------------</td>
<td>-------------</td>
<td>-------------</td>
<td>--------</td>
<td>--------</td>
<td>--------</td>
<td>------</td>
<td></td>
</tr>
<tr>
<td>*L. C. Eaves et al. (2006b)</td>
<td>36-82 mo</td>
<td>ASD (49) vs. Non-ASD (102)</td>
<td>.71</td>
<td>.79</td>
<td>.35</td>
<td>.65</td>
<td>.84</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>Sikora et al. (2008)</td>
<td>16-71 mo</td>
<td>Autism (79) vs. Non-ASD (50)</td>
<td>.53</td>
<td>.54</td>
<td>.73</td>
<td>.76</td>
<td>.29</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>*R. C. Eaves et al. (2006)</td>
<td>3-26 yrs</td>
<td>ASD (111) vs. Non-ASD (23)</td>
<td>.79</td>
<td>.68</td>
<td>.83</td>
<td>.92</td>
<td>.40</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>Mazefsky &amp; Oswald (2006)</td>
<td>22 mo-8 yrs</td>
<td>ASD (59) vs. Non-ASD (19)</td>
<td>.39</td>
<td>.61</td>
<td>.76</td>
<td>.76</td>
<td>.24</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>*Lecavalier (2005)</td>
<td>3-21 yrs</td>
<td>ASD (284)</td>
<td>.38</td>
<td>NR</td>
<td>1.00</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td></td>
</tr>
<tr>
<td>*South et al. (2002)</td>
<td>3-10.5 yrs</td>
<td>ASD (119)</td>
<td>.48</td>
<td>NR</td>
<td>1.00</td>
<td>NR</td>
<td>NR</td>
<td>NR</td>
<td></td>
</tr>
</tbody>
</table>

*Note. SRS = Social Responsiveness Scale; CBCL = Child Behavior Checklist; Wd = Withdrawn; PDP = Pervasive Developmental Problems; CARS = Childhood Autism Rating Scale; SCQ = Social Communication Questionnaire; GARS = Gilliam Autism Rating Scale; Sens = Sensitivity; Spec = Specificity; BR = Base Rate; PPV = Positive Predictive Value; NPV = Negative Predictive Value; AUC = Area Under the Curve; NR = Not Reported; * = Study included in Norris & Lecavalier (2010)
Table 2

Average sensitivity and specificity scores for each screening instrument

<table>
<thead>
<tr>
<th>Checklist</th>
<th>Sensitivity</th>
<th>CI</th>
<th>Specificity</th>
<th>CI</th>
<th># of studies</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Instruments illustrating good validity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SRS</td>
<td>.92</td>
<td>.91–.93</td>
<td>.95</td>
<td>.94–.96</td>
<td>3</td>
</tr>
<tr>
<td>CARS</td>
<td>.92</td>
<td>.89–.95</td>
<td>.92</td>
<td>.89–.95</td>
<td>4</td>
</tr>
<tr>
<td>CBCL Wd</td>
<td>.81</td>
<td>.76–.86</td>
<td>.87</td>
<td>.82–.92</td>
<td>3</td>
</tr>
<tr>
<td>CBCL PDP</td>
<td>.89</td>
<td>.86–.92</td>
<td>.81</td>
<td>.76–.86</td>
<td>4</td>
</tr>
<tr>
<td><strong>Instruments illustrating poor validity</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCQ</td>
<td>.69</td>
<td>.66–.72</td>
<td>.76</td>
<td>.73–.79</td>
<td>9</td>
</tr>
<tr>
<td>GARS</td>
<td>.49</td>
<td>.45–.53</td>
<td>.60</td>
<td>.50–.70</td>
<td>5</td>
</tr>
</tbody>
</table>

*Note.* CI = Confidence Interval; SRS = Social Responsiveness Scale; CARS = Childhood Autism Rating Scale; CBCL = Child Behavior Checklist; Wd = Withdrawn; PDP = Pervasive Developmental Problems; SCQ = Social Communication Questionnaire; GARS = Gilliam Autism Rating Scale
Table 3

A breakdown of items on screening instruments according to DSM-IV-TR diagnostic domains

<table>
<thead>
<tr>
<th>Instruments illustrating good validity</th>
<th>Communication</th>
<th>Social Interaction</th>
<th>RRSBIA</th>
<th>Associated Features</th>
</tr>
</thead>
<tbody>
<tr>
<td>SRS</td>
<td>22 (34%)</td>
<td>31 (48%)</td>
<td>12 (18%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>CARS</td>
<td>2 (14%)</td>
<td>4 (29%)</td>
<td>3 (21%)</td>
<td>5 (36%)</td>
</tr>
<tr>
<td>CBCL Wd</td>
<td>0 (0%)</td>
<td>16 (100%)</td>
<td>0 (0%)</td>
<td>0 (0%)</td>
</tr>
<tr>
<td>CBCL PDP</td>
<td>1 (8%)</td>
<td>6 (46%)</td>
<td>6 (46%)</td>
<td>0 (0%)</td>
</tr>
</tbody>
</table>

| Instruments illustrating poor validity | | | | |
|---------------------------------------| | | | |
| SCQ                                   | 13 (33%)      | 15 (38%)           | 8 (21%) | 3 (8%)              |
| GARS                                  | 14 (33%)      | 14 (33%)           | 14 (33%) | 0 (0%)              |

| ADOS | | | |
|------| | | |
| Module 1 | 8 (28%) | 12 (41%) | 4 (14%) | 5 (17%) |
| Module 2 | 8 (30%) | 11 (40%) | 4 (15%) | 4 (15%) |

| DSM-IV-TR | 4 | 4 | 4 | 1 |

**Note.** DSM-IV-TR = Diagnostic and Statistical Manual of Mental Disorders – Revised 4th edition; RRSBIA = Restricted, Repetitive Behaviors, Interest and Activities; SRS = Social Responsiveness Scale; CBCL = Child Behavior Checklist; Wd = Withdrawn; PDP = Pervasive Developmental Problems; CARS = Childhood Autism Rating Scale; SCQ = Social Communication Questionnaire; GARS = Gilliam Autism Rating Scale; ADOS = Autism Diagnostic Observation Schedule
Table 4

Participant characteristics

<table>
<thead>
<tr>
<th>Gender</th>
<th>Non-ASD</th>
<th>ASD</th>
<th>Total</th>
<th>Non-ASD</th>
<th>ASD</th>
<th>Total</th>
<th>ADHD</th>
<th>DBD</th>
<th>Language</th>
<th>Anxiety</th>
</tr>
</thead>
<tbody>
<tr>
<td>Males</td>
<td>43</td>
<td>67</td>
<td>110 (81%)</td>
<td>11</td>
<td>8</td>
<td>19 (17%)</td>
<td>14</td>
<td>10</td>
<td>14 (25%)</td>
<td>19 (33%)</td>
</tr>
<tr>
<td>Females</td>
<td>14</td>
<td>12</td>
<td>26 (19%)</td>
<td>3</td>
<td>2</td>
<td>5 (4%)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>57 (42%)</td>
<td>79 (58%)</td>
<td>136</td>
<td>14 (25%)</td>
<td>10  (17%)</td>
<td>14 (25%)</td>
<td>19 (33%)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Age</th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean</td>
<td>4 yrs, 10 mo</td>
<td>4 yrs, 2 mo</td>
<td>4 yrs, 6 mo</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SD</td>
<td>1.34</td>
<td>1.16</td>
<td>1.28</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

| Comorbid Diagnoses |         |      |            |         |     |            |      |     |          |         |
| ADHD          | 8       | 4    | 12 (15%)   |         |     |            |      |     |          |         |
| DBD           | 7       | 9    | 16 (20%)   |         |     |            |      |     |          |         |
| Language      | 17      | 1    | 18 (23%)   |         |     |            |      |     |          |         |
| Anxiety       | 5       | 5    | 10 (13%)   |         |     |            |      |     |          |         |
| ID            | 12      | 4    | 16 (20%)   |         |     |            |      |     |          |         |
| DCD           | 2       | 2    | 4 (5%)     |         |     |            |      |     |          |         |
| Other         | 1       | 2    | 3 (4%)     |         |     |            |      |     |          |         |
| Total         | 52 (66%)| 27 (34%)| 79          |         |     |            |      |     |          |         |

*Note. ASD = Autism Spectrum Disorder; ADHD = Attention Deficit/Hyperactivity Disorder; DBD = Disruptive Behavior Disorder; Language = Mixed Receptive-Expressive Language Disorder; Anxiety = Anxiety Disorder Not Otherwise Specified; SD = Standard Deviation; ID = Intellectual Disability; DCD = Developmental Coordination Disorder; Other = Seizure and Adjustment Disorders*
Table 5

A breakdown of item weights per domain for models

<table>
<thead>
<tr>
<th>Model</th>
<th>Communication</th>
<th>Social Interaction</th>
<th>RRSBIA</th>
<th>Executive Functioning</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>33%</td>
<td>33%</td>
<td>33%</td>
<td>0%</td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>25%</td>
<td>50%</td>
<td>25%</td>
<td>0%</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
<td>25%</td>
</tr>
<tr>
<td>Blended Model</td>
<td>30%</td>
<td>40%</td>
<td>15%</td>
<td>15%</td>
</tr>
</tbody>
</table>

*Note. RRSBIA = Restricted, Repetitive Behaviors, Interest and Activities*
Table 6

Descriptive statistics of the *Baseline* and theoretical models

<table>
<thead>
<tr>
<th>Model</th>
<th>Mean</th>
<th>SD</th>
<th>Range</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline Model</td>
<td>88.27</td>
<td>18.86</td>
<td>57 – 128</td>
<td>.379</td>
<td>-.599</td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>97.07</td>
<td>22.32</td>
<td>61 – 141</td>
<td>.290</td>
<td>-.680</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>67.82</td>
<td>24.23</td>
<td>19 – 126</td>
<td>.123</td>
<td>-.321</td>
</tr>
<tr>
<td>Blended Model</td>
<td>40.71</td>
<td>14.42</td>
<td>12 – 76</td>
<td>.121</td>
<td>-.324</td>
</tr>
</tbody>
</table>

*Note.* SD = Standard Deviation
Table 7

Diagnostic validity of models

<table>
<thead>
<tr>
<th>Model</th>
<th>Cutoff</th>
<th>Sens</th>
<th>CI</th>
<th>Spec</th>
<th>CI</th>
<th>PPV</th>
<th>CI</th>
<th>NPV</th>
<th>CI</th>
<th>AUC</th>
<th>CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A Bivariate diagnostic validity results of the models</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Model</td>
<td>85</td>
<td>.56</td>
<td>.48 – .64</td>
<td>.53</td>
<td>.45 – .61</td>
<td>.62</td>
<td>.54 – .70</td>
<td>.46</td>
<td>.38 – .54</td>
<td>.588</td>
<td>.478 – .697</td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>85</td>
<td>.75</td>
<td>.68 – .82</td>
<td>.47</td>
<td>.39 – .55</td>
<td>.66</td>
<td>.58 – .74</td>
<td>.57</td>
<td>.49 – .65</td>
<td>.592</td>
<td>.483 – .701</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>72</td>
<td>.55</td>
<td>.47 – .63</td>
<td>.70</td>
<td>.62 – .78</td>
<td>.72</td>
<td>.65 – .79</td>
<td>.53</td>
<td>.45 – .61</td>
<td>.686</td>
<td>.585 – .786</td>
</tr>
<tr>
<td>Blended Model</td>
<td>40</td>
<td>.72</td>
<td>.65 – .79</td>
<td>.70</td>
<td>.62 – .78</td>
<td>.77</td>
<td>.70 – .84</td>
<td>.65</td>
<td>.57 – .73</td>
<td>.689</td>
<td>.591 – .787</td>
</tr>
<tr>
<td><strong>Panel B Univariate diagnostic validity results of the models, when the specificity is pegged at .70</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Baseline Model</td>
<td>90</td>
<td>.48</td>
<td>.40 – .56</td>
<td>.70</td>
<td>.45 – .61</td>
<td>.69</td>
<td>.61 – .77</td>
<td>.49</td>
<td>.41 – .57</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>92</td>
<td>.52</td>
<td>.44 – .60</td>
<td>.70</td>
<td>.39 – .55</td>
<td>.71</td>
<td>.63 – .79</td>
<td>.51</td>
<td>.43 – .59</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>72</td>
<td>.55</td>
<td>.47 – .63</td>
<td>.70</td>
<td>.62 – .78</td>
<td>.72</td>
<td>.64 – .80</td>
<td>.53</td>
<td>.45 – .61</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Blended Model</td>
<td>40</td>
<td>.72</td>
<td>.65 – .79</td>
<td>.70</td>
<td>.62 – .78</td>
<td>.77</td>
<td>.70 – .84</td>
<td>.65</td>
<td>.57 – .73</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Note.* Sens = Sensitivity; CI = Confidence Interval; Spec = Specificity; PPV = Positive Predictive Value; NPV = Negative Predictive Value; AUC = Area Under the Curve
Table 8

McNemar chi-square test results

<table>
<thead>
<tr>
<th>Model</th>
<th>Baseline Model</th>
<th>Weighted Social Interaction Model</th>
<th>Executive Functioning Inclusion Model</th>
</tr>
</thead>
<tbody>
<tr>
<td>Panel A Bivariate statistical differences between the sensitivity rate distributions of models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>15.00**</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>0.67</td>
<td>12.25**</td>
<td></td>
</tr>
<tr>
<td>Blended Model</td>
<td>3.56*</td>
<td>22.27</td>
<td>9.00*</td>
</tr>
<tr>
<td>Panel B Bivariate statistical differences between the specificity rate distributions of models</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>5.00</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>4.50*</td>
<td>9.31*</td>
<td></td>
</tr>
<tr>
<td>Blended Model</td>
<td>5.44</td>
<td>10.29 *</td>
<td>1.00</td>
</tr>
<tr>
<td>Panel C Univariate statistical differences between the sensitivity rate distributions of models when specificity rates are pegged at .70</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>2.27</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>0.33</td>
<td>0.40</td>
<td></td>
</tr>
<tr>
<td>Blended Model</td>
<td>8.90*</td>
<td>6.55*</td>
<td>9.31*</td>
</tr>
</tbody>
</table>

Note. * = p < .05; ** = p < .001
Table 9

Statistical significance using 95% confidence intervals of differences between the models

<table>
<thead>
<tr>
<th>% correctly identified</th>
<th>Theoretical Model</th>
<th>Baseline Model</th>
<th>P value</th>
<th>Statistical Significance</th>
<th>Difference in % responding</th>
<th>95% CI</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Panel A</strong> Bivariate statistical differences between the sensitivity rate distributions of models using 95% confidence intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>59/79 = 75%</td>
<td>44/79 = 56%</td>
<td>.000</td>
<td>Yes</td>
<td>19%</td>
<td>4% to 34%</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>36/66 = 55%</td>
<td>37/66 = 56%</td>
<td>1.00</td>
<td>No</td>
<td>1%</td>
<td>-15% to 17%</td>
</tr>
<tr>
<td>Blended Model</td>
<td>48/66 = 72%</td>
<td>37/66 = 56%</td>
<td>.046</td>
<td>Yes</td>
<td>16%</td>
<td>0% to 32%</td>
</tr>
<tr>
<td><strong>Panel B</strong> Bivariate statistical differences between the specificity rate distributions of models using 95% confidence intervals</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>27/57 = 47%</td>
<td>30/57 = 53%</td>
<td>.063</td>
<td>No</td>
<td>6%</td>
<td>-12% to 24%</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>33/47 = 70%</td>
<td>25/47 = 53%</td>
<td>.070</td>
<td>No</td>
<td>17%</td>
<td>-3% to 37%</td>
</tr>
<tr>
<td>Blended Model</td>
<td>33/47 = 70%</td>
<td>25/47 = 53%</td>
<td>.070</td>
<td>No</td>
<td>17%</td>
<td>-3% to 37%</td>
</tr>
<tr>
<td><strong>Panel C</strong> Univariate statistical differences between the sensitivity rate distributions of models using 95% confidence intervals when specificity rates are pegged at .70</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Weighted Social Interaction Model</td>
<td>41/79 = 52%</td>
<td>38/79 = 48%</td>
<td>.227</td>
<td>No</td>
<td>4%</td>
<td>-12% to 20%</td>
</tr>
<tr>
<td>Executive Functioning Inclusion Model</td>
<td>36/66 = 55%</td>
<td>32/66 = 48%</td>
<td>.774</td>
<td>No</td>
<td>7%</td>
<td>-10% to 24%</td>
</tr>
<tr>
<td>Blended Model</td>
<td>48/66 = 72%</td>
<td>32/66 = 48%</td>
<td>.004</td>
<td>Yes</td>
<td>24%</td>
<td>7% to 31%</td>
</tr>
</tbody>
</table>
Figure 1. Summary statistics and confidence intervals of the true (sensitivity) and false (1 – specificity) positive rates of ASD screening instruments
Figure 2. Summary statistics and confidence intervals of the true (sensitivity) and false (1 – specificity) positive rates of the Baseline and theoretical models.
Figure 3. Summary statistics and confidence intervals for the sensitivity of the Baseline and theoretical models when the specificity is .70
Figure 4. Receiver operating curve (ROC) for the Baseline and theoretical models
Appendix A

Summary statistics and confidence intervals for the sensitivity of the Executive Functioning Inclusion and Blended Models for each of the broad indices of the BRIEF-P when the specificity is .70

![Graph showing sensitivity and confidence intervals for the BRIEF-P broad indices]

Note. BRIEF-P = Behavior Rating Inventory of Executive Functioning – Preschool Version; ISCI = Inhibitory Self-Control; FI = Flexibility Index; EMI = Emergent Metacognition; GEC = Global Executive Composite

As expected, none of the additional models that included the BRIEF-P indices performed as well as the theoretically derived model. Therefore, only the results of the Executive Functioning Inclusion and Blended Models that include the Flexibility Index are discussed.