2008

Associations Between Intelligence Test Scores and Test Session Behavior in Children with ADHD, LD, and EBD

Stephanie Anne Nelson
University of Vermont

ASSOCIATIONS BETWEEN INTELLIGENCE TEST SCORES AND TEST SESSION BEHAVIOR IN CHILDREN WITH ADHD, LD, AND EBD

A Dissertation Presented

by

Stephanie Anne Nelson

to

The Faculty of the Graduate College

of

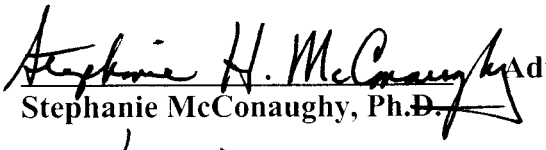
The University of Vermont

In Partial Fulfillment of the Requirements
For the Degree of Doctor of Philosophy
Specializing in Clinical Psychology

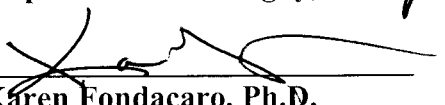
February 2008

Accepted by the Faculty of the Graduate College, The University of Vermont, in partial fulfillment of the requirements for the degree of Doctor of Philosophy, specializing in Clinical Psychology.

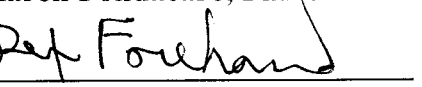
Dissertation Examination Committee:



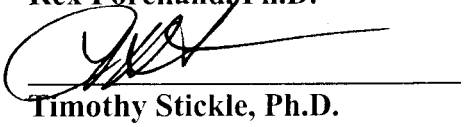
Stephanie McConaughy, Ph.D. Advisor



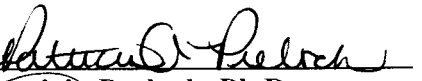
Karen Fondacaro, Ph.D.



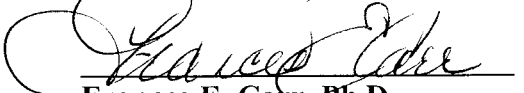
Rex Forehand, Ph.D.



Timothy Stickle, Ph.D.



Patricia Prelock, Ph.D. **Chairperson**



Frances E. Carr, Ph.D. **Vice President for Research and Dean of Graduate Studies**

Date: **December 12, 2007**

ABSTRACT

Individually administered intelligence tests are a routine component of psychological assessments of children who may meet criteria for Attention-Deficit/Hyperactivity Disorder (ADHD), learning disorders (LD), or emotional and behavioral disorders (EBD). In addition to providing potentially useful test scores, the individual administration of an intelligence test provides an ideal opportunity for observing a child's behavior in a standardized setting, which may contribute clinically meaningful information to the assessment process. However, little is known about the associations between test scores and test session behavior of children with these disorders. This study examined patterns of test scores and test session observations in groups of children with ADHD, LD, EBD who were administered the Stanford Binet Intelligence Scales, Fifth Edition (SB5), as well as in control children from the SB5 standardization sample.

Three hundred and twelve children receiving special education services for ADHD (n = 50), LD (n = 234), EBD (n = 28) and 100 children selected from the SB5 standardization sample were selected from a data set of children who were administered both the SB5 and the Test Observation Form (TOF; a standardized rating form for assessing behavior during cognitive or achievement testing of children). The groups were then compared on SB scores and TOF scores. Associations between test scores and TOF scores in children with ADHD, LD, and EBD and normal controls were also examined.

The results of this investigation indicated that children with ADHD, LD, and EBD and normal control children differed on several SB5 and TOF scales. Control children scored higher on all of the SB5 scales than children with LD, and scored higher on many of the SB5 scales than children with ADHD and EBD. Children with EBD demonstrated the most problem behavior during testing, followed by children with ADHD. Children with LD were similar to control children with respect to test session behavior. In addition, several combinations of test scores and test session behavior were able to predict diagnostic group status. Overall, the results of this investigation suggest that test scores and behavioral observations during testing can and should be important components of multi-informant, multi-method assessment of children with ADHD, LD, and EBD.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Psychologists routinely administer intelligence tests to children as part of a comprehensive psychological evaluation to assess for childhood disorders such as Attention-Deficit/Hyperactivity Disorder (ADHD), learning disorders (LD), and emotional or behavioral disorders (EBD) (e.g., Achenbach, 2005; Mayes, Calhoun & Crowell, 1998a). For instance, school psychologists surveyed in 2002 reported conducting approximately 15 ability or achievement tests per month as part of their comprehensive psychological assessment of children in their school districts (Hosp & Reschly, 2002). Another recent survey of practicing school psychologists (Demaray, Schaefer & DeLong, 2003) found that 73% reported using intelligence tests in the assessment of ADHD. The use of intelligence tests in child assessment has two distinct advantages. First, children referred for assessment are often experiencing poor academic performance (Schroeder & Gordon, 2002). Intelligence tests are thought to be useful in determining if the performance is commensurate with the child's ability, generating hypotheses regarding under-performance, and in some cases developing strategies for remediation (Kaufman, 1994; Schwean & Saklofske, 2005; Wallbrown, Vance & Blaha, 1979). In addition, in most states in the U.S., a score on an intelligence test administered within the past 3 years is considered a crucial component to establishing the ability-achievement discrepancy necessary for obtaining special education services for LD, although this is changing as some schools choose the alternative of classifications based on response to intervention (RTI) as allowed by changes to federal law (e.g., No Child Left Behind; IDEA 2004).

Second, an individually administered intelligence test provides an ideal opportunity for observing a child's behavior in a standardized setting. Direct observation of children is considered by many psychologists to be crucial to accurate child assessment (Edwards, 2005; McConaughy 2005). A special section on Evidence-Based Assessment in the *Journal of Clinical Child and Adolescent Psychology* (see Mash & Hunsley, 2005) stressed the importance of observations in the assessment of numerous childhood disorders including ADHD (Pelham, Fabiano & Massetti, 2005), anxiety (Silverman & Ollendick, 2005), and conduct problems (McMahon & Frick, 2005). Landau and Swerdlik (2005) have observed that school psychologists are particularly well-suited to gathering behavioral observations of children. For instance 94% of school psychologists surveyed by Demaray and colleagues (2003) reported using behavioral observations of children in their classrooms as a standard component of an ADHD assessment. Behavioral observations are utilized quite frequently in school settings for a variety of concerns other than ADHD, and surveys have demonstrated that school psychologists conduct 15 or more observations of student behavior in a typical month (Wilson & Reschly, 1996).

However, psychologists who observe children in their classrooms are disadvantaged by the non-standardized nature of these observations. That is, the children are frequently observed informally in different classrooms and participating in various activities ranging from highly constrained, independent academic activities (e.g., solving math problems at their desks) to less formal, group activities (e.g., working on a collaborative art project). Often, children are not observed in any standardized manner

(Konold, Glutting, Oakland & O'Donnell, 1995), which as Hintze (2005) notes fails to meet basic legal requirements or practice standards (e.g., IDEA 2004). Although a recent survey of National Association of School Psychologists (NASP) members revealed that 69% of respondents used some form of systematic observational system on at least 4 of their last 10 referrals for assessment of emotional or behavioral problems (Shapiro & Heick, 2004), even when a standardized system is used, the child's behavior is not compared to normative child classroom behavior or even systematically compared to the behavior of other students in that classroom. In addition, as these ratings are not compared to ratings of the same child's behavior by other raters or in other situations, the reliability of these observations cannot be established (Volpe & McConaughy, 2005).

In contrast, individually administered intelligence tests have a standardized administration, enabling the behavior of an individual child during testing to be compared to the behavior of other children of the same age during the same situation. McConaughy (2005) describes test sessions as "controlled settings" and details the many advantages that observations of children in controlled settings provide, which include the more uniform conditions under which the observations take place and the opportunity to evaluate the impact of specific factors present in the controlled setting (e.g., few distractions, individual attention) on the child's behavior. Intelligence tests also offer a reasonable time period (e.g., typically 45 – 90 minutes) over which observations can be easily gathered without providing undue burden on clinicians, which is important given recent research suggesting that a certain threshold of time spent in observation must be met for the ratings to obtain reasonable correlations with other behavioral measures (e.g.,

McKevitt & Elliott, 2005). Behavioral observations during testing can also be used as checks on the validity of the intelligence scores generated and can provide insight into how students approach and process cognitive tasks (Frisby & Osterlind, 2006; Oakland, Broom & Glutting, 2000). In addition, although empirical support for this position is debatable, many psychologists believe that observations of a child during testing will reveal enduring characteristics or behaviors (e.g., inattentiveness, depressed mood) that the child will also be likely to display in other settings, such as at home or in a classroom (e.g., Kaufman & Lichtenberger, 2000).

Intelligence test scores and observations during testing are well-established components of a comprehensive psychological evaluation of a child, especially when academic difficulties are present. The present study seeks to expand this rich clinical tradition by testing associations between Stanford-Binet Intelligence Scales, Fifth Edition (SB5; Roid, 2003a) test scores and test observations of three diagnostic groups (children with ADHD, LD, and EBD) compared to normal controls selected from the SB5 standardization sample. The diagnostic groups were selected because they represent three of the most common reasons for referral to school psychologists and mental health clinics, and because differential diagnosis in these populations is an important endeavor in psychological assessment (Demaray, Schaefer & DeLong, 2003; Culbertson & Edmonds, 1996). Before discussing the present study, it will be useful to survey current practices in the interpretation of intelligence test scores, including an overview of a number of methodological problems with how test scores are used by psychologists when formulating diagnoses and treatment recommendations, and review what is known about

test scores in children with ADHD, LD, and EBD. Current practices and issues in observations of test behavior, as well as what is known about the test session behavior of children with ADHD, LD, and EBD, will also be discussed in Chapter 2.

CHAPTER 2: LITERATURE REVIEW

Utility of intelligence test profiles

Global intelligence quotient scores such as the Full Scale IQ generated by the Wechsler scales (Wechsler, 1991) or the Total Composite Score generated by the Stanford-Binet scales (Thorndike, Hagen & Sattler, 1986) are known to be stable and to correlate with many important life outcomes, such as school and career achievement (see Groth-Marnat, 1997, for a discussion). The IQ scores may be useful in conjunction with other data for some psychological diagnosis (i.e., mental retardation, learning disability) or for determining appropriate school placement (i.e., gifted/talented). Proponents of IQ testing also advocate using intelligence tests to determine an individual's pattern of cognitive strengths and weaknesses (e.g., Kaufman, 1994; Sattler, 1998). This assertion is based on the theoretical constructs that underlie most popular intelligence tests. These theories posit a more general, higher order factor (such as Spearman's *g*) which incorporates a number of subfactors or subabilities. Thus most IQ tests are composed of several subtests presumed to measure important, but distinct cognitive abilities, that are organized into cognitive domains, with the scores for each domain contributing to the higher-order score. Theoretically, an individual's scores on these cognitive domains or on the individual subtests themselves can be interpreted to reveal areas of higher and lesser ability. This process, referred to generally as *profile analysis*, is thought to reveal useful information about an individual's cognitive processes. An individual can also be compared to profiles generated by exceptional samples, such as persons with LD or ADHD, for assistance in making diagnostic decisions (Kaufman, 1994). Researchers

have been attempting to find useful profiles on commonly used individual intelligence tests such as the Wechsler and Stanford Binet scales for at least 70 years (Oakland et al., 2000). Many major figures in psychological assessment advocate profile analysis (Kaufman, 1994; Sattler, 1998). Although others are strongly opposed to profile analysis, the process is extensively taught in professional training, and is routinely practiced by many psychologists who use intelligence testing (Glutting, McDermott, Watkins, Kush & Konold, 1997; Watkins & Glutting, 2000).

For the Wechsler intelligence scales, the most widely used intelligence tests in the United States and the world (Oakland et al., 2000), more than 75 different subtest profiles have been suggested in the literature as potentially useful for interpretation (McDermott, Glutting, Jones, Watkins & Kush, 1989). For instance, some researchers have found that children with ADHD have lower scores on the Freedom from Distractibility Index (FDI) of the Wechsler Intelligence Scale for Children – Revised (WISC-R; Wechsler, 1974) or WISC – Third Edition (WISC-III; Wechsler, 1991) than children without ADHD (e.g., Andreou, Agapitou & Karapetsas, 2005; Anastopoulous, Spisto & Maher, 1994; Mealer, Morgan & Luscomb, 1996; Prifitera & Dersh, 1993). Although support for this profile is equivocal at best (Mayes, Calhoun & Crowell, 1998a), clinicians still frequently use it for decision-making purposes. Almost 60% of the school psychologists surveyed by Demary et al. (2003) endorsed specifically using the FDI of the WISC-III in ADHD assessments, and 81% of the psychologists who endorsed using intelligence tests routinely indicated that they use the FDI score as part of their ADHD assessment process.

As noted earlier, profile analysis is very popular and has many advocates. For instance, in his widely used book on cognitive assessment of children, Sattler (1998) provides instructions for three different methods of profile analysis. However, Glutting and his colleagues are quite opposed to profile analysis and have presented strong empirical evidence that clinicians should not interpret profiles for strengths and weaknesses (e.g., McDermott, Fantuzzo & Glutting, 1990; Watkins & Glutting, 2000). Some degree of *variability*, or distance between an individual's highest and lowest subtest score that is often interpreted as meaningful, is actually quite common in IQ test profiles (e.g., Kramer, Henning-Stout, Ullman & Schellenberg, 1987). For instance, in a study of 66 children with LD and 51 children without LD, the mean variability in WISC-III subtest scores was approximately 8.5 points (almost 3 standard deviations) for both groups (Mayes, Calhoun & Crowell, 1998b). Variability is more commonly operationalized by clinicians as *scatter* and calculated via a series of comparisons between an individual's mean subtest score and each individual subtest score. Following Sattler's (1998) rules, these differences are then determined to be cognitive strengths or weaknesses if they exceed a critical value (usually 3 points). These putative strengths and weaknesses, however, are extremely common in both exceptional and non-exceptional samples. For instance, in a study by Watkins and Glutting (2000) 92.5% of 1,118 non-exceptional and 538 exceptional students demonstrated at least one statistically significant strength or weakness on the WISC-III. Also, in contrast to more stable global scores, subtest strengths and weaknesses are known to be much less stable and reliable. Sixty percent of strengths and weaknesses are no longer present upon re-testing after only

one month, and a full 80% of the time, specific strengths and weaknesses disappear after three years (McDermott & Glutting, 1997).

In addition, calculation of strengths and weaknesses adds little to the predictive utility of IQ scores. Global IQ scores typically account for one-third to one-half of the variance on achievement scores (e.g., Kline, Snyder, Guilmette & Castellanos, 1993) and a small but potentially meaningful percentage (about 8%) of the variance in learning behavior (McDermott & Glutting, 1997). However, profile scatter accounts for little or no additional variance. In their study using cross-sectional samples from the WISC-III, the Wechsler Individual Achievement Test (WIAT; The Psychological Corporation 1992), and the Differential Ability Scales (Elliot, 1990) standardization samples, McDermott and Glutting (1997) demonstrated that normative subtest scatter (that is, using scaled subtest scores) accounted for only 7% of the unique variance in achievement and 1.7% of the unique variance in learning behavior. Using the more common method of calculating an individual's own relative strengths and weaknesses resulted in a loss of 70% of the unique variance contributed by normative subtest scatter calculations. This commonly used profile analysis method accounted for no unique variance in achievement or learning behavior beyond what was accounted for by global scores and normative scatter.

Strong empirical evidence also exists that clinicians who assess children should not use intelligence test profiles to make diagnostic decisions or formulate diagnostic hypotheses (e.g., Hale & Saxe, 1983; McDermott & Glutting, 1997; Watkins, Kush & Glutting, 1997). Unfortunately, despite the popularity, intuitive appeal, and potential usefulness of subtest profiles, to date they have demonstrated virtually no validity in

predicting a child's emotional, social, or behavioral functioning or diagnosing psychopathology such as ADHD, LD, or EBD (Watkins & Glutting, 2000). One of the major reasons that profiles that appear to be common in samples of children with ADHD, LD, or EBD demonstrate little discriminant validity is that researchers who have advocated particular profiles have failed to account for the base rates of these profiles in the general population (Glutting et al., 1997; McDermott et al., 1989).

As an illustration of the base rate problem, let us return to the example given earlier of children with ADHD reportedly demonstrating low FDI scores on the WISC-R or WISC-III. Other researchers have noted that in addition to low FDI scores, children with ADHD show lower scores on the Processing Speed Index (PSI) of the WISC-III when compared to their other index scores (see Calhoun & Mayes, 2005, for a recent example). This pattern, first demonstrated empirically by Prifitera and Dersh (1993), was named the SCAD profile by Kaufman (1994) in reference to the WISC-III subtests that make up the FDI and PSI (Symbol Search, Coding, Arithmetic, and Digit Span). Similar profiles that substitute or remove various WISC subtests (such as the ACID – Arithmetic, Coding, Information and Digit Span – and CAD – Coding, Arithmetic, and Digit Span profiles) have also been proposed. In general, a lower SCAD (or similar profile) score, which is more indicative of impairment, has been found by some researchers to exist more often in populations with ADHD and/or LD than in normal populations (e.g., Mayes, Calhoun & Crowell, 1998b; Prifitera & Dersh, 1993; Schwean, Saklofske, Yackulic & Quinn, 1993; Watkins et al., 1997). However, at best, these studies demonstrate the phenomenon referred to by Barkley (1996; p. 7) as "high positive

predictive power" but "lousy negative predictive power," meaning that many children with a low SCAD score meet criteria for ADHD, but the absence of the profile does not reliably indicate an absence of ADHD.

Most studies found that while children with ADHD or LD had lower SCAD scores as a group than normal children as a group, few children in the diagnostic group actually show a SCAD (or similar) profile (that is, have 4 of their 5 lowest scores on the subtests that make up the SCAD profile). For instance, while they found lower *mean* ACID and SCAD scores in their sample of 719 children with LD compared to children in the WISC-III standardization sample, Ward, Ward, Hart, Young and Mollner (1995) found that only 4.7% of the children with LD actually showed an ACID profile on the WISC-III. Although this percentage is higher than the percentage of children in the standardization sample who demonstrated the ACID profile, it is certainly of low negative predictive power (95% of children with LD would be false negatives). Ward et al. (1995) found that the SCAD profile (that is, students whose lowest scores were on the four SCAD subtests) was not more common in the children with LD than in the standardization sample. In addition, the ACID and SCAD profiles have also not been able to distinguish between children with ADHD or LD and those without these disorders at a rate exceeding chance or the base rate in the sample (Filippatou & Livaniou, 2005; Ward et al., 1995; Watkins et al., 1997). Attempts to find profiles of other exceptional populations (e.g., EBD) or using other intelligence tests (e.g., SB4, Thorndike et al., 1986; Kaufman Assessment Battery for Children, Kaufman & Kaufman, 1983) have met with similar discouraging results (Kline, Snyder, Guilmette & Castellanos, 1992).

These findings suggest that the common clinical practices of searching for relative strengths and weaknesses or for particular diagnostic profiles are not supported by the empirical literature. In particular, current research does not support efforts to use intelligence test factor scores or subtest scores to diagnose or confirm a diagnosis of LD, ADHD, or EBD in children. However, some clinicians may still feel that profile analysis could be useful for generating hypotheses that might then be tested with other measures and confirmed or disconfirmed (Groth-Marnat, 1997; Hale, Fiorello, Kavanagh, Holdnack & Aloe, 2007). However, Watkins and colleagues (1997) caution against this process due to the high likelihood of making cognitive errors such as only seeking evidence that confirms one's hypothesis, failure to account for base rates, reliance on correlational data, and discounting evidence against one's hypothesis. Such caveats are well taken given the high percentage of school psychologists who endorse using the FDI of the WISC-III in the assessment of ADHD despite overwhelming empirical evidence that the FDI is not a good discriminant measure of that disorder.

While the evidence clearly suggests that the current practice of using IQ test profiles diagnostically is not appropriate or useful, this does not mean that the search for profiles, or for differences between groups on IQ subtests or domains in general, is not a meaningful research question. Group differences, if they exist, can help to conceptualize the general nature of the disorders in question. For instance, both LD and ADHD are thought to be heterogeneous disorders (e.g., Anderson & Stanley, 1992; Rose, Lincoln & Allen, 1992) with many subtypes or variations. Knowledge of how and to what extent groups of children with ADHD or LD differ from non-ADHD and non-LD populations

can help determine the nature of the deficits that characterize these disorders. In addition, profiles that characterize some, but not all, individuals in a particular diagnostic group may (in conjunction with other data) help define subgroups within that diagnosis. For example, perhaps children with ADHD who do show impaired performance on the FDI differ in important ways from children with ADHD who do not demonstrate lower FDI scores.

Failure to find group differences on subtests or domains, when differences are expected, can also lead to useful research hypotheses. For instance, the inability of IQ factor scores to discriminate between LD and ADHD populations has led some researchers to speculate about common factors underlying these disorders, such as problems with attention and memory (e.g., Reid, Hresko & Swanson, 1996). Others have hypothesized that the inability of the FDI of the WISC-III to distinguish between ADHD and non-ADHD groups or to correlate with other measures of attention suggests that the FDI more accurately indexes learning or memory problems than distractibility (e.g., Krane & Tannock, 1992; the FDI index was subsequently renamed the Working Memory Index when the WISC-IV was introduced). Finally, the fact that meaningful profiles have not been discovered with current intelligence tests does not preclude their discovery in future measures (Kline et al., 1992), especially if those measures are designed to be sensitive to the deficits believed to characterize particular disorders. The recent Stanford-Binet Intelligence Scales, Fifth Edition (SB5) were, according to its publisher, “designed with ADHD in mind” (The Riverside Publishing Company, 2004), and the author of the SB5 hopes that it will be able to differentiate children with this disorder from those

without ADHD (Roid, 2003b). It is also hoped that revisions of the SB5 such as the addition of the Working Memory subtests will offer greater utility in the assessment of individuals with learning disorders (Mleko & Burns, 2005). In addition, when the Stanford-Binet Intelligence Scales, Fourth Edition were designed (Thorndike et al., 1986), the SB4 represented a significant departure from previous editions in that it incorporated verbal and nonverbal factors similar to those found on the Wechsler scales. One of the stated rationales for this change was to be sensitive to individuals who might show discrepancies between their abilities in these areas (Mleko & Burns, 2005).

Intelligence test scores for children with ADHD, LD, and EBD

Bearing in mind the above caveat that group differences cannot, at present, be used to predict a specific diagnosis for an individual, we are actually aware of some ways in which children with ADHD, LD, and EBD differ from children without these disorders (when considering only children without mental retardation). The most robust findings exist for global IQ scores, which is consistent with the research summarized thus far on the lesser utility of profile analysis. In general, the mean IQs for children with ADHD, LD, and EBD have been lower than the mean IQ of children from the standardization sample of the IQ test used (Gathercole, Alloway, Willis & Adams, 2005; Kaufman & Lictenberger, 2000; Zimmerman & Woo-Sam, 1997). Children with LD show the greatest deviation from standardization sample means, with mean IQ scores approximately one standard deviation below the mean (e.g., Doll & Boren, 1993; Canivez, 1996; Lavin, 1996; Prewett & Matavich, 1993) although this pattern is not

always found (see Kaufman & Lictenberger, 2000). As a group, children with EBD demonstrate mean IQ scores that are one-half to one standard deviation below the mean (e.g., Connery, Katz, Kaufman & Kaufman, 1996; Javorsky, 1993; Lavin, 1996; Slate & Jones, 1995). Children with ADHD are found to have mean IQ scores one-third to one-half of a standard deviation below the mean (e.g., Barkley, DuPaul, & McMurray, 1990; Faraone et al., 1998; King & Young, 1982; Saklofske, Schwean, Yackulic & Quinn, 1994; see also similar results in a meta-analysis of 18 studies of adults with ADHD in Bridgett & Walker, 2006), although again, this pattern is not always found (see Schuck & Crinella; 2005; Semrud-Clikeman, Hynd, Lorys & Lahey, 1998). In addition, most studies have found no difference in IQ scores between subtypes of ADHD (Hyperactive-Impulsive, Inattentive, and Combined types; see Milich, Balentine & Lynam, 2001, for a review). Although these studies suggest a general trend for lower global IQ scores in children with ADHD, LD, and EBD compared to children from standardization samples, it is important to remember that there are actually wide overlaps between the ranges of the IQ scores for all the diagnostic groups and normal controls.

The literature with regard to specific deficits or strengths that children with ADHD, LD, and EBD might be expected to show on measures of IQ is more difficult to summarize because of the different measures and diverse populations used in the studies. However, some general conclusions, as well as hypotheses as to how these findings might translate to SB5 performance, can be presented. In research using the Wechsler scales, there is some debate regarding whether children with ADHD show higher verbal than nonverbal scores (e.g., Mahone et al., 2003), higher performance than verbal scores

(e.g., Andreou et al., 2005; Saklofske et al., 1994; Saklofske, Schwean & O'Donnell, 1995), or equivalent verbal and nonverbal performance (e.g., Carter, Zelko, Oas & Waltonen, 1990; Naglieri, Goldstein, Iseman & Schwebach, 2003). Some researchers have postulated that these equivocal findings reflect the Wechsler verbal and performance scales' insensitivity to the nature of the deficits shown by children with ADHD. When examining their performance on intelligence tests organized into different factors, such as the K-ABC, Woodcock-Johnson Cognitive (Woodcock, McGrew & Mather, 2001) and the Cognitive Assessment System (CAS; Naglieri & Das, 1997), children with ADHD show clear deficits in Sequential Processing, Working Memory, and Planning (e.g., Ford, Floyd, Keith, Fields, & Schrank, 2003). These findings are consistent with other research demonstrating that regardless of subtype, children with ADHD are noted to have deficits in working memory (McInnes, Humphries, Hogg-Johnson & Tannock, 2003; Shapiro, Hughes, August & Bloomquist, 1993; Stevens, Quittner, Zuckerman & Moore, 2002; Schwean & Saklofske, 2005). The deficits in Sequential Processing and Planning are also consistent with the findings summarized earlier showing lower group mean scores for children with ADHD on the FDI of the Wechsler scales compared to children without ADHD. Interestingly, research using subtests from the Children's Memory Scale (Cohen, 1997) and the Wide Range Assessment of Memory and Learning (Adams & Sheslow, 1990) suggests that children with ADHD are more impaired in spatial working memory than in verbal working memory (McInnes et al., 2003). The Wechsler FDI tasks (which are believed to tap working memory) are entirely verbal. By contrast, the SB5 utilizes both verbal and

nonverbal (spatial) working memory tasks, offering us an opportunity to assess for the subtle distinctions suggested by McInnes and colleagues (2003).

Although speculation regarding global verbal-nonverbal discrepancies on the SB5 for children with ADHD is premature, the available literature does suggest that differences between verbal and nonverbal performance may emerge for children with LD and children with EBD. Research on children with LD (especially LD in reading) has usually found higher nonverbal than verbal IQ scores (see Riccio & Hynd, 2000 for a review). For example, Anderson and Stanley's (1992) cluster analysis of WISC scores of children with LD yielded a five-group solution, with four groups demonstrating verbal IQ scores significantly lower than nonverbal IQ scores. The group which did not show this pattern (in fact, their verbal IQs were significantly higher than their performance IQ scores) displayed lower performance on the Coding, Arithmetic, and Digit Span tests (CAD profile), suggesting to the authors that this group may be an attention problems group. Mayes, Calhoun, and Crowell (1998b) also found higher nonverbal than verbal WISC-III IQ scores in children with LD only, but no significant difference in verbal and performance IQ scores in children with comorbid LD and ADHD. Furthermore, Ottem (2002) has suggested that the design of the Wechsler scales, which is comprised of more complex performance tasks than complex verbal tasks, actually underestimates the verbal-performance discrepancy noted in children with LD, which may account for some researchers' failure to find differences between children with and without LD. In terms of even finer IQ score distinctions, as reported earlier, children with LD are just as likely as children with ADHD to show working memory problems (e.g., Gathercole et al., 2006),

and as such have scored lower on the FDI and PSI than on other indexes of the Wechsler scales (Mayes et al., 1998a; Swanson, 2005). In contrast, children with EBD (without comorbid ADHD or LD) have not been found to demonstrate working memory problems (Carter et al., 1990; Naglieri et al., 2003). However, the pattern of higher mean nonverbal IQ scores that is seen in children with LD is also seen in children with EBD (e.g., Lipsitt, Buka & Lipsitt, 1990), especially if the EBD is externalizing in nature. Children with more internalizing symptoms may actually show higher verbal than nonverbal IQ scores (e.g., Naglieri et al., 2003).

It is possible that newer intelligence tests, such as the SB5, that target both verbal and nonverbal working memory may be more sensitive to group differences between children with ADHD, LD, and EBD. For instance, children with ADHD and children with LD are expected to be similar on working memory indexes, but may differ in terms of the discrepancies between their verbal and nonverbal IQ scores. Children with externalizing EBDs may appear similar to children with LD in terms of verbal-nonverbal discrepancies, but show no working memory difficulties. Group differences like these, if they exist, could inform future research into the nature of these disorders and in what ways they differ from, or are similar to, other childhood disorders. Additionally, if these discrepancies exist not just between groups, but also appear when looking at the level of the individual child, the patterns an individual child obtains might help clinicians feel more confident in the diagnoses they provide. This process would represent a much more refined, empirically supported use of test data than current subtest profiling practices.

Some research, reported in the SB5 *Technical Manual* (Roid, 2003a), is available on the SB5 test scores of children with ADHD, LD, and EBD compared to each other and to normal controls. As part of the process of establishing criterion-related validity, the SB5 was administered to children from a number of exceptional groups, including children with ADHD (n = 94), LD (n = 300), and EBD (n= 48). Children were identified as eligible for the ADHD, LD, or EBD group if they were receiving special education services under that primary classification. The group means on the five SB5 factor scores and the Verbal, Nonverbal, and Full-Scale IQ are reproduced here in Table 1 and presented graphically in Figure 1. Some of the patterns anticipated above, such as lower mean IQ scores for diagnostic groups compared to the standardization sample and children with ADHD showing significantly lower scores on Working Memory than on all other factors, can be seen in Table 1 and Figure 1. However, there are a number of problems with the scores reported by Roid (2003a). The diagnostic groups used in the studies included a large number of children with Full Scale IQs in the mental retarded (MR) range, which could have significantly impacted the group mean scores. In addition, there was a substantial amount of participant overlap between the diagnostic groups. For instance, children with ADHD and comorbid LD were included in both the ADHD and LD samples, children with 2 or more types of LD were included in each LD category, and children with ADHD, LD, or EBD who also met criteria for another exceptional group (e.g., Autism, Speech and Language Disorders) were included in both groups. The diagnostic groups also included children under age 5 and some individuals over age 19. Additionally, gender and age groups were not included in the analyses. For these reasons,

the first aim of the present study is to compare SB5 scores in non-MR children ages 6-18 with ADHD-only, LD-only, or EBD-only to scores from non-MR children ages 6-18 selected from the SB5 standardization sample, considering gender and age group in the analyses. These results will provide a clearer picture of how SB5 intelligence test scores differ among diagnostic groups and normal controls.

Utility of observations of test behavior

Because assessment and differential diagnosis of children with ADHD, LD, and EBD are highly relevant to school-based and child mental health clinicians, it is important to consider all of the data generated from the administration of an intelligence test to evaluate its clinical usefulness. So far, I have summarized the research to date on the actual test scores produced by children with ADHD, LD, and EBD. I turn now to a review of the second type of information available from an individually administered IQ test: observations of the child during testing.

During testing sessions, children completing a standardized series of tasks can be observed by trained examiners who have observed many other children complete the identical series of tasks in similar settings. There are many advantages to clinician observations of the child during test sessions. In general, clinicians have the potential to be less interested in the outcome of the observations than other parties such as parents or teachers (because they are likely less invested in the child receiving particular services or a specific diagnostic label), to be more familiar with behavior and development across childhood than parents, and to be more knowledgeable about behavioral observations

than either teachers or parents (Glutting, Youngstrom, Oakland & Watkins, 1996). In addition, because the test session involves a standardized administration, the clinician can compare the child's behavior to the behavior of other children of the same age and gender who were exposed to the same stimuli under very similar circumstances and who completed the same tasks. As Glutting and his colleagues note, "None of the major contexts of child development (e.g., home, school, and community) offers as high a level of professional expertise, observational control, or uniformity of conditions as the context of individual test-taking" (1996; p. 94). Unfortunately, the potential utility of test observations is often squandered when test observations are not gathered in any standardized fashion. Typically, clinicians simply record a narrative of the child's behavior, with no objective reference to typical child behavior during testing nor use of any coherent, empirically based system for integrating that behavior into an overall picture of the child's behavior. Narrative descriptions of test session behavior obtained in this manner have not been systematically investigated in any research studies to date (Frisby & Osterlind, 2006).

Many of the published assessment measures for children's test behavior offer no normative information. If the system for recording the child's test behavior is organized at all, it is usually by way of a checklist-type system with intuitive, but not empirically based, categories of test session behavior. An example is the Behavior and Attitude Checklist (Sattler, 2001), which has 28 items organized into 12 intuitive domains such as Attitude toward examiner, Reaction to failure, and Gross motor skills. Another example is the 1986 version Stanford Binet Observation Schedule (SBOS) that was included with

the record form for the SB4 (Thorndike et al., 1986; a similar form was included with the record form for the SB3, Terman & Merrill, 1960). The 1986 version of the SBOS has 16 items, rated on a 5-point scale, that are grouped into 5 rationally-derived domains (Attention, Reactions During Test Performance, Emotional Independence, Problem-Solving Behavior, and Independence of Examiner Support). A more recent example, the Test Session Observation Checklist (TSOC) designed for use with the WJ-III COG, contains 7 items derived through rational analysis (e.g., Level of Cooperation, Care in Responding), and the clinician is asked to choose which of five descriptive statements for each item best describes the student (e.g., for the item Care in Responding, choices range from "very slow and hesitant in responding" to "impulsive and careless in responding"). Although the TSOC has the advantage of some research on its correlation with the WJ-III COG (Frisby & Osterlind, 2006), all of these rationally-derived checklists are hampered by a lack of standardization, lack of normative samples, and little or no data on reliability or validity.

Two exceptions to these intuitive, non-standardized systems are the Guide to the Assessment of Test Session Behavior (GATSB; Glutting & Oakland, 1993) and the Test Observation Form (TOF; McConaughy & Achenbach, 2004), both of which are standardized rating forms with empirically derived scales and normative samples. The GATSB has 29 items measuring a child's approach to testing and interaction with the examiner. Each GATSB item is rated on a 3-point scale. Example items include "Exhibits rigid and inflexible approach to tasks", "Hesitates when giving answers", "Does not look examiner in the eye", and "Listens attentively to directions and test items." The GATSB

has three scales derived through factor analysis: Avoidance, Inattentive, and Uncooperative Mood. The TOF has 125 items that measure a range of behaviors that children might exhibit during a test session. Items are rated on a 4-point scale. Sample TOF items include “Slow to warm up”, “Concrete thinking”, “Asks for feedback about performance”, “Tries to manipulate examiner”, and “Doesn’t concentrate or pay attention for long on tasks, questions, topics.” The TOF has five syndrome scales derived through factor analysis: Withdrawn/Depressed, Language/Thought Problems, Anxious, Oppositional, and Attention Problems, plus Internalizing, Externalizing, and Total Problems, and a scale measuring problems associated with ADHD (see Methods section for a full description of the TOF).

Assuming test session observations are collected using empirically derived scales with normative samples, behavior observations could potentially contribute two important types of information: (1) test session behavior may inform clinicians as to the validity of the cognitive or achievement test scores generated by the test, and (2) test session behavior may provide the clinician with information that could illuminate how the child might act in non-test situations. Glutting and colleagues (1996) refer to the first type of information as *intrasession validity* while the second type of information is termed *exosession validity*. Intrasession validity is important because it serves as a guideline for how much confidence can be placed in the scores that an individual child achieves on intelligence tests. Authors of intelligence tests have long been aware that optimum performance is obtained when the child is comfortable in the testing situation, understands test expectations (including the expectation to do well), understands the

instructions, and is motivated, attentive, and cooperative (e.g., Konold et al., 1995). However, little research has assessed how behavior not conducive to optimal testing systematically impacts test scores.

Research on intrasession validity could provide clinicians with expectations regarding how test scores can be affected (or not affected) by particular patterns of behavior, such as lack of cooperation, test anxiety, or inattentiveness. Current practice by most clinicians involves making a clinical judgment that test scores likely were, or were not, impacted by the child's behavior during testing. Researchers have shown some support for the premise that more behavior problems during testing is associated with lower IQ scores. For example, Scores on the GATSB Total Score and the three GATSB scales correlated -.21 to -.39 with WISC-III Full Scale IQ scores (Glutting & Oakland, 1993) and -.32 to -.41 with the Woodcock-Johnson Revised (WJ-R, Woodcock & Johnson, 1989; Daleiden, Drabman & Benton, 2002), and scores on the TSOC Global Impressions score correlated -.47 with the Global General Ability index on the WJ-III COG. Scores on the TOF Total Problems scale correlated -.29 with SB5 Full Scale IQ Scores, and the TOF syndrome scales correlated -.21 to -.40 with SB5 FSIQ scores (except Anxious, no correlation; McConaughy & Achenbach, 2004). It is important to note that despite the common clinical practice of ascribing causality to this relationship such that poor test session behavior leads to a child's underperformance on IQ tests relative to their true ability, the associations found in the research are only correlational. That is, it is equally possible that the correlation results from lower IQ scores leading to increased problems during testing (for example, a child being avoidant or oppositional

when continually confronted with tasks beyond his/her ability), or that a common factor underlies both scores (for example, subtle cognitive deficits lead to higher scores on the TOF Language/Thought Problems scale as well as lower IQ scores).

Although intrasession validity is undeniably important, clinicians are arguably more interested in the potential exosession validity of test observations. Every clinician who records test session behavior and includes this information in an assessment is implicitly hypothesizing that the sample of behaviors demonstrated during the context of the test session will generalize to other contexts, and that the observational data, either alone or in conjunction with additional behavior ratings, will yield important insights into the difficulties the child may be having. As an example, an examiner may observe that a child referred for evaluation because of oppositional behavior displays a particular pattern of test behavior including perseveration on topics and difficulty shifting his cognitive set to the next task. The examiner might hypothesize that this behavior is also apparent when the child is at home or at school, and that some of his reported oppositional behaviors in these settings are due to cognitive or emotional difficulties adjusting to transitions. The clinician might recommend interventions targeted towards helping this child successfully negotiate the transitions he encounters. As a second example, an examiner might observe that a different child referred for evaluation because of academic difficulties has great difficulty concentrating and sustaining attention during testing. This same child might show significant restless and fidgety behavior during the test session. The clinician might hypothesize that the concentration difficulties and

restlessness are present at home and school as well, and may therefore conclude that further assessment of possible ADHD is warranted for this child.

Although intuitively appealing, the clinician is currently without solid empirical grounding in generating such hypotheses and recommending interventions, because little is known about the exosession validity of test behavior observations. Using both nonstandardized test observation measures and the GATSB, Glutting and his colleagues have argued that the test session behavior of children is only moderately correlated with teacher ratings classroom behavior (e.g., average r = .12 for normal controls and r = .16 for referred children; Glutting et al., 1996). Teacher ratings of learning style showed average correlations of -.25 with test session behavior in non-referred children (Glutting & McDermott, 1988). Gordon, DiNiro, Mettelman and Tallmadge (1989) also found that examiner ratings of a child's behavior during a continuous performance task correlated .25 with teacher ratings of ADHD problems. Daleiden and colleagues (2002) found that the GATSB showed only modest correlations with parent ratings of child behavior on the Child Behavior Checklist (CBCL; Achenbach 1991), with the Total Problems score on the CBCL correlating .23 with the Total score on the GATSB.

While these authors interpret their results as evidence of the lack of exosession validity of test session behaviors, other explanations have been offered. For instance, Kaplan (1993) has noted that the interrater reliability of the GATSB has not been adequately established. If examiners complete the GATSB in idiosyncratic ways, then low correlations between GATSB scores and other ratings of child behavior would be expected due to the unreliability of the GATSB. An even more compelling argument is

that Glutting and his colleagues are comparing the scores of two different raters of a child's behavior in two separate contexts (e.g., examiner and teacher). A growing body of research is attesting to only moderate correlations between ratings of children's behavior by different raters across different settings. For instance, a meta-analysis conducted by Achenbach, McConaughy, and Howell (1987) found an average correlation of .28 between informants rating children's behavior in different contexts. Thus, the correlations between test session behavior as rated by examiners and classroom behavior as rated by teachers or home behavior as rated by parents are quite consistent with other correlations between multiple informants. Glutting and his colleagues' results suggest that test session behavior, like child behavior in other major contexts of development, is specific to the situation and correlates only modestly with behavior in other situations. However, just as clinicians consider information from both parents and teachers to be important sources of meaningful data despite modest correlations between teacher and parent reports, test session observations may be useful information despite only moderate associations with other ratings of the child's behavior.

In contrast to the GATSB, the TOF has the advantage of demonstrating adequate interrater reliability (McConaughy & Achenbach, 2004; see Methods section for more details). Another advantage of the TOF is that information is available regarding the correlations between the TOF and other theoretically similar measures from the Achenbach System of Empirically Based Assessment (ASEBA). There were: 39 significant correlations (22 medium and 17 small effects) between the TOF and the CBCL; 50 significant correlations (27 medium and 23 small effects) between the TOF

and the Teacher's Report Form (TRF; Achenbach, 1991); and 7 correlations (1 large and 6 medium effects) between the TOF and the Youth Self Report (YSR; 1991c). The magnitude of these correlations was .26 to .43, which is consistent with the correlation of .28 between different types of informants in Achenbach et al.'s (1987) meta-analysis.

Test behavior observations in children with ADHD, LD, and EBD

Given the dearth of research on test session behavior overall, it is perhaps not surprising that little is known about the test session behavior of children with ADHD, LD, and EBD. Research using informal observations of test session behavior suggests that children with ADHD move more frequently than children without ADHD while being tested (with a continuous performance task; Teicher, Ito, Glod & Barber, 1996). Similarly, EBD/LD children who frequently looked away, verbalized, and got out of their seats (measures of inattention and hyperactivity) during testing were more likely to be classified as abnormal on teacher ratings of attention problems and on a continuous performance task than EBD/LD children who were rated as less inattentive and hyperactive during the test session (Gordon et al., 1989). Using the GATSB Inattentive score, Glutting, Robins, and deLancey (1997) were able to correctly classify 71% of children with ADHD and 90% of children without ADHD (81% overall correct classification rate). Interestingly, these results suggest higher exosession validity than Glutting and his colleagues typically ascribe to the GATSB (e.g., Glutting et al., 1996). McConaughy and Achenbach (2004) reported significantly higher mean scores for referred children compared to nonreferred children on all TOF empirically-based and

DSM-oriented scales. Discriminant analyses also showed that a weighted combination of the five TOF syndrome scales correctly classified 50 – 59% of referred children and 81 – 91% of nonreferred children (71% overall correct classification rate). However, as the referred group contained children referred for myriad psychological difficulties, more specific information about TOF scores for children with ADHD, LD, or EBD as rated by the TOF is not yet available. Overall, the limited research that is available regarding test session behavior suggests that children with ADHD, LD, and EBD may have different patterns of behavior during test sessions that may be clinically informative. Thus, the second major aim of this study is to compare patterns of test session behavior as measured by the TOF in children with ADHD, LD, and EBD and normal controls.

The relationship between gender and test session behavior of children with ADHD, LD, and EBD and normal children has also received little study. In their studies using the GATSB, Glutting and colleagues found no differences in test session behavior or GATSB factor structure between boys and girls nor significant item bias that could be attributed to gender (Glutting, Oakland & Konold, 1994; Oakland & Glutting, 1990; Konold et al., 1995). However, the lack of differences in scores across gender may be an artifact of the type of items that comprise the GATSB (i.e., items that concern only the child's approach to the test and towards the examiner) and may not adequately capture different patterns of behavior displayed by boys and girls during test sessions. For example, boys demonstrate more behavior problems during testing than girls when rated with the TOF (McConaughy & Achenbach, 2004). However, specific information about gender differences in test session behavior of children with ADHD, LD, and EBD is

lacking. The relationship between age and test session behavior in children with ADHD, LD, and EBD has also received little, if any, study. Younger children in the TOF standardization sample and the referred sample displayed more problem behavior during testing than older children (McConaughy & Achenbach, 2004), but again, research on the associations between age and test session behavior for children with ADHD, LD, and EBD is not currently available. Thus, an important component of the second aim of this study is to include gender and age group as predictors in the analyses examining the relationships between test session behavior and diagnostic status.

Test scores and test session behavior in children with ADHD, LD, and EBD

There is a clear need for additional research on the patterns of IQ test scores in children with ADHD, LD, and EBD on newer IQ tests designed to be sensitive to the deficits thought to underlie these disorders. In addition, further research on the test session behavior of children with ADHD, LD, and EBD using a standardized measure with a normative sample is also important. However, clinicians are presumably more interested in the associations *between* test scores and test session behavior in children of different diagnostic groups compared to other diagnostic groups and normal children. Psychologists are encouraged by major figures in assessment (e.g., Kaufman, 1994; Sattler, 1998) to pay considerable attention to both test scores and test session behavior when conducting a psychological assessment, and the implicit assumption is that the integration of these two sources of information is more meaningful than either source of data alone. For instance, it is reasonable to suspect that a clinician might feel more

comfortable making a diagnosis of ADHD if the child behaved in an inattentive, impulsive, and hyperactive manner during testing and that child's test scores showed a relative weakness on tasks measuring memory or processing speed. Unfortunately, there is little, if any, research that examines associations between IQ scores and test session behavior in normal children or in children with ADHD, LD, or EBD. In fact, few researchers have combined any behavioral ratings or observational data with standardized test data to predict group membership, although some recent studies suggest that this may be a worthwhile endeavor. For example, Vile Junod and colleagues found that adding standardized classroom behavioral observations to achievement test results and SES status combined accurately predicted group status (ADHD vs. peer controls) at 92% (Vile Junod et al., 2006).

Another important question that has yet to be examined is whether the combination of test scores and test session behavior will more usefully inform diagnosis than either test scores alone or test session behavior alone. As a hypothetical example, a clinician may be more likely to accurately diagnose a child as having LD if her SB5 scores show a large verbal-nonverbal discrepancy and the child scored in the clinical range on the TOF Language/Thought Problems scale than they would be if they based her diagnosis on test scores only or test session behavior only. Thus the third aim of this study is to test various combinations of scores for test session observation and intelligence test scores for predicting diagnostic group status.

Present study

Given the unique potential advantages of utilizing individually administered intelligence tests in the assessment of childhood psychopathology, namely the opportunities to obtain important cognitive information and to observe the child under standard conditions, it is imperative for psychologists to develop empirically-supported methods for interpreting and integrating test scores and test observations. The recent publications of the SB5, with its strong theoretical basis and statistical support, and the comprehensive, standardized TOF, co-normed with the SB5, provide an excellent opportunity to examine associations between patterns of intelligence test scores and patterns of test session behavior for children with cognitive, emotional, and behavioral difficulties. The aims of this study are threefold:

Aim 1: To compare the patterns of IQ scores of children with ADHD, LD, and EBD, and normal controls on the SB5.

Although some differences in performance on IQ tests between children with ADHD, LD, and EBD and normal controls have been documented in the literature, little research exists that directly compares these diagnostic groups to each other and to children who do not receive special education services. In addition, most of the existing research on these samples uses outdated intelligence tests (e.g., the WISC-R). Although group means for samples of children with ADHD, LD, and EBD were reported in the SB5 *Technical Manual* (see Table 1), there was considerable participant overlap and comorbidity in these samples and many of the participants had IQs in the MR range.

Thus, no information yet exists about the SB5 performance of children with normal intelligence who qualify for special education services for non-comorbid ADHD, LD, and EBD. This research may be especially important because the SB5 was specifically designed to be sensitive to the deficits displayed by children with ADHD and LD (Riverside Publishing Company, 2004). In addition, little research is available regarding the relationships between gender, age, and IQ scores in children with ADHD, LD, and EBD, although what literature is available suggests that gender at least may be important predictor (e.g., Preiss & Lenka, 2006). Therefore an important component of this aim is to include gender and age group in these analyses. Consistent with previous research, it is hypothesized that children in the three diagnostic groups (ADHD, LD, and EBD) will score lower than children from the standardization sample on all SB5 scores. It is also hypothesized that the SB5 will be sensitive to deficits experienced by children with ADHD and LD; for example, children with ADHD and LD are expected to have lower Working Memory scores than Control children, while this pattern is not anticipated for children with EBD. Post hoc tests will provide information on additional SB5 test score patterns.

Aim 2: To compare patterns of test session behavior among children with ADHD, LD, and EBD and normal controls on the TOF.

The TOF measures a wide range of behaviors a child may display during a testing session, including behaviors that may be clinically informative but not directly related to the child's approach to testing. The second aim of this study is to compare TOF scores

across diagnostic groups and to TOF scores of normal controls. Because of the limited research on test session behavior of children with ADHD, LD, and EBD, specific hypotheses are difficult to formulate. However, a few hypotheses seem appropriate based on previous findings and the theoretical constructs discussed above. First, children with ADHD are expected to score higher than children with LD or EBD and normal children on the TOF Attention Problems syndrome and the DSM-oriented Attention Deficit Hyperactivity Problems scale and its Hyperactivity-Impulsivity and Inattentive subscales. Children with LD are expected to score higher on the TOF Language/Thought Problems syndrome. Children with EBD are expected to score higher on the TOF Oppositional and Withdrawn/Depressed syndromes compared to other diagnostic groups and normal children. Post hoc tests will provide information regarding additional patterns.

Aim 3: To determine if a combination of SB5 scores and TOF scores will differentially predict clinical diagnostic group status.

Previous research has suggested that to date, test scores cannot be used to predict diagnostic group status or to inform diagnostic decisions. Observations of test session behavior fare better, in that a weighted combination of TOF syndrome scores correctly classified 71% of children as referred or nonreferred (McConaughy & Achenbach, 2004) and the GATSB Inattentive score correctly classified 81% of children as ADHD or not-ADHD (Glutting et al., 1997). However, the diagnostic utility of intelligence test scores combined with test session behavior for discriminating children with ADHD, LD, and EBD has yet to be tested. The third aim of this study is therefore to determine if a

combination of intelligence test scores and observations of test session behavior will be clinically meaningful in terms of predicting diagnostic group status. Because the SB5 was designed to be sensitive to deficits shown by children with ADHD, LD, and EBD, and given that several TOF factors (e.g., Language/Thought Problems, Attention Problems, Withdrawn/Depressed, Oppositional Behavior) directly assess for behaviors conceptually related to ADHD, LD, and EBD, it is hypothesized that a weighted combination of intelligence test scores and scores for test session observations will differentially predict diagnostic group status. The results regarding these hypotheses are expected to have important clinical utility for assessing these three groups of children.

CHAPTER 3: METHOD

Participants

Referred sample. For establishing validity in exceptional populations, SB5 was administered to 3,000 individuals with special needs, such as individuals who are deaf, physically disabled, or learning disabled (the SB5 exceptional sample). A portion of the SB5 exceptional sample, containing SB5 and TOF data, was provided to Dr. Stephanie McConaughy by the Riverside Publishing Company. For the present study, a subsample of the data set provided to Dr. McConaughy was created by selecting cases representing children between the ages of 6 and 18 who were eligible for special education services in their schools for ADHD, LD, or EBD. Whether or not a participant qualified for special services was determined by the examiner by examining the child's school records (Riverside Publishing Company, personal communication, 2005). Children who meet criteria for special education services must have been evaluated by a school multidisciplinary team (or an independent evaluator), have been found to have a disability, and must have been found in need of special education services. For the purposes of this study, the classification under which the child was receiving Special Education was determined by the examiner's response to section VII of the TOF, which states "Does the child meet criteria for a special education disability, Section 504 plan, or other service category?" and then asks the examiner to review the categories under which the child receives services (e.g., by examining the child's Individualized Education Program), and check the categories that apply. The available categories on the TOF are ADD/ADHD; LD; mental retardation, developmental delay, learning impaired;

perceptual-motor disability, physical therapy, occupational therapy; EBD; speech or language impairment or delay; gifted, advanced, accelerated, enrichment; counseling, guidance, therapy; chronic health impairment (not ADD or ADHD); and other. Study participants were selected if one (and only one) of ADHD, LD, or EBD was selected, and participants were excluded if mental retardation or speech or language impairment was also selected in addition to ADHD, LD, or EBD. One participant who met criteria for the sample was excluded from further analyses through preliminary data screening for outliers; that child's TOF Total Problem score was 157, which is over 6 standard deviations above the mean TOF Total Problems score for the referred sample (Mean = 13.88, SD = 21.92).

The resulting sample consisted of 312 children receiving special education services for ADHD (n = 50), LD (n = 234), or EBD (n = 28). Children were excluded from the subsample if their Full Scale IQ Score on the SB5 was less than 70. There were more males (n = 193) than females (n = 119) in the sample, which is consistent with research demonstrating that male children are referred 2-9 times more frequently for ADHD, LD, and disruptive behavior disorders than females (American Psychiatric Association, 2000). Further details on the three groups is provided below and in Table 2. In addition, mean SB5 Factor scores and TOF Syndrome scores for each diagnostic group and for the nonreferred sample are presented in Tables 3 and 4.

ADHD group. The ADHD group consisted of 50 children and adolescents (32% female) receiving special education services related to a diagnosis of ADHD who were administered the SB5. Mean age was 10.74 years (SD = 3.48). Ethnicity for this sample

was 70% Caucasian, 17% African-American, 10% Hispanic, and 3% other. The Mean FSIQ on the SB5 was 96.70 (SD = 15.48). Mean TOF Total Problems was 23.16 (SD = 27.95).

Of the 50 children in the ADHD group, 30 children were reported to be on medication and 20 were reported to be not on medication during administration of the SB5 and TOF (based on the examiner's response to section VI of the TOF, "Was the child on medication when tested?"). There were no significant differences between the medicated and unmedicated children on any SB5 scale. Children with ADHD who were not on medication scored significantly higher than children with ADHD who were on medication during testing on three TOF scales: the Attention Problems scale (no medication M = 12.79; SD = 10.27; on medication M = 7.23; SD = 6.96), the DSM-oriented Inattentive scale (no medication M = 7.90; SD = 4.76; on medication M = 4.33, SD = 4.91), and the DSM-oriented ADHP scale (no medication M = 15.05; SD = 11.52; on medication M = 8.53; SD = 9.45). Children with ADHD who were not on medication during testing did not score significantly differently from children with ADHD who were on medication during testing on any of the other TOF scales.

LD group. The LD group consisted of 234 students (40% female) who were receiving special education services for a learning disability in math, reading, or writing who were administered the SB5. The mean age for the LD sample was 10.53 years (SD = 2.44). Ethnicity for this sample was 50% Caucasian, 11% African-American, 35% Hispanic, and 4% other. Mean FSIQ was 89.18 (SD = 10.59). Mean TOF Total Problems was 8.61 (SD = 14.49).

EBD group. The EBD group consisted of 28 students receiving special education services for serious emotional disturbance (36% females) who were administered the SB5. The mean age was 11.71 years (SD = 2.85). Ethnicity for this sample was 48% Caucasian, 25% African American, 17% Hispanic, and 10% other. Mean FSIQ was 91.86 (SD = 15.47). Mean TOF Total Problems was 41.32 (SD = 33.55).

Nonreferred sample. Because the TOF and the SB5 were co-normed, a large percentage of children in the standardization sample of the SB5 were also rated on the TOF (TOF/SB5 sample; n = 2,442). For the present study, a nonreferred sample was created by randomly selecting 100 cases from the TOF/SB5 sample of children between the ages of 6-18 who obtained an SB5 Full Scale IQ ≥ 70. Mean FSIQ for the nonreferred (Control) sample was 101.73 (SD = 12.33). Mean TOF Total Problems was 12.20 (SD = 20.14).

Measures

Stanford-Binet Intelligence Scales, Fifth Edition. The SB5 is an individually administered intelligence scale for children and adults, with norms available for ages 2-80+. The SB5 is composed of 16 subtests, although not all are administered to any given examinee. From the subtest scores, five factor-analytically derived factor scores can be generated: Fluid Reasoning (FR), Knowledge (KN), Quantitative Reasoning (QR), Visual-Spatial Processing (VS), and Working Memory (WM). The SB5 also provides the more traditional Verbal IQ (VIQ), Nonverbal IQ (NVIQ), and Full-Scale IQ (FSIQ) scores. Factor and IQ scores have a mean of 100 and an SD of 15. The SB5 was normed

on 4,800 individuals in the United States. Individuals were excluded from the standardization sample if they had severe medical conditions, limited proficiency in English, severe sensory/communication deficits, or severe behavioral or emotional disturbance, or if they were enrolled in special education services for more than 50% of their school day. The standardization sample was 51% female and was representative of the U.S. population in terms of ethnicity (69.1% Caucasian, 12.2% African-American, 12.3% Hispanic, 3.8% Asian, 2.7% Other), geographic region, and years of education completed by parents or guardians (Roid, 2003a).

The SB5 was revised in 2003 and represents a fairly significant revision from earlier versions. An overview of the organization of the SB5 will be useful to readers not familiar with the Stanford-Binet scales or with this version of the Stanford-Binet test. Subtests are organized into two larger domains (Verbal and Nonverbal), each with five factors (FR, KN, QR, VS, and WM) each. This results in 10 subtest areas (e.g., Verbal Fluid Reasoning, Nonverbal Fluid Reasoning, Verbal Knowledge, Nonverbal Knowledge, etc.). In general, one subtest is administered for each subtest area. However, for some areas, the exact subtest administered may vary depending on the age and ability level of the child or adult being tested or more than one subtest might be administered (e.g., for Verbal Fluid Reasoning, an examinee might be administered the Early Reasoning, Verbal Absurdities, and/or Verbal Analogies subtests). See Appendix A for a graphic presentation of the SB5 organization. The SB5 has demonstrated strong reliability. Internal consistency for the Full Scale IQ across age groups was .98. Test-retest reliability with a median test-retest interval of 8 days for individuals ages 6-20 was

.93 for the Full Scale IQ. Interscorer agreement was examined for all SB5 items that involve the examiner making a subjective scoring judgment (e.g., Vocabulary items that are scored 0, 1, or 2). Across the entire SB5, median interscorer agreement for these items was .90. The SB5 also demonstrates good construct validity, including correlations of .84 between SB5 and WISC-III Full Scale IQ scores. Further details on the reliability and validity of the SB5 are available in the *Technical Manual* (Roid, 2003a).

Test Observation Form. The TOF is a standardized rating form for assessing behavior during cognitive or achievement testing of children between the ages of 2-18. It is a part of the Achenbach System of Empirically Based Assessment (ASEBA) and designed to be consistent with other ASEBA forms such as the Child Behavior Checklist, the Teacher's Report Form, and the Youth Self-Report for school-age children (Achenbach & Rescorla, 2001). The TOF provides space in which the examiner records a narrative of the child's test session behavior during testing. At the end of the test session, the examiner rates the child on 124 items describing specific behavior that a child may demonstrate during testing (e.g., asks for feedback on performance, difficulty understanding language, interrupts) plus one open-ended item (item 125). Items are rated on a 4-point scale: 0 = no occurrence; 1 = very slight or ambiguous occurrence; 2 = definite occurrence with mid to moderate intensity/frequency and less than 3 miniutes total duration; and 3 = definite occurrence with severe intensity, high frequency, or 3 or more minutes total duration. Explicit guidelines for rating the items are available in the TOF *Manual* (McConaughy & Achenbach, 2004).

The TOF has five factor analytically derived syndrome scales (Withdrawn/Depressed, Language/Thought Problems, Anxious, Oppositional, and Attention Problems), as well as overall Internalizing (problems from the Withdrawn/Depressed and Language/Thought Problems), Externalizing (problems from the Oppositional and Attention Problems), and Total Problems scales. The titles of the syndrome scales represent the types of behaviors comprised by the scales; however, it should be noted that the Anxious scale appears to capture behaviors related to test anxiety (rather than more generalized anxiety) and is not part of either the Internalizing nor Externalizing scales (McConaughy & Achenbach, 2004). These scales were derived from ratings of 3,400 clinically referred children who were administered the SB5, the WISC-III, or the Woodcock-Johnson III Tests of Cognitive Abilities (WJ III COG, Woodcock, McGrew & Mather, 2001). In addition, some TOF items that were analogous to items from other ASEBA scales that had been selected as consistent with the DSM-IV diagnostic criteria of ADHD by experts in child psychology were used to create the DSM-oriented Attention Deficit/Hyperactivity Problems (ADHP) scale and its two subscales, Inattentive and Hyperactivity-Impulsivity (see McConaughy & Achenbach, 2004, for a more detailed description of the creation of the TOF syndrome scales and DSM-oriented scales).

The TOF profile provides raw scores, T scores, and percentile scores for each scale. On the TOF syndrome scales and the ADHP scale, T scores between 65-69 (between the 93rd and 97th percentiles) are considered to be in the borderline clinical range, while T scores of 70 and above (above the 97th percentile) are considered in the

clinical range. On the Internalizing, Externalizing, and Total Problem scales, T scores of 60-63 (between the 84th and 90th percentiles) are considered to be in the borderline clinical range, while T scores of 64 and above (above the 90th percentile) are considered to be in the clinical range. Cut-off points for the borderline and clinical ranges on the Internalizing, Externalizing, and Total Problems scales are lower than for the syndrome scales and the ADHP scale because there are more items on these scales.

The TOF was standardized on 3,943 children (51% female) between the ages of 2 and 18. Of this sample, 2,442 children were part of the SB5 standardization sample, while 1,501 children were administered the SB5 but were not part of the SB5 standardization sample. Children were excluded from the TOF standardization sample if they qualified for special education due to behavioral, emotional, or developmental problems, if they had been referred for mental health services in the past 12 months, or if their Full Scale IQ on the SB5 was lower than 75. The TOF standardization sample was representative of the U.S. population in terms of ethnicity (64% Caucasian, 14% African-American, 14% Hispanic, 8% Mixed or Other), SES, and geographic region.

Test-retest reliability was established on a sample of 130 children tested by the same examiner over an average interval of 10 days. Test-retest reliability for all scales was good, ranging from .53 for Anxious to .87 for Attention Deficit/Hyperactivity Problems, with the mean test-retest *r* of .80 for all scales. Interrater reliability was established by having trained lay observers view test sessions for 43 children through a one-way mirror. Interrater reliability between observers and examiners for all scales except Anxious (r = .12) was moderate to good, ranging from .42 for Total Problems to

.78 for Externalizing. Mean interrater reliability for all scales was .62. Examiners' ratings of children were significantly higher than observers' ratings on Language/Thought Problems, Anxious, Internalizing, and Total Problems, suggesting that clinicians may be more sensitive to subtle difficulties such as test anxiety or language problems than lay observers (McConaughy & Achenbach, 2004). Because most of the TOF scales were empirically derived, internal consistency of the scales is good to excellent, ranging from .74 for Anxious to .95 for Total Problems. The mean Cronbach's alpha of all scales was .84.

CHAPTER 4: RESULTS

Analysis of Aim 1

To test group differences on the SB5, a series of 2 x 2 x 4 multivariate analyses of variance (MANOVA) were performed (SPSS General Linear Model, 2000). Gender, age group (6-11; 12-18), and diagnostic group (ADHD, LD, EBD, and Control) were treated as between-subject variables and sets of SB5 scales (VIQ and NVIQ; FR, KN, QR, VS, and WM) were treated as dependent variables. Univariate ANOVAs and Scheffé post-hoc tests were performed following each MANOVA to identify group differences on the dependent variables. In addition, a 2 x 2 x 4 univariate analysis of variance (ANOVA) was performed, with gender, age group, and diagnostic group treated as between-subject variables and the SB5 Full Scale IQ score treated as the dependent variable; Scheffé post-hoc analyses followed this univariate ANOVA. Effect sizes, as indicated by partial Eta^2 (which can be directly translated into percent of variance explained), were evaluated according to Cohen's (1988) criteria: effect sizes accounting for 1 to 5.8% of the variance are small; 5.9 to 13.7% of the variance are medium; and greater than 13.8% of the variance are large. Table 5 summarizes the results of these analyses, while Figure 2 shows the mean scores obtained by the 4 diagnostic groups on each SB5 scale.

SB5 VIQ and NVIQ. The overall MANOVA for VIQ and NVIQ showed significant main effects of diagnostic group, $F(6,792) = 12.88$, $p < .001$, but no significant effects of age group, gender, diagnostic group x age group, diagnostic group x gender, or three-way interaction. A univariate ANOVA showed significant effects of diagnostic group for both VIQ, $F(3, 396) = 26.19$, $p < .001$, and NVIQ, $F(3, 396) = 21.45$,

$p < .001$. Effect sizes were large for both VIQ and NVIQ (16.6% and 14.0% of the variance, respectively). The Scheffé pairwise post-hoc analyses indicated that for VIQ, the Control group scored significantly higher ($p < .05$) than all three exceptional groups (ADHD, LD, EBD), while the ADHD and EBD group scored significantly higher ($p < .05$) than the LD group. The ADHD and EBD groups did not differ significantly from each other. For NVIQ, the Control group scored significantly higher ($p < .05$) than the LD and EBD groups, but did not differ significantly from the ADHD group. The ADHD group scored significantly higher ($p < .05$) than the LD group on NVIQ.

SB5 FSIQ. The overall ANOVA showed significant main effects for diagnostic category, $F(3, 396) = 27.21$, $p < .001$, as well as a significant three-way interaction of diagnostic group x gender x age group, $F(3, 396) = 2.65$, $p < .05$. The main effect of diagnostic category was a large effect, accounting for 17.1% of variance. Subsequent pairwise post-hoc analyses indicated that the Control group scored significantly higher ($p < .05$) than the other three groups. In addition, the ADHD group scored significantly higher ($p < .05$) than the LD group, but was not significantly different from the EBD group. For the three-way interaction, post-hoc tests indicated that amongst males, older children scored significantly higher ($p < .05$) than younger children; however amongst females the opposite pattern was observed: younger children scored significantly higher ($p < .05$) than older children for FSIQ. Figure 3 is a plot of the cell means displaying the pattern of scores for older and younger male and female children on FSIQ.

SB5 Factor Scores. The overall MANOVA showed significant main effects of diagnostic group, $F(15, 1182) = 7.09$, $p < .001$, gender, $F(5, 392) = 3.81$, $p = .002$, and

diagnostic group x age group, $F(15, 1182) = 2.37$, $p = .002$. As shown in Table 5, subsequent univariate ANOVAs indicated significant medium to large effects for diagnostic category on all five SB5 factors (all $p < .001$, effect sizes ranging from 8.6% to 16.2%). Pairwise post-hoc analyses showed that the Control group scored significantly higher than the LD group on all 5 SB5 factors, higher than the EBD group on QR, VS and WM, and higher than the ADHD group on WM. The ADHD group scored significantly higher than the LD group on FR, KN, QR, and VS, but not WM. The EBD group scored significantly higher than the LD group only on QR (all $p < .05$).

For the main effect of gender, subsequent univariate analyses indicated a significant effect only for FR, $F(1, 396) = 5.17$, $p < .05$. Pairwise post-hoc tests indicated that females scored significantly higher ($p < .05$) than males on FR; however, this effect size was small (1.3 % of the variance). For the interaction of diagnostic category x age group, univariate ANOVAs indicated that this effect was significant only for KN, $F(3, 396) = 4.23$, $p = .006$. Subsequent univariate analyses and post-hoc tests indicated that within the ADHD group, older children scored significantly higher ($p < .05$) than younger children on KN. Within the Control group, the opposite pattern was observed: younger children scored significantly higher ($p < .05$) than older children on KN. Effect sizes for age group were large (15.3% of the variance) in the ADHD group but small (5.2% of the variance) in the Control group. Within the LD and EBD groups, there were no significant effects of age group on KN. These results are shown in Figure 4.

Analysis of Aim 2

To test group differences on the TOF, a series of 2 x 2 x 4 MANOVAs were performed, with gender, age group, and diagnostic group treated as between-subject variables and sets of TOF scales (Internalizing and Externalizing; Withdrawn/Depressed, Language/Thought Problems, Anxious, Oppositional, and Attention Problems; and DSM Inattentive and DSM Hyperactivity-Impulsivity) treated as dependent variables. Univariate ANOVAs and Scheffé post-hoc tests were performed following each MANOVA to identify group differences on the dependent variables. In addition, two 2 x 2 x 4 univariate ANOVAs were performed, with gender, age group, and diagnostic group treated as between-subject variables and the TOF Total Problems score and DSM ADHD Problems treated as dependent variables.

TOF Syndromes. The overall MANOVA indicated significant main effects of diagnostic group, $F(15,1134) = 8.83$, $p < .001$ and age group, $F(5,376) = 6.16$, $p < .001$, and significant interaction effects of diagnostic category x age group, $F(15,1134) = 2.65$, $p = .001$, diagnostic category x gender, $F(15,1134) = 1.83$, $p = .026$, age group x gender $F(5,376) = 3.71$, $p = .003$, and the three-way interaction of diagnostic category x age group x gender, $F(15,1134) = 1.74$, $p = .039$.

For the main effect of diagnostic group, univariate ANOVAs indicated that this effect was significant for all five TOF syndromes (all $p < .001$). Post-hoc analyses indicated that the EBD group scored significantly higher (all $p < .05$) than the ADHD, LD, and Control groups for the Withdrawn/Depressed, Language/Thought Problems, Anxious, and Oppositional syndromes. On the Attention Problems syndrome, the EBD

group scored significantly higher ($p < .05$) than the LD and Control groups, but was not significantly different from the ADHD group. The ADHD group scored significantly higher ($p < .05$) than the LD group on the Anxious, Oppositional, and Attention Problems syndromes, and significantly higher ($p < .05$) than the Control group on those three syndromes (Anxious, Oppositional, and Attention Problems) as well as the Language/Thought Problems syndrome. Table 6 summarizes the results of these analyses and Figure 5 shows the mean scores obtained by the four diagnostic groups on each TOF syndrome.

For the main effect of age group, univariate ANOVAs indicated significant effects for the Oppositional, $F(1,380) = 5.82$, $p = .016$, and Attention Problems, $F(1,380) = 19.33$, $p < .001$, syndromes. Younger children scored significantly higher than older children on both the Oppositional syndrome ($p < .05$) and the Attention Problems syndrome ($p < .001$).

The interaction effect of diagnostic group x age group was significant for the Withdrawn/Depressed and Attention Problems syndrome. The cell means for older and younger children on Withdrawn/Depressed are shown in Figure 6. Amongst both younger and older children, the EBD group scored significantly higher (all $p < .05$) on the Withdrawn/Depressed syndrome than the other three groups; however, the effect size was small (3.0 % of the variance) for younger children and large (30.1% of the variance) for older children. The cell means for older and younger children on Attention Problems are shown in Figure 7. Amongst younger children, the EBD group and the ADHD group scored significantly higher (all $p < .001$) than the LD and Control groups on the Attention

Problems syndrome, but the EBD and ADHD groups did not differ significantly from each other. For older children, the ADHD group scored significantly higher (all $p < .05$) than the EBD, LD, and Control groups.

Univariate ANOVAs indicated that the interaction effect of diagnostic group x gender was significant for the Withdrawn/Depressed syndrome, $F(3,380) = 4.39$, $p = .005$. For males, the EBD group scored significantly higher (all $p < .001$) than the ADHD, LD, and Control groups; however, no significant differences were found amongst the diagnostic groups for females. Figure 8 displays the cell means for males and females on the Withdrawn/Depressed syndrome.

The interaction effect of age group x gender was significant for the Language/Thought Problems, $F(1,380) = 5.27$, $p = .014$, and Oppositional, $F(1,380) = 3.89$, $p = .049$, syndromes. However, these were very small effects (2.2 and 1.0% of variance, respectively) and post hoc tests failed to show any significant gender differences within the two age groups on either syndrome.

The three-way interaction of diagnostic category x age group x gender was significant for Oppositional, $F(3,380) = 2.63$, $p = .05$; this was also a small effect (2.0 % of the variance). Younger males in the EBD group scored significantly higher than all other children in the EBD group and in all three diagnostic groups (all $p < .05$). Younger males in the ADHD group scored significantly higher than all other children in the ADHD group and in the LD and Control groups (all $p < .05$). Figure 9 is a plot of the cell means displaying the pattern of scores for older and younger male and female children on the Oppositional syndrome.

TOF Internalizing and TOF Externalizing. The overall MANOVA showed significant main effects of diagnostic group, $F(6,768) = 16.62$, $p < .001$, and age group, $F(6,383) = 12.89$, $p < .001$, and interaction effects of diagnostic group x age group, $F(6,768) = 5.02$, $p < .001$, and diagnostic group x gender, $F(6,768) = 2.15$, $p = .046$. Subsequent univariate ANOVAs showed significant main effects of diagnostic group for both TOF Internalizing, $F(3,394) = 12.18$, $p < .001$, and TOF Externalizing, $F(3,384) = 27.62$, $p < .001$, but significantly main effects of age group only for TOF Externalizing, $F(1,384) = 19.93$, $p< .001$. For the main effect of diagnostic group, pairwise post-hoc analyses indicated that the EBD group scored significantly higher than the other three diagnostic groups (ADHD, LD, and Control) on TOF Internalizing (all $p < .001$). Both the EBD and the ADHD group scored significantly higher (all $p < .001$) than the LD and Control groups for TOF Externalizing; however, the EBD and ADHD groups did not differ significantly from each other. For the main effect of age group, pairwise post-hoc analyses indicated that younger children scored significantly higher than older children on TOF Externalizing ($p < .001$). Figure 10 shows the mean raw scores of the four diagnostic groups on TOF Internalizing, TOF Externalizing, and TOF Total Problems.

Subsequent univariate ANOVAs examining the interaction effect of diagnostic group x age group indicated that this effect was significantly only for TOF Externalizing, $F(3,384) = 2.15$, $p = .004$. Post-hoc analyses showed that amongst younger children, the EBD group scored significantly higher (all $p < .05$) than the ADHD, LD, and Control groups, while the ADHD group scored significantly higher (all $p < .001$) than the LD and Control groups. Amongst older children, the EBD and ADHD groups scored significantly

higher (all $p < .05$) than the LD and Control groups, but the EBD and ADHD groups did not differ significantly from each other. This pattern of scores is displayed in Figure 11, which plots the cell means for older and younger children with ADHD, LD, and EBD and Control children on TOF Externalizing.

Univariate ANOVAs examining the interaction effect of diagnostic group x gender indicated that this effect was significant only for TOF Internalizing, $F(3,384) = 4.00$, $p = .008$. Post-hoc analyses revealed that males in the EBD group scored significantly higher (all $p < .001$) than males in the other three diagnostic groups; this was a large effect (16.5% of the variance). However, none of the diagnostic groups differed significantly from the other groups for females. This pattern of scores is displayed in Figure 12.

TOF Total Problems. The overall ANOVA indicated significant main effects of diagnostic group, $F(3,396) = 27.51$, $p < .001$, and age group ($F(1,396) = 5.24$, $p = .023$). For the main effect of diagnostic group, the EBD group scored significantly higher than the ADHD, LD, and Control groups, while the ADHD group scored significantly higher than the LD and Control groups (all $p < .001$). As shown in Table 6, this was a large effect (17.2% of the variance). The mean raw scores obtained by the 4 diagnostic groups on TOF Total Problems is shown in Figure 10. For the main effect of age group, younger children scored significantly higher than older children ($p < .05$); however, this was a small effect (1.3% of the variance).

DSM-oriented Inattentive and Hyperactivity-Impulsivity Subscales. The overall MANOVA indicated significant main effects of diagnostic group, $F (6,786) = , p < .001$,

and age group, F (2,392) = , $p < .001$, and a significant interaction effect of diagnostic group x age group, F (6,786) = , $p = .004$. The main effect of diagnostic group was significant for both the DSM-oriented Inattentive subscale, F(3,393) = 27.29, $p < .001$, and Hyperactivity-Impulsivity subscale, F(3,393) = 30.41, $p < .001$. These effects were large (17.2% and 18.8% of the variance, respectively). Subsequent pairwise analyses indicated that the ADHD and EBD groups scored significantly higher than the LD and Control groups on both the DSM-oriented Inattentive and Hyperactivity-Impulsivity subscales (all $p < .001$). Figure 13 displays the mean raw scores obtained by the 4 diagnostic groups on the DSM-oriented Inattentive and Hyperactivity-Impulsivity subscales.

The main effect of age group was also significant for both the DSM-oriented Inattentive subscale, F(1,393) = 17.85, $p < .001$, and Hyperactivity-Impulsivity subscale, F(1,393) = 22.23, $p < .001$, but these effects were small (4.3% and 5.4% of the variance, respectively). Pairwise post-hoc analyses indicated that younger children scored significantly higher than older children on both DSM-oriented subscales (all $p < .001$).

The interaction effect of diagnostic group x age group was also significant for both the DSM-oriented Inattentive subscale, F(3,393) = 4.32, $p = .005$, and Hyperactivity-Impulsivity subscale F(3,393) = 4.74, $p = .003$. Again, effect sizes were small (3.2% and 3.5% of the variance, respectively). Amongst younger children, the EBD and ADHD groups scored significantly higher than the LD and Control groups (all $p <$ .001) on the DSM-oriented Inattentive subscale, but the EBD and ADHD groups did not differ significantly from each other. In addition, the LD group scored significantly higher

(p = .031) than the Control group. Amongst older children, the ADHD group scored significantly higher than the LD, EBD, and Control groups (all p < .05) on the DSM-oriented Inattentive subscale, while the EBD group scored significantly higher than the LD and Control groups (all p < .05) on the DSM-oriented Inattentive subscale. On the DSM-oriented Hyperactivity-Impulsivity scale, younger children in the ADHD and EBD groups scored significantly higher than the LD and Control groups (all p < .001). Amongst older children, the ADHD group scored significantly higher than the other three diagnostic groups (all p < .05). Figures 14 and 15 show these patterns of scores.

DSM-oriented Attention Deficit/Hyperactivity Problems (ADHP) Total. The overall ANOVA indicated significant main effects of diagnostic group, $F(3,393) = 34.53$, p < .001, age group, $F(1,393) = 24.05$, p < .001, and age group x diagnostic group, $F(3,393) = 5.18$, p = .002. For the main effect of diagnostic group, the EBD group scored significantly higher (all p < .001) than the ADHD, LD, and Control groups on the ADHP Total score, while the ADHD group scored significantly higher (all p < .001) than the LD and Control groups. In addition, the LD group scored significantly higher (p = .044) than the Control group. As shown in Table 6, the main effect of diagnostic group was a large effect (20.9% of the variance). Mean raw scores for all four diagnostic groups on the DSM-oriented ADHP Total score can be seen in Figure 13. For the main effect of age group, younger children scored significantly higher than older children (p < .001), which was a small effect (5.8% of the variance).

For the interaction of age group x diagnostic group, amongst younger children, the ADHD group and the EBD group scored significantly higher (all p < .001) on the

ADHP Total score than the LD and Control groups, but the ADHD and EBD group did not differ significantly from each other. Amongst older children, the ADHD group scored significantly higher than the LD, EBD, and Control groups and the EBD group scored significantly higher than the LD and Control groups (all $p < .05$). This was a small effect (3.8% of the variance). Figure 16 is a plot of the cell means displaying the pattern of scores for older and younger children with ADHD, LD, and EBD and Control children on the ADHP Total score.

Analysis of Aim 3

A series of forward stepwise logistic regressions were performed to test the contribution of the SB5 scales, TOF syndromes, and combinations of SB5 and TOF scores to discriminating between the diagnostic groups (ADHD, LD, and EBD) and the control group as well as amongst the diagnostic groups. In each logistic regression, the relevant predictors were entered as a block allowing the program to add variables to the model according to Rao's efficient score statistic. Table 7 presents the significant predictors, sensitivity (percent of the ADHD group correctly classified by the predictors), specificity (percent of the Control group correctly classified), false positives (percent of the ADHD group incorrectly classified), false negatives (percent of the Control group incorrectly classified), overall correct classification rates, and Nagelkerke R^2 for the ADHD group versus the Control group. Nagelkerke R^2 ranges from 0 to 1 and serves a measure of the strength of association between the predictors and the categorical outcome variable (ADHD versus Control). It can be interpreted as an estimate of variance

accounted for by the model. Tables 8 and 9 present equivalent information for the LD versus the Control group and the EBD versus the Control group, respectively. Table 10 presents information equivalent information for predictors that significantly discriminated amongst the diagnostic groups (ADHD versus LD, ADHD versus EBD, and LD versus EBD).

Children with ADHD versus Control children. Three separate forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, and SB5 FSIQ to discriminating between children with ADHD and Control children (see top part of Table 7). Of the five SB5 factors, WM and QR together correctly classified 42.0% of the ADHD group and 89.0% of the Control group, with an overall correct classification rate of 73.3%. The model including these two factors produced an estimated 28.2% of the variance in discriminating the ADHD group versus the Control group. For the two SB5 composite scores, VIQ emerged as specific but not sensitive predictor, correctly classifying only 10% of the ADHD group but 96% of the Control group. VIQ produced only an estimated 5.1% of the variance in discriminating between the ADHD group and the Control group. FSIQ also emerged as specific but not sensitive predictor, correctly classifying 10% of the ADHD group and 99% of the Control group. FSIQ produced only an estimated 4.2% of variance.

Five separate forward stepwise logistic regressions were conducted to determine the contribution of the TOF syndromes, the TOF INT and TOF EXT scales, the TOF TOT scale, the DSM-IN and DSM-HI scales, and the DSM-ADHP scale to discriminating between children with ADHD and Control children. Of the five TOF

syndromes, Attention Problems emerged as the only significant predictor, correctly classifying 53.2% of the ADHD group and 95% of the Control group, with an overall correct classification rate of 81.6%. TOF Attention Problems produced an estimated 40.9% of the variance in discriminating the ADHD versus Control groups. When considering the DSM-oriented subscales, DSM-IN correctly classified 56.0% of the ADHD group and 92.0% for the Control group, with an overall correct classification rate of 80.0%. DSM-IN produced an estimated 47.2% of the variance in discriminating the ADHD versus Control groups. Among the two TOF broad scales, TOF EXT score emerged as a significant predictor, correctly classifying 54.2% of the ADHD group and 96.0% of the Control group, with an overall correct classification rate of 82.4%. TOF EXT produced an estimated 40.6% of the variance in discriminating the two groups. The TOF TOT score correctly classified only 36% of ADHD group but 94.0% of the Control group, with an overall correct classification rate of 74.7%. TOF TOT produced an estimated 22.3% of the variance in discriminating between the two groups.

An additional forward stepwise logistic regression was conducted entering SB5 WM, SB5 QR, and TOF Attention Problems as predictors of ADHD versus Control groups (see bottom of Table 7). The model with all three predictors correctly classified 61.2% of the ADHD group and 92.0% of the Control group, with an overall correct classification rate of 81.9%. The three-predictor model produced an estimated 53.4% of the variance. A separate forward stepwise logistic regression was conducted entering SB5 WM, SB5 QR, and TOF DSM-IN as predictors of ADHD versus Control groups. The model with these three predictors correctly classified 58.0% of the ADHD group and

92.0% of the Control group, with an overall correct classification rate of 80.7%. This model produced an estimated 56.1% of the variance. An additional forward stepwise logistic regression was conducted entering SB5 FSIQ and TOF TOF as predictors of ADHD versus Control groups. SB5 FSIQ and TOF TOT correctly classified 40.0% of the ADHD group and 94.0% of the Control group, with an overall correct classification rate of 76.0%. This model produced an estimated 26.3% of the variance.

Children with LD versus Control children. Three separate forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, and SB5 FSIQ to discriminating between children with LD and Control children (see top part of Table 8). Of the five SB5 factors, WM and QR correctly classified 91.5% of the LD group and 45.0 % for the Control group, with an overall correct classification rate of 77.5%. WM and QR produced an estimated 32.0% of the variance in discriminating between the LD group and the Control group. When VIQ and NVIQ were entered into a separate forward stepwise logistic regression, VIQ emerged as a significant predictor, correctly classifying 91.5% of the LD group and 42.0% of the Control group, with a 76.6% overall correct classification rate. VIQ produced an estimated 28.6% of variance. FSIQ correctly classified 91.9% of the LD group and 40.0% of the Control group, with an overall correct classification rate of 76.3%. FSIQ produced an estimated 28.7% of the variance in discriminating the LD versus Control groups.

When the five TOF syndromes were entered into a forward stepwise logistic regression to determine which syndromes were predictors of group status, Anxious and Attention Problems emerged as significant predictors (see middle of Table 8). Although

this model correctly classified 99.1% of the LD group, it correctly classified only 4.0% of the Control group, with an overall correct classification of 69.7%. These two predictors produced only an estimated 6.5% of the variance in discriminating between the LD group and the Control group.

When SB5 WM, SB5 QR, TOF Anxious, and TOF Attention Problems were entered into a forward stepwise logistic regression, only a three-predictor model significantly predicted group status (see bottom of Table 8). The combination of SB5 WM, SB5 QR, and TOF Attention Problems correctly classified 90.9% of the LD group and 49.0% of the Control group, with an overall correct classification rate of 78.2%. These predictors produced estimated 34.1% of the variance in discriminating between the LD group and the Control group.

Children with EBD versus Control children. Three separate forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, and SB5 FSIQ to discriminating between children with EBD and Control children (see top part of Table 9).Of the five SB5 factors, only WM emerged as a significant predictor. WM correctly classified 17.9% of the EBD group and 97.0% of the Control group, with an overall correct classification rate of 79.7%. WM produced an estimated 19.2% of the variance in discriminating between the EBD and Control groups. When VIQ and NVIQ were entered into a forward stepwise logistic regression, VIQ correctly classified only 10.7% of the EBD group, in contrast to 96.0% of the Control group, with an overall correct classification rate of 77.3%. VIQ produced an estimated 13.5% of the variance in discriminating between EBD and Control groups. Similarly,

FSIQ correctly classified only 7.1% of the EBD group but 98.0% of the Control group, with an overall correct classification rate of 78.1%. FSIQ produced an estimated 14.4% of the variance between the two groups.

Of the five TOF syndromes, Language/Thought Problems and Oppositional (see middle of Table 9) correctly classified 51.9% of the EBD group and 99.0% of the Control group, with an overall correct classification rate of 89.0%. The model with these two predictors produced an estimated 48.3% of the variance in discriminating the EBD group versus the Control group. The TOF INT and TOF EXT scores correctly classified 57.1% of the EBD group and 96.0% of the Control group, with an overall correct classification rate of 87.5%. TOF INT and EXT together produced an estimated 52.2% of the variance. TOF TOT correctly classified 46.4% of the EBD group and 96.0% of the Control group, with an overall correct classification rate of 85.2%. TOF TOT produced an estimated 44.9% of the variance in discriminating between the EBD group and the Control group.

When SB5 WM, TOF Language/Thought Problems, and TOF Oppositional were entered into a forward stepwise logistic regression, WM did not contribute significantly to the model. The resulting two-predictor model of TOF Language/Thought Problems and TOF Oppositional produced an estimated 50.1% of the variance in discriminating between the EBD versus Control groups, similar to the estimate of 48.3% of variance with these predictors alone. An additional forward stepwise logistic regression was conducted entering SB5 FSIQ and TOF TOT as predictors of EBD versus Control group. The model with these two predictors correctly classified 50.0% of the EBD group and

95.0% of the Control group, with an overall correct classification rate of 85.2%. This model produced an estimated 50.9% of the variance.

Children with ADHD, children with LD, and children with EBD. Three forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, and SB5 FSIQ to predicting ADHD versus LD status (see top part of Table 10). Of the SB5 factors, QR correctly classified 10.0% of the ADHD group and 97.9% of the LD group, resulting in an overall correct classification rate of 82.4%. QR produced an estimated 13.9% of the variance in discriminating between the ADHD group and the LD group. VIQ, NVIQ, and FSIQ did not emerge as significant predictors of group status. Another three forward stepwise logistic regressions were conducted to determine the contribution of the five TOF syndromes, the TOF INT and TOF EXT scales, and the TOF TOT scales to predicting ADHD versus LD status. A two-predictor model of Attention Problems and Withdrawn/Depressed correctly classified 23.4% of the ADHD group and 97.8% of the LD group, resulting in an overall correct classification rate of 84.8%. This model produced an estimated 24.3% of the variance in discriminating the ADHD versus LD groups. As the one-predictor model of SB5 QR and the two-predictor model of TOF Attention Problems and TOF Withdrawn/Depressed accounted for significant percentages of the variance in discriminating between the ADHD versus LD groups, a separate forward stepwise logistic regression was conducted to determine which of these three variables (SB5 QR, TOF Attention Problems, and TOF Withdrawn/Depressed) were significant predictors of group status. QR and Attention Problems correctly classified 32.7% of the ADHD group

and 96.5% of the LD group, with an overall correct classification rate of 85.3%. The two-predictor model produced an estimated 33.7% of the variance in discriminating the ADHD versus LD groups.

Three forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, and SB5 FSIQ to predicting ADHD versus EBD status (see middle part of Table 10). Of the SB5 factors, a two-predictor model of WM and QR correctly classified 42.0% of the ADHD group and 89.0% of the EBD group, with an overall correct classification rate of 73.3%. This model produced an estimated 28.2% of the variance in discriminating between the ADHD versus EBD groups. SB5 VIQ, NVIQ, and FSIQ did not emerge as significant predictors of group status. Three forward stepwise logistic regressions were conducted to determine the contribution of the five TOF syndromes, the TOF INT and TOF EXT scales, and the TOF TOT scales to discriminating between the ADHD and EBD groups. TOF Withdrawn/Depressed correctly classified 93.6% of the ADHD group and 40.7% of the EBD group, with an overall correct classification rate of 74.3%. TOF Withdrawn/Depressed produced an estimated 33.6% of the variance in discriminating the ADHD group from the EBD group.

Finally, six forward stepwise logistic regressions were conducted to determine the contribution of the five SB5 factors, SB5 VIQ and NVIQ, SB5 FSIQ, the five TOF syndromes, the TOF INT and TOF EXT scales, and the TOF TOT scales to discriminating between the LD and EBD groups (see bottom part of Table 10). Only the two-predictor model of TOF Oppositional and TOF Anxious emerged as a significant

predictor of group status, correctly classifying 98.2% of the LD group, 29.6% of the EBD group, and 90.8% overall. TOF Oppositional and TOF Anxious produced an estimated 28.1% of the variance in discriminating between the LD group and the EBD group.

CHAPTER 5: DISCUSSION

Intelligence tests are often routine components of assessment for children referred to school psychologists and mental health clinics due to suspected ADHD, LD, or EBD (Achenbach, 2005; Mayes et al., 1998a). The administration of intelligence tests during child assessment offers at least two potentially useful sources of information. First, the test scores themselves are thought to be useful in generating hypotheses regarding the challenges that the child is experiencing (Kaufman, 1994). Second, the standardized environment provided by an individually administered cognitive test offers an ideal setting for observing the child's behavior, which many clinicians consider an essential component of a thorough assessment (Edwards, 2005; McConaughy, 2005). Behavioral observations obtained in a reliable manner, such as through the use of standardized observation procedures, can provide information regarding the validity of the test scores that are obtained, or the *intrasession validity* of the test scores (Oakland et al., 2000). In addition, reliable behavioral observations may reveal more enduring characteristics or behaviors that the child is likely to display in other settings, which represents the *exosession validity* of the test scores (Glutting et al., 1996).

This study differed from previous research on test scores and test session behavior in exceptional populations in several ways. First, this is the first independent research to specifically examine the performance of children with ADHD, LD, and EBD on the SB5. The Stanford-Binet Intelligence Scales were revised in 2003, and one of the stated goals of this revision was to create a test that was more sensitive to the cognitive deficits thought to underlie attention and learning problems. Existing research related to test

scores in children has focused on measures of intellectual functioning that were not specifically designed to take into account the strengths and weaknesses of children in these populations. Additionally, much of the available research has focused on tests that are outdated (e.g., the WISC-R). Each year since its revision in 2003, the SB5 has been included in an increasing number of research studies; however, none of the research to date has focused specifically on SB5 scores in children with attention, learning, and/or emotional problems. Thus, this study advances previous research on test scores by examining the scores of children with psychoeducational difficulties on a current, widely-used intelligence test that was designed to be sensitive to the deficits they are experiencing.

Second, this study used a comprehensive, standardized rating form that was co-normed with the SB5 (the TOF) to assess test session behavior in children with ADHD, LD, and EBD. Although behavior observations during testing are considered a crucial element of a comprehensive child assessment, research on the behavior of children during testing or on the diagnostic utility of this information is limited. Research that has examined the utility of test session behavior assessed by standardized test behavior rating scales suggests that this is a promising area of study (e.g., Glutting et al., 1997; McConaughy & Achenbach, 2004).

Finally, this study diverges from previous research by examining the associations *between* test scores and test session behavior in children with ADHD, LD, and EBD. At present, there is very little research exploring the associations between IQ scores and test session behavior in children from exceptional populations, and no known research exists

which utilizes test scores and test session behavior together to predict diagnostic group status. This study therefore represents the first investigation into whether combinations of intelligence test scores and test session behavior can meaningfully predict diagnostic group status in children with ADHD, LD, and EBD versus normal controls.

Summary of SB5 Findings

The first aim of this study was to compare the patterns of IQ scores of children with ADHD, LD, and EBD, and normal controls on the SB5. The results indicated that, as anticipated, children in the Control group scored significantly higher than children in the LD group on all five SB5 factors, as well as the overall VIQ, NVIQ, and FSIQ scales. As can be seen in Table 3, children with LD scored approximately three-fifths to four-fifths of a standard deviation (that is, about 9-12 points) lower than Control children on all SB5 scales, which is generally consistent with previous research demonstrating that children with LD obtain mean IQ scores approximately one standard deviation below the mean of normal controls (e.g., Canivez, 1996; Doll & Boren, 1993; Lavin, 1996; Prewett & Matavich, 1993).

It was hypothesized that children in the Control group would also score significantly higher than children in the EBD groups on all domains of the SB5, but this hypothesis received only partial support. Children in the Control group scored significantly higher than children in the EBD group on the QR, VS, and WM factors, as well as the VIQ, NVIQ, and FSIQ scales. As shown in Table 3, children in the EBD group scored approximately one-half to two-thirds of a standard deviation (that is, about

7-10 points) below the Control group on these scales, which is consistent with previous research demonstrating that children with EBD score approximately one-half to one standard deviations below the mean on IQ measures (e.g., Connery et al., 1996; Javorsky, 1993; Slate & Jones, 1995). There were no significant differences between the Control group and the EBD group on the FR and KN factors.

While it was also hypothesized that children in the Control group would score higher than children in the ADHD group on all of the SB5 scales, this hypothesis received limited support. Control children scored significantly higher than children in the ADHD group on the WM factor, the VIQ scale, and the FSIQ. Children with ADHD scored approximately one-third standard deviations (that is, about 5 points) below children in the Control group on the VIQ and FSIQ, and approximately two-thirds standard deviations (about 10 points) below Control children on WM, as shown in Table 3. Again, this is generally consistent with previous literature documenting that children with ADHD are found to have mean IQ scores one-third to one-half of a standard deviation below the mean (e.g., Barkley et al., 1990; Farone et al., 1998; Saklofske et al., 1994). There were no significant differences between children with ADHD and children in the Control group on NVIQ or the other four SB5 factors.

That children from both the ADHD and LD groups scored significantly lower than children in the Control group on WM is consistent with one of the test publisher's stated rationales for revising the SB5, as the revised test was designed to be sensitive to putative deficits displayed by children with these disorders (Riverside Publishing Company, 2004). It was not anticipated that children with EBD would also score lower

than control children on the WM factor, as previous research has not found that children with EBD demonstrate working memory problems (Carter et al., 1990; Naglieri et al., 2003). However, given that problems concentrating are known symptoms of several emotional disorders (e.g., Major Depressive Disorder and Generalized Anxiety Disorder; American Psychiatric Association, 2000), this finding may reflect the greater sensitivity of the WM factor on the SB5 to the concentration difficulties demonstrated by children with EBD as compared to measures of working memory used by previous researchers. One difference between the SB5 WM factor and other measures of working memory is the inclusion of nonverbal working memory tasks in addition to verbal working memory tasks, which may be important, as research using both verbal and nonverbal tasks has suggested that some groups of children may show greater deficits in nonverbal working memory than in verbal working memory (McInnes et al., 2003). If future research replicates the finding that Control children score significantly higher than children with ADHD, LD, and EBD on the SB5 WM scale, this could have important implications for future research on the cognitive correlates of emotional, behavioral, and learning problems, as well as research on the similarities between these disorders.

When comparing SB5 scores amongst the three diagnostic groups, the results of this study indicated that children in the ADHD group scored significantly higher than children in the LD group on the FR, KN, QR, and VS factors, but not on the WM factor. This pattern suggests that while children with ADHD obtained significantly higher mean scores than children with LD on most SB5 factors, weaknesses in working memory are common to both attention problems and learning disabilities. Recent research supports

this position (e.g., Gathercole et al., 2006). Children with ADHD in the present investigation also scored higher than children with LD on the VIQ, NVIQ, and FSIQ scales. These results are consistent with robust body of previous research (described above) documenting the greater discrepancy on cognitive measures between children with LD and Control children, as compared to the discrepancy observed between children with ADHD and Control children.

Children with ADHD did not score higher than children with EBD on any SB5 scale. This finding is slightly unexpected, considering that, as already noted, previous research has demonstrated that children with EBD typically score one-half to one-standard deviation below the mean, while children with ADHD score one-third to one-half standard deviations below the mean of normal controls (Barkley et al., 1996; Farone et al., 1998; Javorsky, 1993; Saklofske et al., 1994; Slate & Jones, 1995). However, children with EBD scored higher than children with LD on the QR factor and the VIQ scale. As already discussed, children in the ADHD and Control groups also scored significantly higher than children in the LD groups on both QR and VIQ. These particular SB5 factors reflect cognitive processes that would intuitively be anticipated to be particular weaknesses in children with LD, with individuals with an LD in language-based subjects such as reading and/or writing demonstrating verbal reasoning weaknesses, and individuals with an LD in math demonstrating quantitative reasoning weaknesses when compared to children without learning problems.

As SB5 scores are normed by age, few findings that were specific to a particular age group were anticipated. It was not known whether male and female children in the

diagnostic groups would demonstrate different patterns of test scores on the SB5, although the limited available research suggested that significant differences could emerge (Preiss & Leska, 2006). Results specific to a particular age group and/or gender were found only for the FR and KN factors and the FSIQ scale. Females scored slightly higher than males on the FR factor. For KN, older children with ADHD scored higher than younger children with ADHD, and older children in the Control group scored higher than younger children. For FSIQ, older males scored significantly higher than younger males, whereas the opposite pattern was observed in females.

Summary of TOF Findings

The second aim of this study was to compare patterns of test session behavior as assessed by the TOF amongst children with ADHD, LD, and EBD and Control children. As can be seen from Table 4, children in the LD and Control groups displayed few problem behaviors during testing (e.g., mean TOF Total Problems scores of 8.61 and 6.97, respectively), while children in the ADHD group and EBD groups demonstrated many more problem behaviors during testing (e.g., mean TOF Total Problems scores of 23.16 and 41.32, respectively). Close examination of Table 4 also reveals that children with EBD obtained significantly higher scores than the LD and Control groups on all 11 TOF syndromes and scales, and higher mean scores than children in the ADHD group on 7 of 11 of the TOF syndromes and scales. There were no significant differences between the EBD and ADHD groups on the TOF Attention Problems, Externalizing, DSM-oriented Inattentive, or DSM-oriented Hyperactivity-Impulsivity scales.

The finding that the children in the EBD group scored higher than children in the ADHD, LD, and Control groups on most of the TOF scales was unexpected. It was hypothesized that children with EBD would score higher than children in the other groups on the Oppositional and Withdrawn/Depressed syndromes, as behaviors captured by these syndromes are clearly related to emotional and behavioral difficulties. However, as there is very little current research available regarding the test session behavior of children with EBD, it was difficult to formulate hypotheses regarding the Anxious and Language/Thought Problems scales. The findings that children with EBD demonstrated more test anxiety and more language/thought difficulties on average than children in the ADHD, LD, or Control groups is therefore a novel finding that may have important implications for clinical practice and future research.

It was also surprising that children in the EBD group scored higher than children in the ADHD group on the DSM-oriented ADHP scale, and were not significantly different from children in the ADHD group on the Attention Problems syndrome or DSM-oriented Inattentive or Hyperactivity-Impulsivity scales. There are a number of possible explanations for these results. The most likely explanation is that the ADHP scale contains more items than the Attention Problems syndrome or the other DSM-oriented scales. As such, there is a greater range of variability on the ADHP scale, which provides increased power for finding significant group differences. Other possible interpretations of these results relate to potential limitations of this study (which are described in more detail later in this chapter). One of these possible explanations is that children in the EBD group may have been experiencing comorbid attention problems.

Other possible explanations are related to the small size of the EBD group and the limited information available regarding the children in the EBD group, such as what specific emotional/behavioral problems they were experiencing and whether or not they were receiving treatment for their difficulties. More specifically, the results discussed above may have been driven by small numbers of young children and male children with extremely high levels of problem behavior during testing. Within the EBD group, younger children scored higher than older children on the Withdrawn/Depressed and Attention Problems syndromes and the TOF Externalizing scale, as well as the DSM-oriented Inattentive, Hyperactivity-Impulsivity, and ADHP scales. In addition, male children in the EBD group scored higher than female children on the Withdrawn/Depressed syndrome and TOF Internalizing scale, while young male children with EBD scored higher than all other groups of children on the Oppositional scale. While these specific findings are interesting, they may be artifacts of a small and possibly very heterogeneous sample of children with EBD, and replication of these results would be warranted before possible interpretations are offered.

The finding that children with EBD scored much higher than children without EBD on most of the TOF syndromes and scales is consistent with previous research demonstrating that children receiving special education services for EBD scored higher on most of the scales on the Child Behavior Checklist and Teacher Report Form than children receiving special education services for LD (McConaughy, Mattison, and Peterson, 1994). Nevertheless, it was somewhat unexpected that the high degree of emotional and behavioral problems reported by parents and teachers of children receiving

services for EBD would also be apparent during a relatively brief individual testing session. One compelling explanation for these findings is that children who receive special education services related to emotional and/or behavioral problems may be experiencing an exceptionally high level of emotional and behavioral dysregulation. Children who receive education services must demonstrate a need for services; that is, their condition(s) must impair their ability to benefit from traditional classroom instruction. Thus, it would stand to reason that children whose emotional/behavioral challenges are sufficient to substantially impact their academic functioning would demonstrate a very high number of problem behaviors in the classroom as well as in classroom-like settings, such as during cognitive testing. Children who meet criteria for special education services under the category of EBD might represent a subgroup of children with emotional and behavioral problems, with the subgroup demonstrating many more problem behaviors in academic environments than the larger group of children with emotional/behavioral disorders would be expected to display. If this were shown to be the case, the finding that children receiving services for EBD display higher levels of problem behavior during testing than other referred and nonreferred children who do not receive EBD services would not be surprising. This will be an important direction for future research, as these results hint that research on children who receive special education services for EBD could prove to have limited external validity for children with emotional and/or behavioral problems who do not qualify for school services.

Although the most salient findings on the TOF relate to the children in the EBD group, the results of this study yielded some interesting patterns of test results for

children in the ADHD and LD groups as well. As anticipated, children in the ADHD group scored significantly higher than children in the LD and Control groups on the Attention Problems syndrome and the DSM-oriented Inattentive, Hyperactivity-Impulsivity, and ADHP scales. However, children with ADHD also scored significantly higher than children in the LD and Control groups on the Anxious and Oppositional syndromes, as well as the TOF Externalizing and TOF Total Problems scales. Children with ADHD also scored higher than children in the Control group on the Language/Thought Problems scale. While the high degree of comorbidity between ADHD and behavior problems is a likely explanation for the higher scores on the Oppositional syndrome, it is not clear why children with ADHD would demonstrate more test anxiety or language/thought problems than other children, and this may be a fruitful area for future research.

Interestingly, children with LD did not differ significantly from Control children on any TOF syndrome or scale except the DSM-oriented ADHP scale. Furthermore, the finding that children in the LD group scored significantly higher than children in the Control group on the DSM-oriented ADHP scale may have been significant by chance (Sakoda, Cohen, & Beall, 1954). Although it was hypothesized that children in the LD group would demonstrate more language/thought problems during testing than children in the other groups, this hypothesis was not supported. Thus, although children with LD scored significantly lower than Control children on the SB5 factor measuring their verbal reasoning ability, this discrepancy did not translate into increased problems with language formulation that were observable to examiners completing the TOF.

The results indicating that children with LD demonstrated essentially the same level of problem behaviors during testing as Control children is particularly noteworthy in light of the discrepancy between these two groups in their scores on the SB5. Research on intrasession validity has been plagued by difficulties determining whether the negative correlations observed between test session behavior and IQ scores reflects: (1) the impact of behavior problems during testing on test scores, (2) the impact of cognitive deficits on test session behavior, or (3) a common factor that contributes to both lower test scores and increased problem behaviors. The results of this study suggest that the lower SB5 scores obtained by children in the LD group as compared to children in the Control group cannot be ascribed to either a greater number of behavior problems during testing or to a factor that contributes to both lowered test scores and increased problem behavior. As this finding would have important implications for clinical practice and research on these populations, replication of these results would be an important first step for researchers interested in this area.

Summary of SB5 and TOF Findings

The third aim of this study was to determine if combinations of SB5 scores and TOF scores could differentially predict clinical diagnostic group status. As discussed at length in Chapter 2, previous research has demonstrated that to date, cognitive scores alone cannot be used to predict diagnostic group status, although none of the existing research on this topic has investigated tests that were designed to be sensitive to deficits demonstrated by specific populations of children. Observations of test session behavior

have been shown to be reliable predictors of group status; for example, McConaughy and Achenbach (2004) demonstrated that a weighted combination of TOF syndrome scores produced correctly classified 50 – 59% of referred children and 81 – 91% of nonreferred children (overall correct classification rate of 71%). However, no existing research is available regarding the diagnostic utility of a combination of test scores and test session behaviors.

This study therefore examined the predictive associations between test scores and test session behavior and group status. It may be helpful to evaluate the results of this study within a framework of the *sensitivity* and the *specificity* of the results. In describing a classification system in which there are two possible results (e.g., disorder and no disorder), sensitivity refers to how effective the system is at correctly classifying individuals with the disorder, whereas specificity refers to how effective the system is at correctly classifying individuals without the disorder. For our purposes, when comparing children from a specific diagnostic group to children in the Control group, high sensitivity is achieved when test scores, test session behavior, or a combination thereof correctly classifies a high percentage of children with the diagnosed disorder. In contrast, high specificity is achieved when the predictors correctly classify a high percentage of the Control children. Ideally, both sensitivity and specificity should be high for the classification system to be widely useful; however, there are circumstances under which high sensitivity and low specificity might be acceptable (e.g., in designing broad-based interventions designed for children “at risk” for a particular outcome), or where low

sensitivity but high specificity could be acceptable (e.g., when allocating high-cost or intensive interventions to a particular group of children).

Several SB5 and TOF scales emerged as predictors of ADHD versus Control group status. A combination of the SB5 WM and QR factors correctly classified 42% of the ADHD group and 89% of the Control group, whereas the SB5 VIQ and FSIQ correctly classified 10% of the ADHD group and 96 – 99% of the Control group. The TOF Attention Problems syndrome and the DSM-oriented Inattention, TOF Externalizing, and TOF Total Problems scales correctly classified 36 – 56% of the ADHD group and 92 – 96% of the Control group. Although the overall correct classification rates were quite respectable for these scores (ranging from 67 – 82%), when considered separately, the SB5 scores and TOF scores demonstrated high specificity but low sensitivity. In other words, using these predictors, few Control children were incorrectly identified as being in the ADHD group (false positives); however, many children from the ADHD group were incorrectly classified as Control children (false negatives). Intriguingly, using a weighted combination of SB5 WM, SB5 QR, and TOF Attention Problems increased the sensitivity without sacrificing the overall correct classification rate. The model using these three predictors correctly classified 61% of the ADHD group and 92% of the Control group, with an overall correct classification rate of 82%. In addition, the strength of association between these predictors and the categorical group classification was a very respectable 53%, which is moderately to substantially higher than the strength of association between ADHD versus Control group status that was accounted for by SB5 scores alone (4 – 28%) or TOF scores alone (22 – 47%). The

classification results are similar to the correct classification rates that Glutting and colleagues (1997) obtained through discriminant analysis of GATSB scores, which correctly classified 71% of children with ADHD and 90% of children without ADHD (overall correct classification rate 81%).

The same combination of SB5 factors that discriminated between ADHD and Control children, namely WM and QR, also emerged as predictors of LD versus Control group status, as did VIQ and FSIQ. However, in contrast to the pattern found for ADHD versus Control children, these SB5 scales demonstrated high sensitivity, correctly classifying 91% of children in the LD group, but low specificity, correctly classifying only 40 – 45% of Control children. The combination of TOF scores that emerged as predictors of LD versus Control group status, Anxious and Attention Problems, also demonstrated high sensitivity, correctly classifying over 99% of the LD group, but dismal specificity, correctly classifying only 4% of the Control group. In other words, most children with LD were correctly classified as belonging to the LD group, but few of the Control children were correctly classified, thus producing very high false positive rates for LD. These results reflect Barkley's concern (1996; p. 7) regarding the "high positive predictive power" but "lousy negative predictive power" of some predictors. The poor negative predictive power of TOF syndromes for discriminating between children in the LD and Control groups clearly reflects the fact that, as discussed earlier, the children from the LD and Control groups did not differ significantly on any of the TOF syndromes. However, it is interesting to note that the weighted combination of SB5 WM, SB5 QR, and TOF Attention Problems retained high sensitivity for LD (correctly

classifying 91% of LD children) while also producing improvements in specificity (correctly classifying 49% of the Control children), with an overall correct classification rate of 78%. In addition, the strength of association between the SB5 and TOF scales taken together as predictors was larger (34%) than for the TOF scales alone (7%).

In discriminating the children in the EBD group from the children in the Control group, SB5 WM, VIQ, and FSIQ emerged as significant predictors, correctly classifying 7 – 18% of the EBD group and 96 – 98% of the Control group. As with the ADHD versus Control group results, these results reflect excellent specificity, but poor sensitivity. Using the TOF syndromes or broad scales as predictors resulted in much higher sensitivity. For instance, the combination of TOF Language/Thought Problems and Oppositional syndromes correctly classified 52% of the EBD group, while the combination of the TOF Internalizing and Externalizing scales correctly classified 57% of children in the EBD group. The TOF Total Problems scale alone correctly classified 46% of the children in the EBD group. Specificity for these combinations of TOF scales and syndromes was high, resulting in correct classification rates of 96 – 99% of the Control children. The weighted combination of Language/Thought Problems and Oppositional as predictors resulted in an overall correct classification rate (89%),while the weighted combination of the TOF Internalizing and Externalizing scores and the TOF Total Problems score alone produced overall correct classification rates of 88% and 85%, respectively. However, combining SB5 test scores and TOF scores did not increase the overall correct classification rates above those obtained using TOF scales alone. The TOF

predictors showed higher strength of associations with EBD versus Control group status (45 – 52%) than did the SB5 predictors (14 – 19%)

While information regarding the utility of test scores, test session behavior, and combinations of test scores and behavior in distinguishing groups of children receiving special education services from Control children has important implications for clinical practice, individuals who assess children from these populations must also discriminate children with ADHD, LD, and EBD from each other. The results of this study indicated that SB5 QR discriminated children in the ADHD group from children in the LD group with an overall correct classification rate of 82%; however, while almost 98% of the children in the LD group were correctly classified, only 10% of the children in the ADHD were correctly classified. The weighted combination of SB5 WM and QR correctly classified 42% of the ADHD group and 89% of the EBD group, resulting in an overall correct classification rate of 73%. In considering the TOF scales, Withdrawn/Depressed correctly classified 94% of children with ADHD and 41% of children with EBD (overall correct classification rate of 74%), while combination of Attention Problems and Withdrawn/Depressed correctly classified 23% of children in the ADHD group and 98% of children in the LD group (overall correct classification rate of 85%). The overall correct classification rate for the weighted combination of Oppositional and Anxious in classifying children in the LD group versus children in the EBD group was 91%; however, while 98% of the LD group was correctly classified, only 30% of the EBD was classified correctly. Combining QR and Attention Problems correctly classified 33% of children with ADHD and 97% of children with LD, with an

overall correct classification rate of 85%. The predictors that significantly discriminated the ADHD, LD, and EBD groups from one another accounted for 14 – 34% of the variance in discriminating between the groups. While this may seem modest, the ability to discriminate between two groups of children receiving special education using only test scores, test session behavior, or a combination thereof presents a strong diagnostic challenge. As such, these findings are quite exciting and suggest that further research in this area is likely to be a promising undertaking.

Implications for Clinical Practice

The overall results of this investigation suggest that cognitive assessment and behavioral observations during testing can and should be important components of multi-informant, multi-method assessment of children with ADHD, LD, and EBD. In particular, the processes of assessing cognitive functioning with a test designed to be sensitive to the deficits demonstrated by children in these populations, and obtaining standardized information regarding the test session behavior of these children during testing, offer excellent opportunities to obtain information that may meaningfully contribute to diagnosis. The results of this study are consistent with previous research in indicating that it would be unwise to utilize test scores or test session behavior, either alone or in combination, as the sole diagnostic tool. No cognitive test score or behavioral observation score demonstrated acceptably high sensitivity and specificity to suggest that clinicians routinely consider that factor when making diagnostic decisions, nor did any combination of test scores and observations. Of course, major figures in child assessment

would caution that no clinician should use one and only once source of information (e.g., just parent report) when formulating a diagnosis (e.g., Sattler, 1998), and the results of this study are therefore consistent with best practices in child assessment. However, these results do suggest that clinicians should not overlook the information that is provided through test scores and test session behavior, provided that the information is interpreted with due caution.

More specifically, the results of this study emphasized the utility of behavior observations obtained in a standardized format, such as through the TOF. Direct observations of children are critical to the assessment of many childhood disorders, including ADHD (Pelham et al., 2005), anxiety (Silverman & Ollendick, 2005) and conduct problems (McMahon & Frick, 2005), and researchers have noted that "none of the major contexts of child development (e.g., home, school, and community) offers as high a level of professional expertise, observational control, or uniformity of conditions as the context of individual test-taking" (Glutting et al., 1996, p. 94). However, as discussed in detail in Chapter 2, the potential utility of behavior observations is often squandered by clinicians, as examiners typically do not observe children in any reliable or valid manner. The results of this study underscore the importance of the information that clinicians are failing to obtain when they do not take advantage of standardized test observation systems. Clearly, test session behavior can contribute meaningfully to the assessment of children from exceptional populations as well as to differentiating among these children and normal controls, and this information should be one component of a multi-method, multi-informant assessment.

As noted above, children with ADHD, LD, and EBD obtained lower SB5 FSIQ scores than normal controls, which is consistent with previous research. However, this finding may also have implications for clinical practice, depending on how representative the children who receive special education services are of the larger group of children with psychoeducational disorders. Children who qualify for special education services are often considered to be demonstrating achievement that is not commensurate with their intellectual functioning. If children who receive special education services for ADHD, LD, and/or EBD are similar in their cognitive functioning to their peers with these disorders who do not qualify for services, children from exceptional groups who perform at a level commensurate with their cognitive functioning would be nevertheless be disadvantaged in the classroom when compared to normal controls (that is, their academic functioning, like their cognitive functioning, would fall approximately one-third to two-thirds of a standard deviation below the mean). In contrast, if children who receive services differ in terms of their cognitive functioning from children who experience attentional, learning, or emotional problems but who do not receive services, then perhaps children who receive special education services could be more accurately described as children with learning/attention/emotional and cognitive challenges, rather than as children who are only struggling with their learning, attentional, or emotional functioning. Resolving these questions could have substantial implications for our understanding of children in mainstream and special educational classrooms.

Limitations

This study has several limitations. One important limitation concerns the method in which the ADHD, LD, and EBD samples were defined. The exceptional samples were not selected specifically for this study. Rather, the participants were selected from a larger sample of children from exceptional populations, to whom the SB5 was administered in order to establish validity in these populations prior to its publication. All of the participants in this study were receiving special education services for either ADHD, LD, or EBD, and the determination regarding their eligibility for special education services were made by multidisciplinary team decisions in accordance with IDEA 2004. However, further information regarding these participants was not available. It is not known, for instance, with which subtype of ADHD most of the participants in the ADHD sample had been diagnosed. Similarly, it is not known what type of learning disability the children in the LD sample experienced, nor is it known with which of the multitude of possible emotional and behavioral disorders the children in the EBD sample had been diagnosed. Furthermore, while children who were receiving special education services under two or more categories at the same time (e.g., ADHD and LD simultaneously) were purposely excluded in this study, it cannot be stated definitively that study participants did not meet criteria for comorbid conditions. For instance, given the high rate of comorbidity between LD and ADHD, it is highly likely that some of the participants in the LD sample met diagnostic criteria for ADHD; however, they may have received educational services only under the category of LD, potentially limiting the

ability to find meaningful differences between children with LD and children with ADHD.

These are important limitations to this investigation, as the results of this study may obscure important differences between the children within the samples. As one example, children with the Primarily Inattentive subtype of ADHD may demonstrate different SB5 and TOF scores from children with the Combined subtype of ADHD. If a small portion of children in a sample differed in some important way from the larger sample (for instance, if a small number of the ADHD sample was diagnosed with the Primarily Inattentive subtype while the majority were diagnosed with the Combined subtype), the results of this study would likely demonstrate poor external validity for children similar to that subsample. As another example, perhaps children with language-based learning disabilities differ substantially from children with learning disabilities in math (e.g., perhaps children with an LD in reading would obtain lower VIQs than children with an LD in math, while the opposite pattern would be observed for NVIQ). If the LD group contained a large percentage of children with one type of LD and a smaller group of children with a different type of disability, grouping these children together into a larger LD sample may have hidden important within-group differences.

Another limitation to this study concerns the small sample size of the ADHD and EBD samples. While the overall number of study participants was quite adequate, most of the participants were from the LD and Control groups. This limitation is especially important because this study used age and gender as between-group variables. When conducting analyses that considered the ADHD and EBD groups, this resulted in cells

with very small numbers of participants; for instance, there were very few females (from 4 to 10) in each age group in both of these samples. The small samples reduced the power to detect age and gender effects, particularly within the ADHD and EBD groups. The sample size limitation also precluded analyses of whether the combinations of SB5 and TOF scores that discriminated between groups differed by age and/or gender. In addition, this limitation also precluded analyses of associations between test scores and test session behavior in children with ADHD who were on medication during testing versus children with ADHD who were not on medication during testing.

An additional limitation of this study concerns the challenges associated with conducting a large number of analyses. Because of the large number of tests conducted for this investigation, it is likely that some of the statistically significant results were chance effects. For example, when examining patterns of SB5 scores, pairwise comparisons (e.g., ADHD versus Control; ADHD vs. LD, etc.) were conducted for 8 scales (FSIQ, VIQ, NVIQ, and the five factors). Following Sakoda et al.'s (1954) criteria for the .05 significance level and the .05 protection level, within each set of pairwise comparisons for SB5 scores, 2 significant effects may have been due to chance. For example, when comparing children with LD to Control children, 8 significant differences emerged (Control children scored higher on all SB5 scales) and 2 of these effects may have been chance effects. Similarly, within each set of pairwise comparisons conducted for the 11 TOF syndromes and scales, 2 effects may have been chance effects according to Sakoda et al.'s criteria. At the same time, several patterns emerged consistently across analyses (e.g., WM was consistently lower in exceptional children as compared to the

Control group), and the consistency of these results suggests that those observed patterns are less likely to be due to chance.

It is also important to consider that this study utilized standardized behavioral observations. It is possible that the TOF is not sensitive to some observable behaviors which could meaningfully discriminate between groups. As one example, although there are many items on the TOF that address task approach, these items are not specific to the type of task (e.g., verbal or quantitative) that the child is engaging in; for instance, perhaps children who complete *quantitative* problems slowly and effortfully are more likely to be children from the LD group than children from the Control group. Behaviors that are highly specific to a particular group of children but which may have a very low base-rate within the larger normative sample of children engaged in testing (e.g., expressing suicidal ideation, which may be highly specific to children with EBD but unlikely to occur frequently within the normative sample) may be particularly likely to be missed by the TOF, as low frequency items were omitted when the final TOF item set was selected. While nonstandardized behavioral observations do have some advantages over standardized behavioral observations, such as the ability to discern these low base-rate behaviors, the advantages of standardized behavioral observations far outweigh the disadvantages. This is especially true in research designs such as the present study, which combine observations from multiple examiners to analyze patterns of test behavior (and combinations of test behavior and cognitive test scores) which can distinguish between groups of children. The reliability of such studies is immeasurably enhanced by the use of standardized observations. Therefore, while the use of standardized observations may be

considered a minor limitation of this study, it can also be viewed a considerable strength of the present investigation.

Future Directions for Research

It will be important for future research on the test scores and test session behavior of children with ADHD, LD, and EBD to avoid the limitations of the present study. A preliminary first step in future research would be to replicate the results of this study using larger samples that are more clearly defined. For instance, the particular subtype of ADHD, type of LD, and specific emotional or behavioral disorder with which the children in each sample have been diagnosed, represents crucial information for determining the applicability of these results to the broader population of children with these disorders. Research is currently underway investigating associations between test scores (on the WISC-IV) and test session behavior in carefully-defined samples of children with ADHD, clinic-referred children without ADHD, and nonreferred children (S.H. McConaughy, personal communication).

While the results of this study are limited by sample size and possible sample heterogeneity, the results do suggest that consideration of age group and gender will be necessary in future research on test scores and test session behavior in these populations, perhaps particularly for children with emotional and behavioral problems. Extending the research questions from this study to other exceptional populations (such as gifted children, children with speech/language impairment, children with mental retardation, and other groups) and examining differences between children who are on medication

during testing versus children who are not on medication during testing are also potentially useful avenues for future research. In addition, examining patterns of verbal and nonverbal test scores may also be interesting. The SB5 factor scores are standard scores that are calculated based on a composite of the individual's scaled scores on the verbal and nonverbal task that comprise that factor (e.g., the Working Memory factor score is calculated based on a composite of the scaled scores on the Verbal Working Memory and Nonverbal Working Memory tasks). As this study was the first study to examine test scores on the SB5 in children with ADHD, LD, and EBD, it was not feasible to examine the data using these scaled scores in addition to the factor scores. However, as some research has suggested that children with learning and attention difficulties may differ in terms of their verbal and nonverbal profiles of strengths and weaknesses (e.g., McInnes et al., 2003), researchers may find this a compelling area for exploration.

The results of this study also illuminate potential directions for future research regarding similarities and differences between children with ADHD, LD, and EBD. Two factors from the SB5, namely Working Memory and Quantitative Reasoning, consistently emerged as significant predictors of diagnostic group status. Direct comparison of the diagnostic groups indicated that children with ADHD, LD, and EBD performed similarly to each other but different from children in the Control group on WM, while children with LD and EBD performed lower than children in the Control group on QR (children in the LD group also performed lower than children in the ADHD and EBD groups on QR). Thus, while children in the LD group were similar to children in the ADHD group on WM, the children in the LD group were very different from the ADHD group on QR.

These patterns offer potential clues about the shared correlates of attention, learning, and emotional problems as well as factors which may be specific to particular groups of children. Investigating whether this group pattern represents a potential "profile" that has diagnostic utility for distinguishing between individual children with ADHD and children with LD might also be worthwhile.

Similarly, the results of this study suggest potential directions for future research on the similarities and differences in test session behavior amongst children with ADHD, LD, and EBD. This study found that children with EBD differed significantly from children with LD or ADHD and normal controls in many aspects of their test session behavior, including their overall level of problem behavior during testing, as well as the amount of withdrawn behavior, language/thought problems, test anxiety, oppositionality, and attentional difficulties that they displayed. Children with ADHD were similar to children with EBD in terms of the level of attention problems they displayed during testing, but were otherwise dissimilar from children with EBD, while children with LD were generally indistinguishable from normal controls in terms of their test session behavior. In addition to replicating these results, future researchers may wish to explore how these patterns differ from children with comorbid attention, learning, and/or emotional issues and from children with ADHD, LD, or EBD who do not qualify for special education services.

Finally, attempting to clarify the contribution of the relationship between test scores and test session behavior to both exosession and intrasession validity is likely to be a difficult, but worthwhile, endeavor. As noted earlier, clinicians routinely use test

session behavior to generate hypotheses regarding both the validity of the test scores that are obtained and the external validity of the behavior observed during testing. However, for any given child who obtains test scores below the mean and who demonstrates a high level of problem behavior during testing, it is difficult to determine whether the test session behavior contributed to the low test scores or vice versa, or whether a third factor contributed to both poor test scores and elevated levels of problem behavior; nevertheless, these are essential considerations for clinicians. The results of this study offer exciting clues regarding these questions, in that they suggest that for children with LD as a group, behavior during testing is unlikely to have contributed significantly to the discrepancy between the test scores that these children obtained compared to the scores obtained by normal controls. Future research should examine whether these patterns hold true for other groups of children as well as for individual children.

Table 1

SB5 Standardized Scores of Children with ADHD, LD, and EBD as Reported in the SB5 Technical Manual (Roid, 2003a)

	Diagnostic Group		
SB5 Factor	Children with ADHD (N = 94)	Children with LD-Reading (N = 212)	Children with EBD (N = 48)
VIQ	92.3 (16.6)	84.3 (14.3)	87.9 (16.9)
NVIQ	93.1 (16.6)	85.6 (13.5)	84.9 (18.3)
FSIQ	92.2 (16.1)	84.1 (13.8)	85.4 (17.9)
FR	93.4 (17.5)	86.8 (13.8)	90.9 (14.6)
KN	92.7 (16.5)	85.0 (14.2)	87.4 (16.1)
QR	95.9 (15.7)	87.1 (11.0)	88.7 (17.4)
VS	95.1 (14.6)	88.1 (15.3)	86.0 (16.9)
WM	90.2 (13.7)	85.6 (15.2)	86.0 (18.4)

Note. Numbers in parentheses are standard deviations. VIQ = Verbal IQ, NVIQ = Nonverbal IQ, FSIQ = Full Scale IQ, FR = Fluid Reasoning, KN = Knowledge, QR = Quantitative Reasoning, VS = Visual-Spatial Processing, WM = Working Memory.

Table 2

Age Group (6-11 or 12-18) and Gender by Diagnostic Group

	Diagnostic Group					
	ADHD (N = 50)		LD (N = 234)		EBD (N = 28)	
Gender	6-11	12-18	6-11	12-18	6-11	12-18
Males	22	12	103	38	12	6
Females	10	6	61	32	4	6
Total	32	18	164	70	16	12

Table 3

SB5 Standardized Scores by Diagnostic Group

	Diagnostic Group			
SB5 Factor	ADHD (N = 50)	LD (N = 234)	EBD (N = 28)	Control (N = 100)
VIQ	96.80 (15.84)$_{a}$	90.03 (10.90)$_{b,c,d}$	93.46 (14.12)	102.32 (11.54)
NVIQ	97.26 (15.10)	89.51 (10.98)$_{b,c}$	91.25 (16.56)$_{e}$	101.11 (13.51)
FSIQ	96.70 (15.48)$_{a}$	89.18 (10.59)$_{b,c}$	91.86 (15.47)$_{e}$	101.73 (12.33)
FR	97.70 (17.22)	91.41 (11.56)$_{b,c}$	94.93 (12.48)	101.71 (12.56)
KN	97.36 (15.98)	89.75 (12.48)$_{b,c}$	92.86 (15.18)	100.73 (12.55)
QR	100.42 (16.15)	90.38 (10.82)$_{b,c,d}$	94.32 (16.46)$_{e}$	101.01 (12.67)
VS	97.82 (12.88)	92.95 (11.92)$_{b,c}$	92.36 (14.54)$_{e}$	101.33 (13.37)
WM	93.34 (13.14)$_{a}$	90.02 (12.05)$_{b}$	91.50 (16.70)$_{e}$	103.57 (12.17)

Note. Numbers in parentheses are standard deviations. VIQ = Verbal IQ, NVIQ = Nonverbal IQ, FSIQ = Full Scale IQ, FR = Fluid Reasoning, KN = Knowledge, QR = Quantitative Reasoning, VS = Visual-Spatial Processing, WM = Working Memory. Significant mean differences are represented by subscripts: $_{a}$ = Control > ADHD, $_{b}$ = Control > LD, $_{c}$ = ADHD > LD, $_{d}$ = EBD > LD, and $_{e}$ = Control > EBD (see also Table 5).

Table 4

TOF Raw Scores by Diagnostic Group

	Diagnostic Group			
TOF Syndrome	ADHD (N = 50)	LD (N = 234)	EBD (N = 28)	Control (N = 100)
WDD	1.18 (2.15)	1.81 (5.11)	6.82 (8.53)$_{d,e,f}$	0.91 (2.31)
LTP	1.73 (2.12)$_{a}$	1.16 (2.73)	3.18 (2.96)$_{d,e,f}$	0.53 (1.10)
ANX	4.00 (4.43)$_{a,b}$	1.58 (3.31)	6.81 (6.27)$_{d,e,f}$	1.82 (3.47)
OPP	2.74 (7.50)$_{a,b}$	0.50 (1.54)	6.14 (9.04)$_{d,e,f}$	0.32 (1.29)
ATT	9.39 (8.73)$_{a,b}$	2.74 (4.81)	8.86 (8.60)$_{d,f}$	1.38 (3.37)
TOF INT	3.00 (3.01)	3.03 (4.81)	10.00 (10.57)$_{d,e,f}$	1.44 (2.84)
TOF EXT	12.12 (13.88)$_{a,b}$	3.27 (5.90)	15.00 (16.04)$_{d,f}$	1.70 (4.19)
TOF TOT	23.16 (27.95)$_{a,b}$	8.61 (14.49)	41.32 (33.55)$_{d,e,f}$	6.97 (11.76)
DSM-IN	5.76 (5.11)$_{a,b}$	1.73 (3.31)	5.93 (5.96)$_{d,f}$	0.74 (1.78)
DSM-HI	5.38 (6.26)$_{a,b}$	1.18 (2.62)	4.68 (5.12)$_{d,f}$	0.71 (1.92)
DSM ADHP	5.38 (6.26)$_{a,b}$	1.18 (2.62)$_{c}$	4.68 (5.12)$_{d,e,f}$	0.71 (1.92)

Note. Numbers in parentheses are standard deviations. WDD = Withdrawn/Depressed, LTP = Language/Thought Problems, ANX = Anxious, OPP = Oppositional, ATT = Attention Problems, INT = TOF Internalizing, TOF EXT = TOF Externalizing, TOF TOT = TOF Total Problems, DSM-IN = DSM Inattentive, DSM-HI = DSM Hyperactivity-Impulsivity, DSM-ADHP = DSM Attention Deficit/Hyperactivity Problems Total. Significant mean differences are represented by subscripts: $_{a}$ = ADHD > Control, $_{b}$ = ADHD > LD, $_{c}$ = LD > Control, $_{d}$ = EBD > Control, $_{e}$ = EBD > ADHD, and $_{f}$ = EBD > LD (see also Table 6).

Table 5

Significant Group Differences and Effect Sizes on SB5 Scales for Children with ADHD, LD, and EBD, and Control Children

SB5 Scale	F	p	Eta^2	Group Differences[a]
VIQ	F(3,396) = 26.19	< .001	.166	Control > ADHD, LD, EBD ADHD, EBD > LD
NVIQ	F(3,396) = 21.45	< .001	.140	Control > LD, EBD ADHD > LD
FSIQ	F(3,396) = 27.21	< .001	.171	Control > ADHD, LD, EBD ADHD > LD
FR	F(3,396) = 15.69	< .001	.106	ADHD, Control > LD
KN	F(3,396) = 19.93	< .001	.131	ADHD, Control > LD
QR	F(3,396) = 21.41	< .001	.140	ADHD, EBD, Control > LD Control > EBD
VS	F(3,396) = 12.37	< .001	.086	ADHD, Control > LD Control > EBD
WM	F(3,396) = 25.45	< .001	.162	Control > ADHD, LD, EBD

[a] Scheffe pairwise tests, $p < .05$.

Note. VIQ = Verbal IQ, NVIQ = Nonverbal IQ, FSIQ = Full Scale IQ, FR = Fluid Reasoning, KN = Knowledge, QR = Quantitative Reasoning, VS = Visual-Spatial Processing, WM = Working Memory.

Table 6

Significant Group Differences and Effect Sizes on TOF Syndromes and Scales for Children with ADHD, LD, and EBD, and Control Children

TOF Scale	F	p	Eta^2	Group Differences[a]
WDD	F(3,380) = 10.24	< .001	.075	EBD > ADHD, LD, Control
LTP	F(3,380) = 9.84	< .001	.072	EBD > ADHD, LD, Control ADHD > Control
ANX	F(3,380) = 17.19	< .001	.119	EBD > ADHD, LD, Control ADHD > LD, Control
OPP	F (3,380) = 8.07	< .001	.060	EBD > ADHD, LD, Control ADHD > LD, Control
ATT	F (3,380) = 39.61	< .001	.238	ADHD, EBD > LD, Control
TOF INT	F(3,384) = 12.18	< .001	.087	EBD > ADHD, LD, Control
TOF EXT	F(3.384) = 27.62	< .001	.177	ADHD, EBD > LD, Control
TOF TOT	F(3,396) = 27.51	< .001	.172	EBD > ADHD, LD, Control ADHD > LD, Control
DSM-IN	F(3,393) = 27.29	< .001	.172	ADHD, EBD > LD, Control
DSM-HI	F(3,393) = 30.41	< .001	.188	ADHD, EBD > LD, Control
DSM ADHP	F(3,393) = 34.52	< .001	.209	EBD > ADHD, LD, Control ADHD > LD, Control LD > Control

[a] Scheffe pairwise tests, $p < .05$.

Note. WDD = Withdrawn/Depressed, LTP = Language/Thought Problems, ANX = Anxious, OPP = Oppositional, ATT = Attention Problems, INT = TOF Internalizing, TOF EXT = TOF Externalizing, TOF TOT = TOF Total Problems, DSM-IN = DSM Inattentive, DSM-HI = DSM Hyperactivity-Impulsivity, DSM-ADHP = DSM Attention Deficit/Hyperactivity Problems Total.

Table 7

SB5 and TOF Predictors of ADHD versus Control Children

	Sensitivity	Specificity	False Positives	False Negatives	Overall Correct	Nagelkerke R^2
SB5 Scales						
WM & QR	42.0	89.0	11.0	58.0	73.3	.28
VIQ	10.0	96.0	4.0	90.0	67.3	.05
FSIQ	10.0	99.0	1.0	90.0	69.3	.04
TOF Scales						
ATT[a]	53.2	95.0	5.0	46.8	81.6	.41
DSM-IN	56.0	92.0	8.0	44.0	80.0	.47
EXT[b]	54.2	96.0	4.0	45.8	82.4	.41
TOT	36.0	94.0	6.0	64.0	74.7	.22
SB5 & TOF Scales						
WM & QR & ATT[c]	61.2	92.0	8.0	38.8	81.9	.53
WM & QR & DSM-IN	58.0	92.0	8.0	42.0	80.7	.56
FSIQ & TOT	40.0	94.0	6.0	60.0	76.0	.26

Note. ADHD $N = 50$ except for [a]($N = 47$), [b]($N = 48$), and [c]($N = 49$), Control $N = 100$. WM = Working Memory, QR = Quantitative Reasoning, VIQ = Verbal IQ, FSIQ = Full Scale IQ, ATT = TOF Attention Problems, EXT = TOF Externalizing, TOT = TOF Total Problems, DSM-IN = TOF DSM-Inattentive

Table 8

SB5 and TOF Predictors of LD versus Control Children

	Sensitivity	Specificity	False Positives	False Negatives	Overall Correct	Nagelkerke R^2
SB5 Scales						
WM & QR	91.5	45.0	55.0	8.5	77.5	.32
VIQ	91.5	42.0	58.0	8.5	76.6	.29
FSIQ	91.9	40.0	60.0	8.1	76.3	.29
TOF Scales						
ANX & ATT[a]	99.1	4.0	96.0	0.8	69.7	.07
TOF & SB5 Scales						
WM & QR & ATT[b]	90.9	49.0	51.0	9.1	78.2	.34

Note. LD $N = 234$ except for [a]($N = 223$) and [b]($N = 230$), Control $N = 100$. WM = Working Memory, QR = Quantitative Reasoning, VIQ = Verbal IQ, FSIQ = Full Scale IQ, ANX = TOF Anxious, ATT = TOF Attention Problems.

Table 9

SB5 and TOF Predictors of EBD versus Control Children

	Sensitivity	Specificity	False Positives	False Negatives	Overall Correct	Nagelkerke R^2
SB5 Scales						
WM	17.9	97.0	3.0	82.1	79.7	.19
VIQ	10.7	96.0	4.0	89.3	77.3	.14
FSIQ	7.1	98.0	2.0	92.9	78.1	.14
TOF Scales						
LTP & OPP[a]	51.9	99.0	1.0	48.1	89.0	.48
INT & EXT	57.1	96.0	4.0	42.9	87.5	.52
TOT	46.4	96.0	4.0	53.6	85.2	.45
TOF & SB5 Scales						
FSIQ & TOT	50.0	95.0	5.0	50.0	85.2	.51

Note. EBD N =28 except for [a](N = 27), Control N = 100. WM = Working Memory, VIQ = Verbal IQ, FSIQ = Full Scale IQ, LTP = Language/Thought Problems, OPP = TOF Oppositional, INT = TOF Internalizing, EXT = TOF Externalizing, TOT = TOF Total Problems.

Table 10

SB5 and TOF Predictors of Children with ADHD, LD, and EBD

	ADHD versus LD			
	Percent Correctly Classified ADHD	Percent Correctly Classified LD	Overall Correct	Nagelkerke R^2
SB5 Scales				
QR	10.0	97.9	82.4	.14
TOF Scales				
ATT & WDD[a]	23.4	97.8	84.8	.24
TOF & SB5 Scales				
QR & ATT[b]	32.7	96.5	85.3	.34

	ADHD versus EBD			
	Percent Correctly Classified ADHD	Percent Correctly Classified EBD	Overall Correct	Nagelkerke R^2
SB5 Scales				
WM & QR	42.0	89.0	73.3	.28
TOF Scales				
WDD[c]	93.6	40.7	74.3	.34

	LD versus EBD			
	Percent Correctly Classified LD	Percent Correctly Classified EBD	Overall Correct	Nagelkerke R^2
TOF Scales				
OPP & ANX[d]	98.2	29.6	90.8	28.1

Note. ADHD N =50, LD N = 234, EBD N =28 except for [a](ADHD N = 47, LD N = 233), [b](ADHD N = 49, LD N = 230), [c](ADHD N = 47, EBD N = 27), and [d](LD N = 223, EBD N = 27). QR = Quantitative Reasoning, WM = Working Memory, ATT = TOF Attention Problems, WDD = TOF Withdrawn/Depressed, OPP = TOF Oppositional, ANX = TOF Anxious.

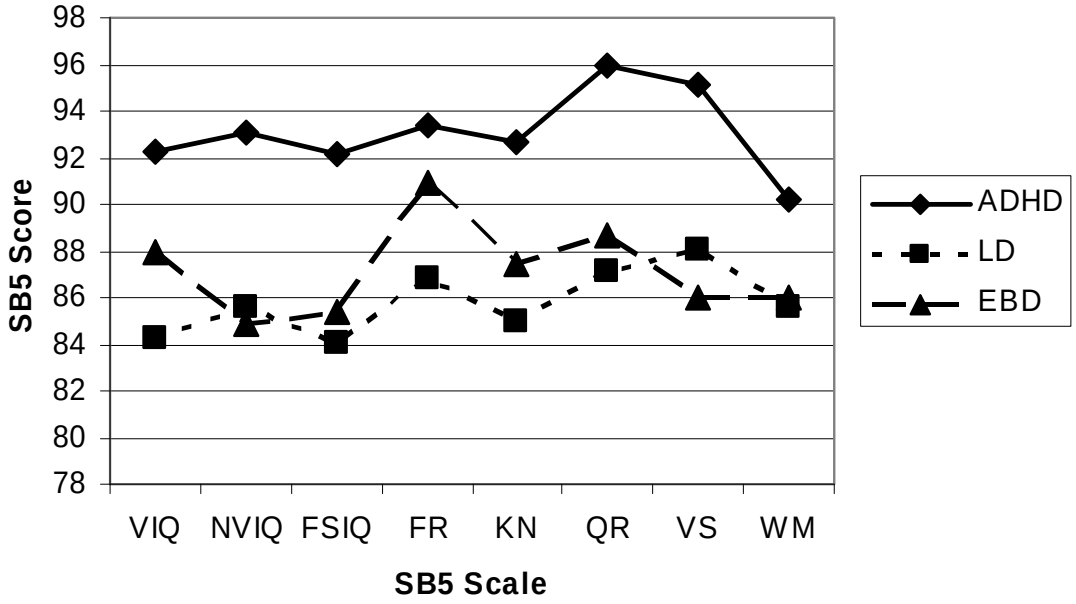


Figure 1.

SB5 Standardized Scores of Children with ADHD, LD, and EBD as Reported in the SB5 Technical Manual (Roid, 2003a).

Note. VIQ = Verbal IQ, NVIQ = Nonverbal IQ, FSIQ = Full Scale IQ, FR = Fluid Reasoning, KN = Knowledge, QR = Quantitative Reasoning, VS = Visual-Spatial Processing, WM = Working Memory.

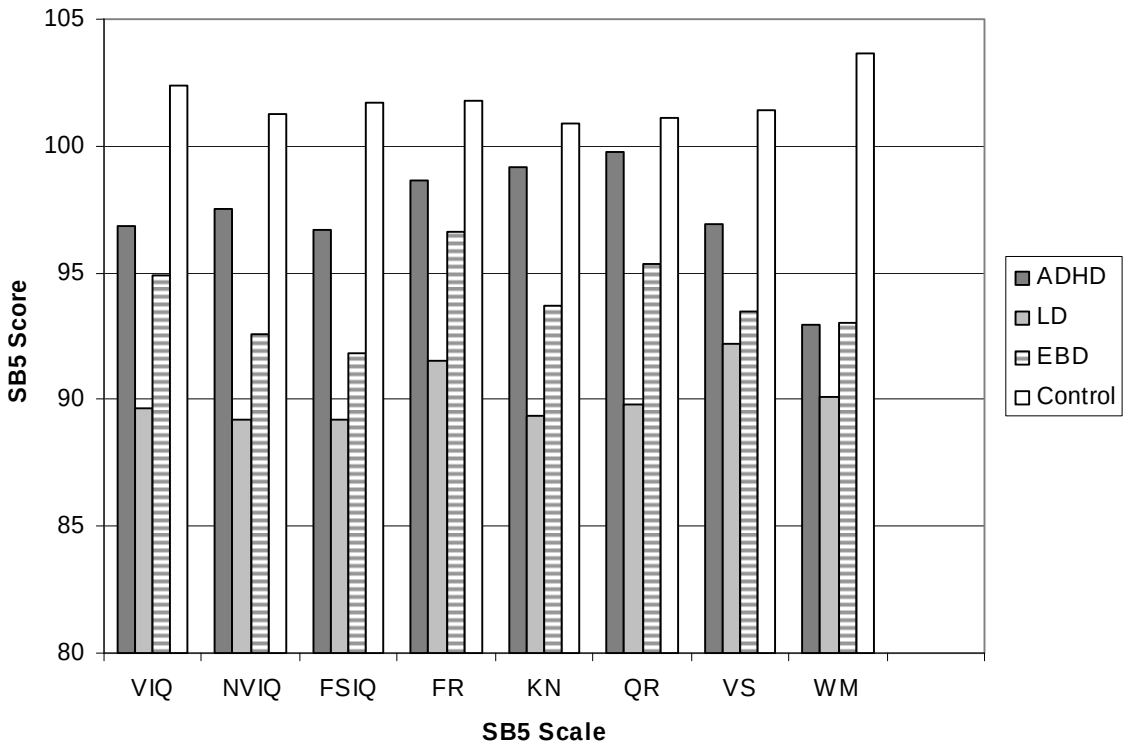


Figure 2.

Standardized Scores on the SB5 VIQ, NVIQ, FSIQ, and the 5 SB5 Factors by Diagnostic Group.

Note. N = 412. VIQ = Verbal IQ, NVIQ = Nonverbal IQ, FSIQ = Full Scale IQ, FR = Fluid Reasoning, KN = Knowledge, QR = Quantitative Reasoning, VS = Visual-Spatial Processing, WM = Working Memory.

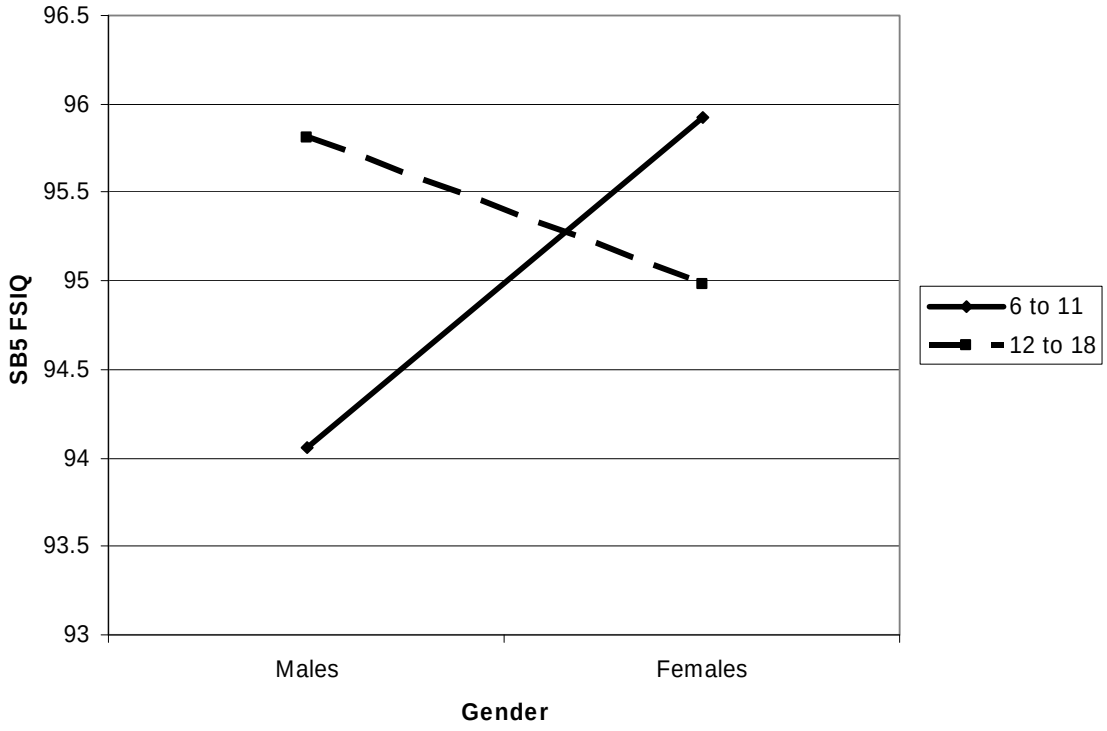


Figure 3.

SB5 Full-Scale IQ Standardized Scores by Age Group and Gender.

Note. N = 412.

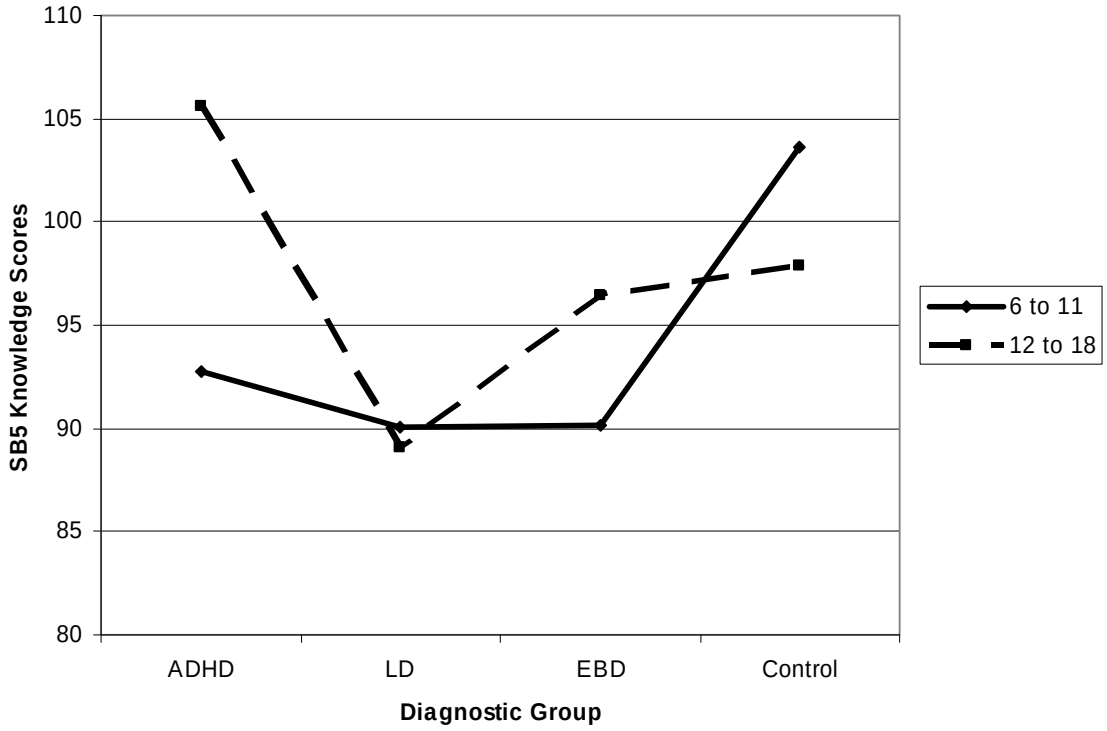


Figure 4.

SB5 Knowledge Standardized Sscores by Age Group and Diagnostic Group.

Note. N = 412.

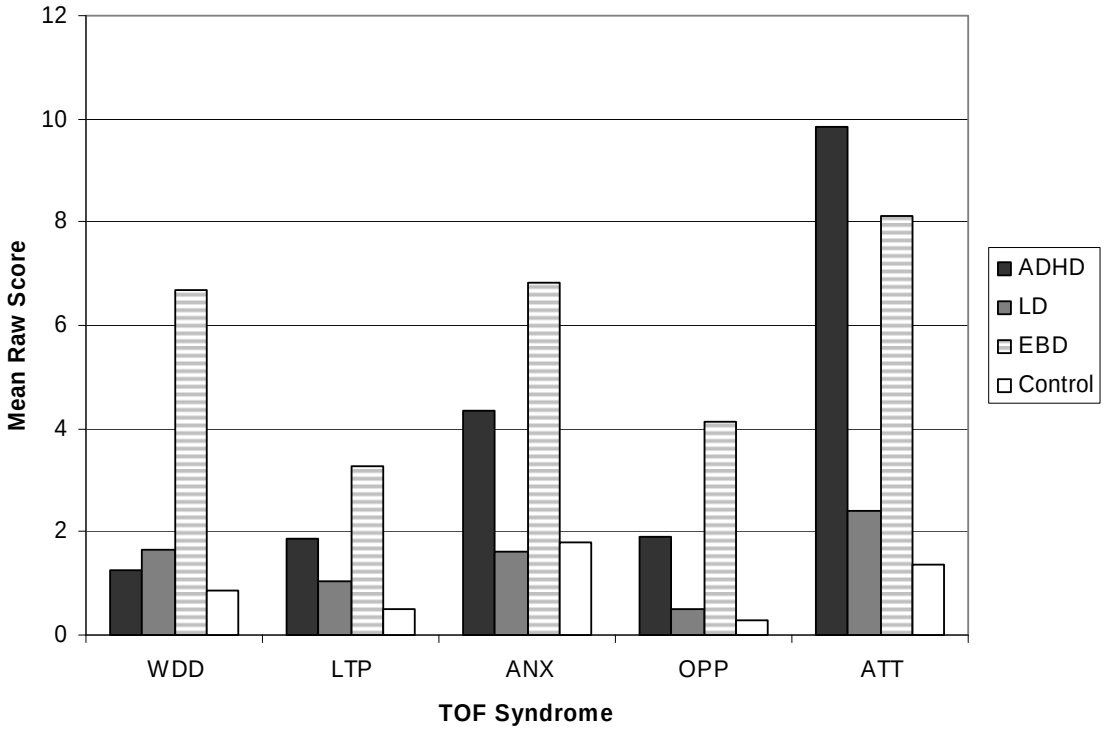


Figure 5.

Raw Scores on the TOF Syndromes by Diagnostic Group.

Note: N = 412. WDD = Withdrawn/Depressed, LTP = Language/Thought Problems, ANX = Anxious, OPP = Oppositional, ATT = Attention Problems.

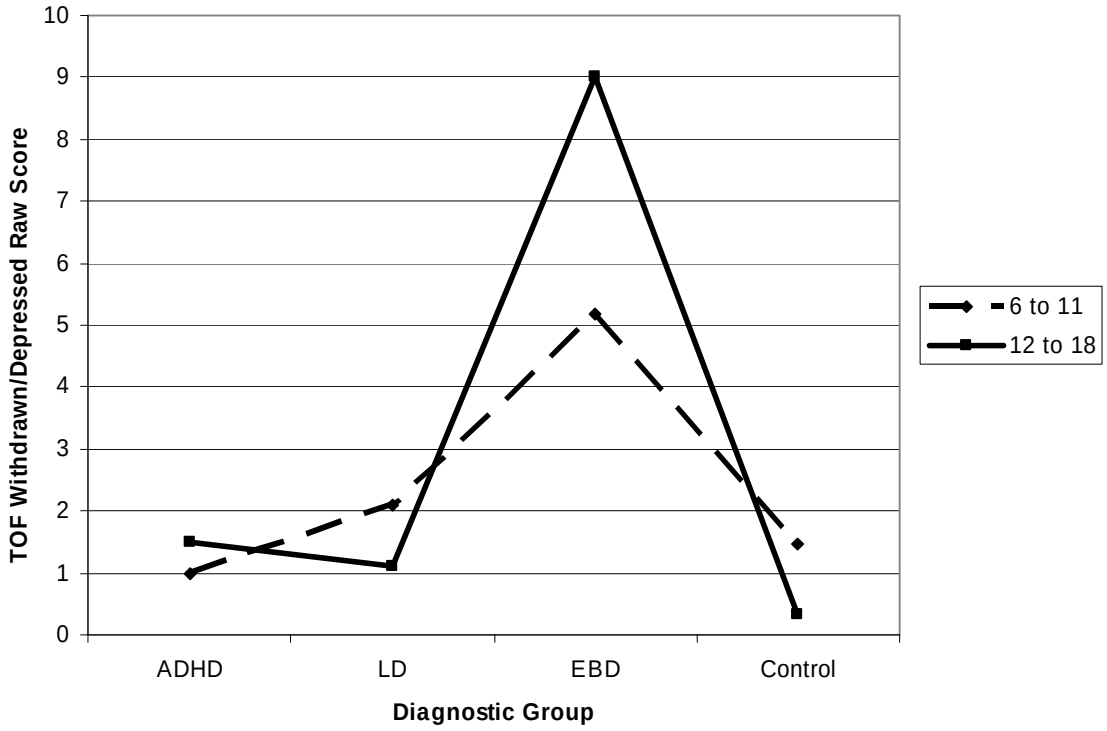


Figure 6.

TOF Withdrawn/Depressed Raw Scores by Age Group and Diagnostic Group.

Note. N = 412.

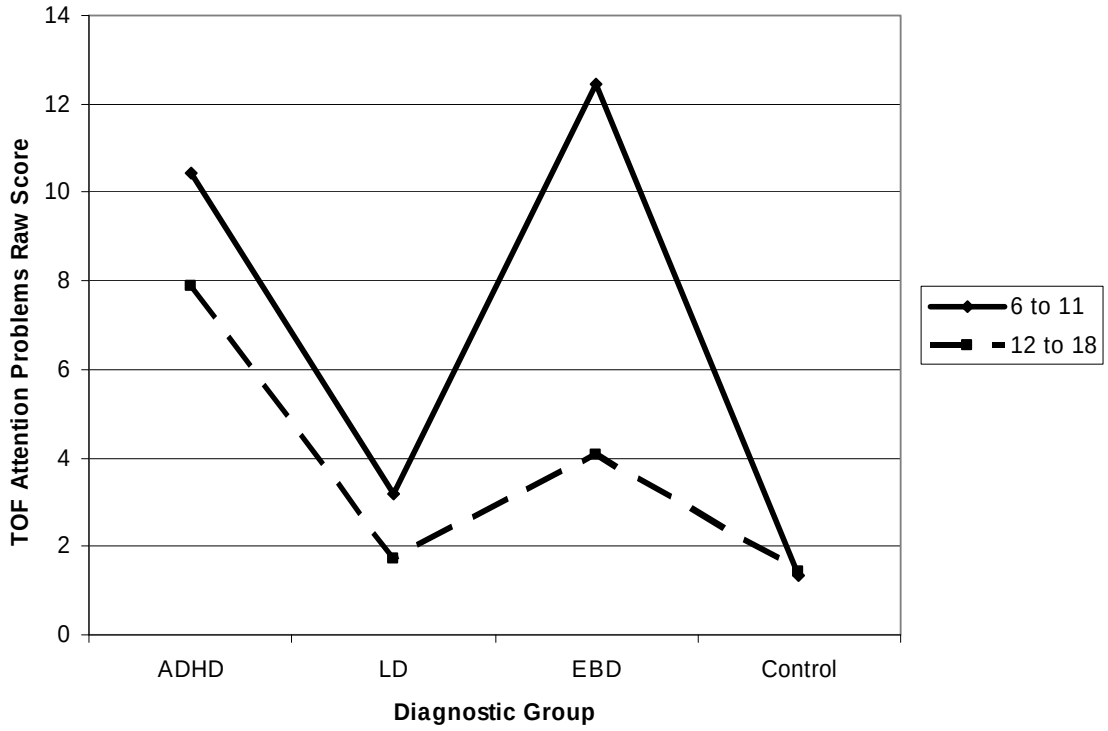


Figure 7.

TOF Attention Problems Raw Scores by Age Group and Diagnostic Group.

Note: N = 413.

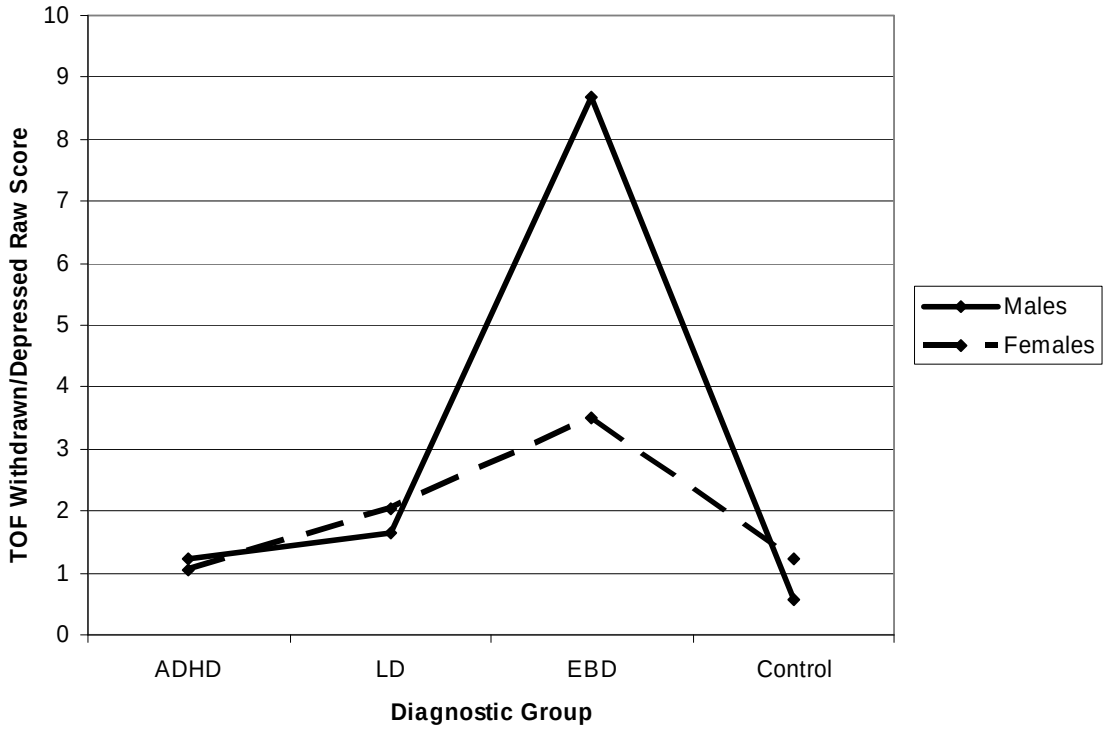


Figure 8.

TOF Withdrawn/Depressed Raw Scores by Gender and Diagnostic Group.

Note. N = 412.

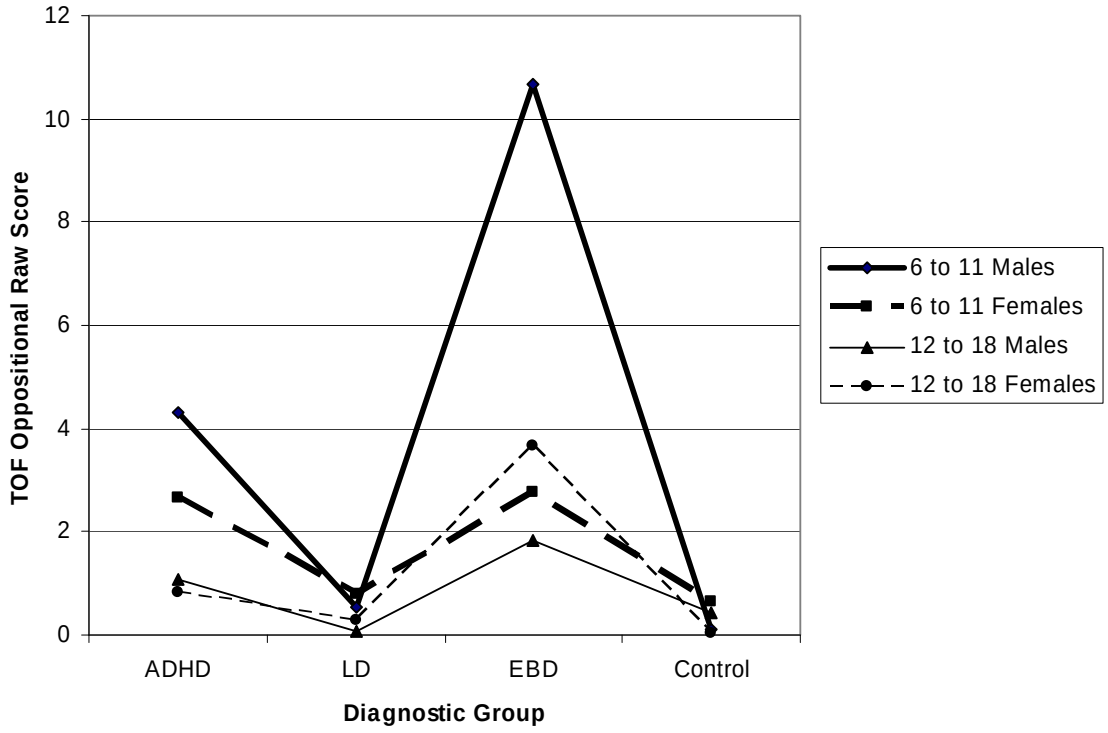


Figure 9.

TOF Oppositional Raw Scores by Age Group, Gender, and Diagnostic Group.

Note. N = 412.

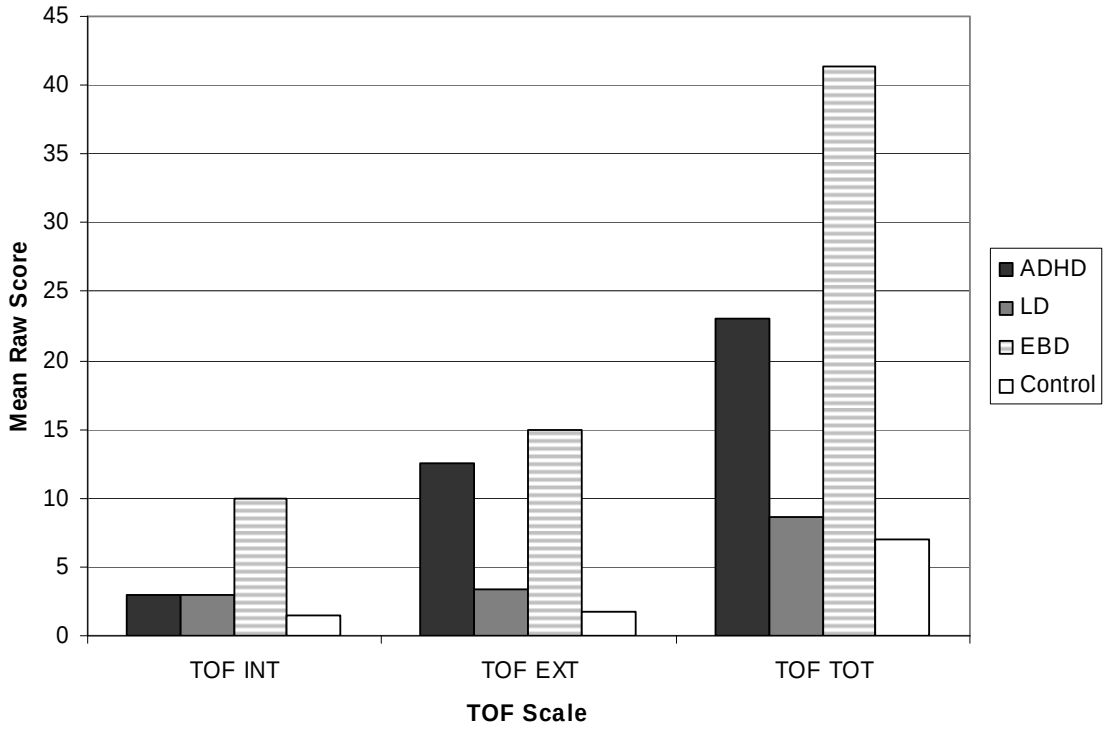


Figure 10.

Raw Scores on TOF Internalizing, TOF Externalizing, and TOF Total Problems by Diagnostic Group.

Note. N = 412. TOF INT = TOF Internalizing Problems, TOF EXT = TOF Externalizing Problems, TOF TOT = TOF Total Problems.

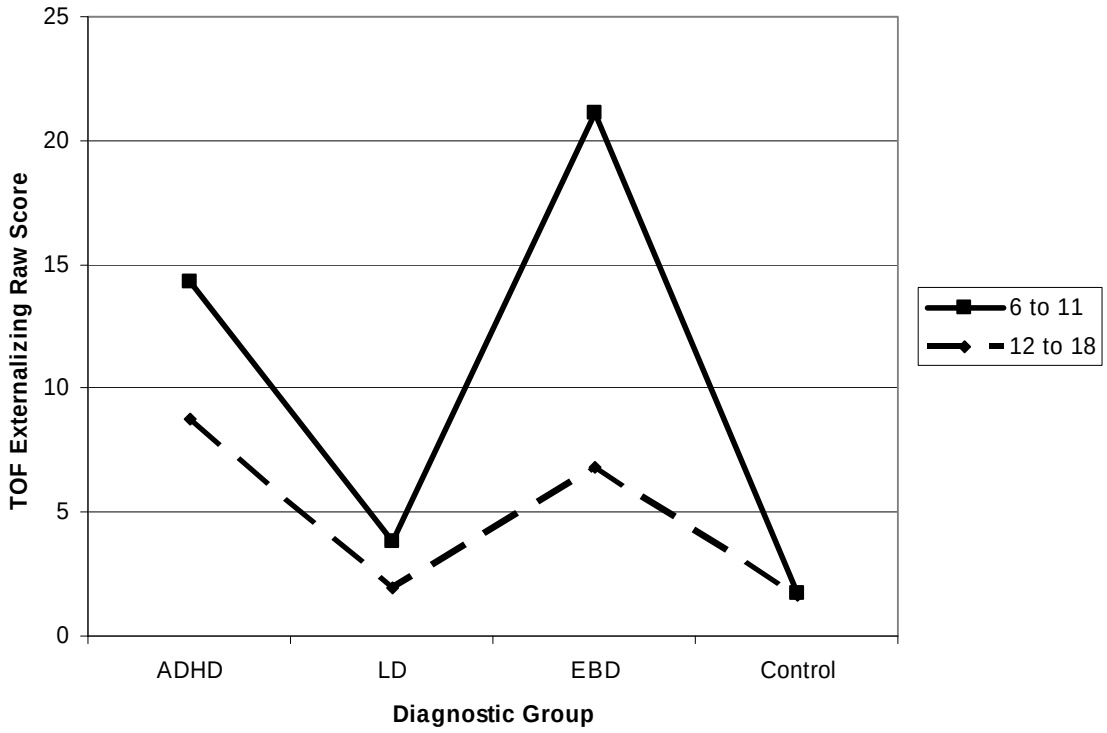


Figure 11.

TOF Externalizing Raw Scores by Age Group and Diagnostic Group.

Note. N = 412.

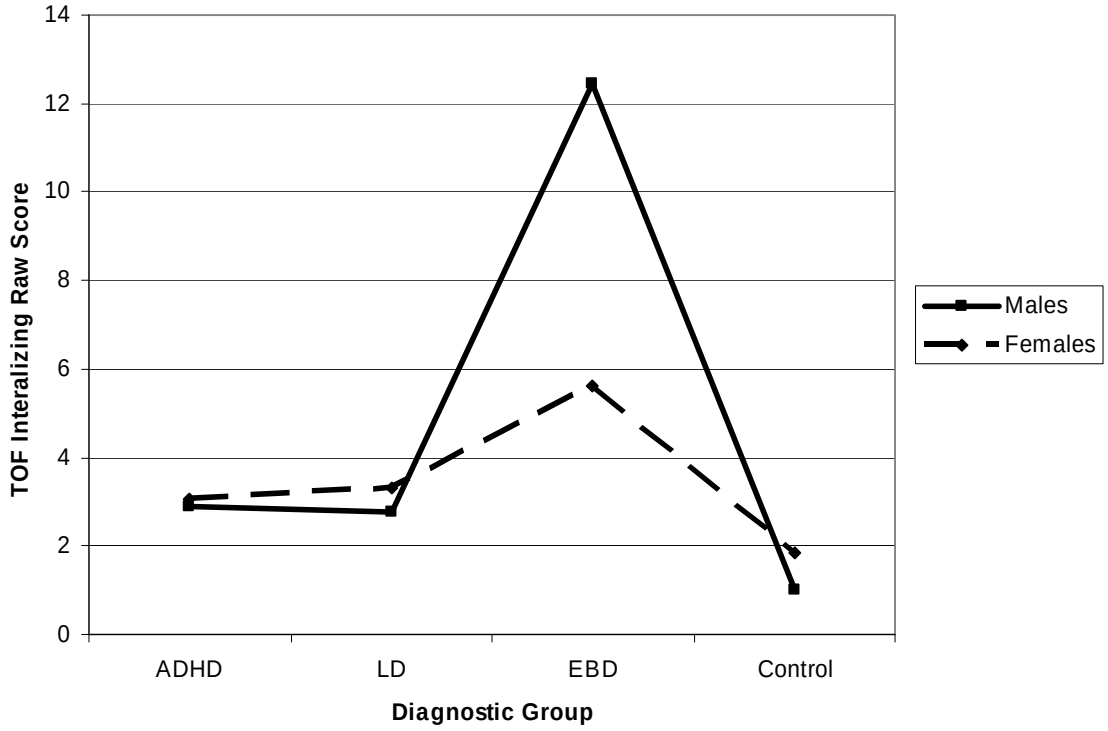


Figure 12.

TOF Internalizing Raw Scores by Gender and Diagnostic Group.

Note. N = 412.

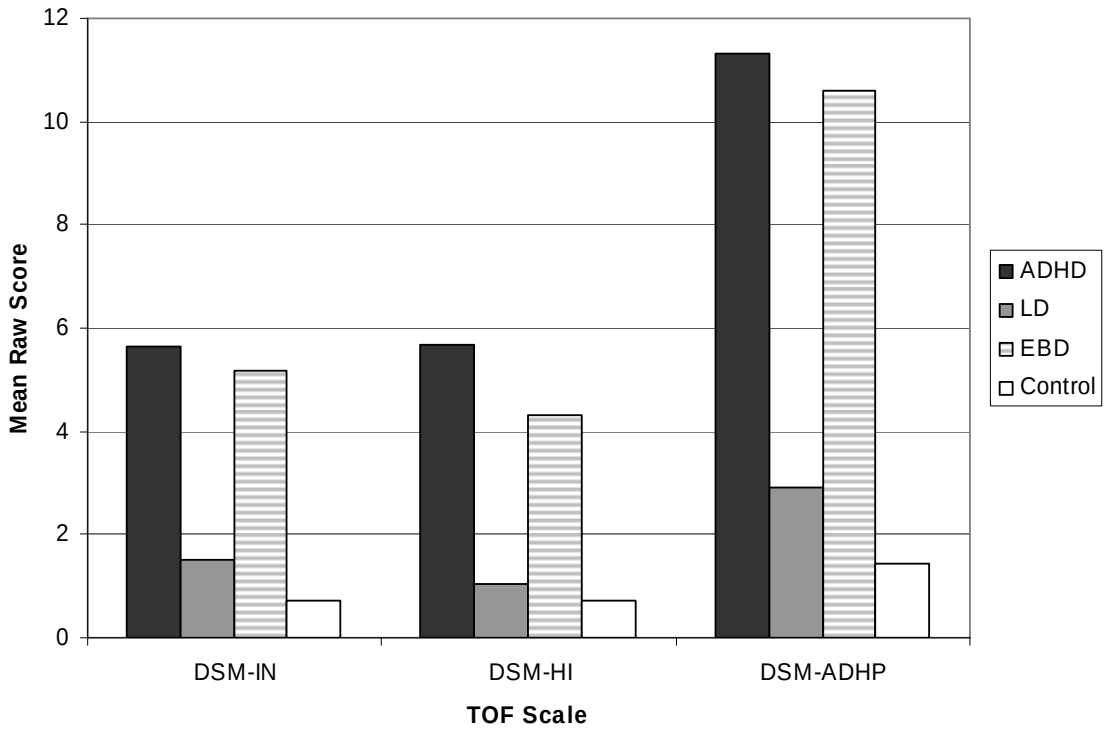


Figure 13.

Raw Scores on DSM-oriented Hyperactivity-Impulsivity, DSM-oriented Inattentive, and DSM-oriented Attention Deficit/Hyperactivity Problems Total Score by Diagnostic Group.

Note. N = 412. DSM-IN = DSM Inattentive, DSM-HI = DSM Hyperactivity-Impulsivity, DSM-ADHP = DSM Attention Deficit/Hyperactivity Problems Total.

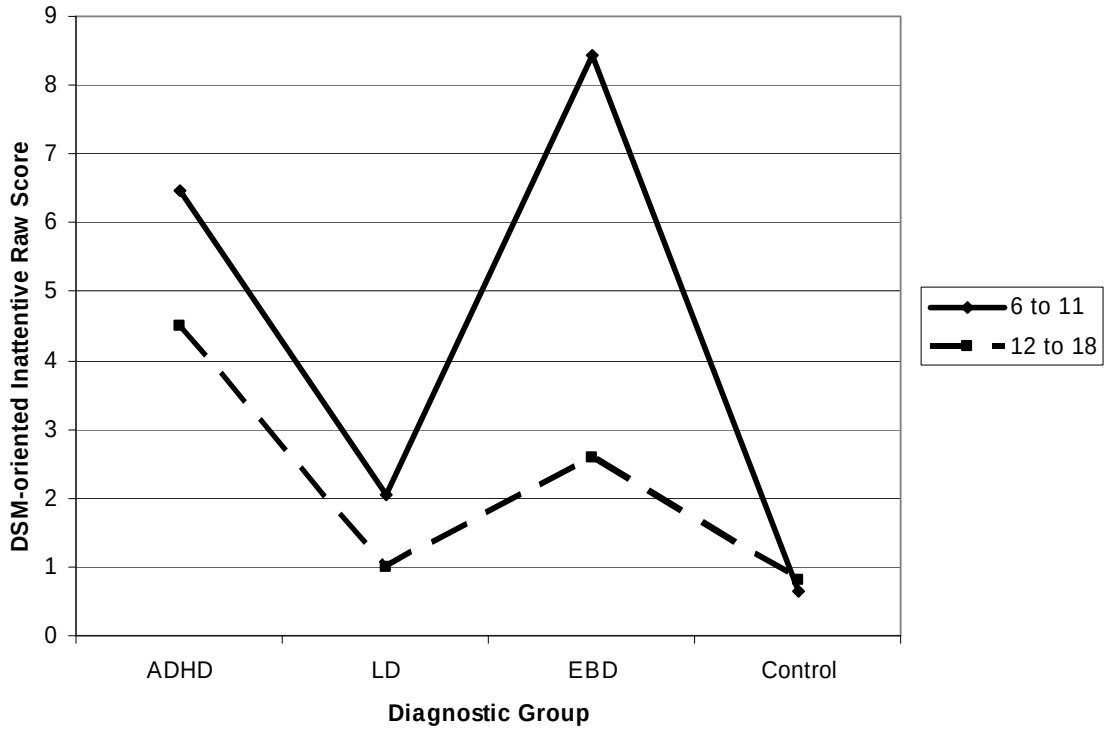


Figure 14.

DSM-oriented Inattentive Raw Score by Age Group and Diagnostic Group.

Note: N = 412.

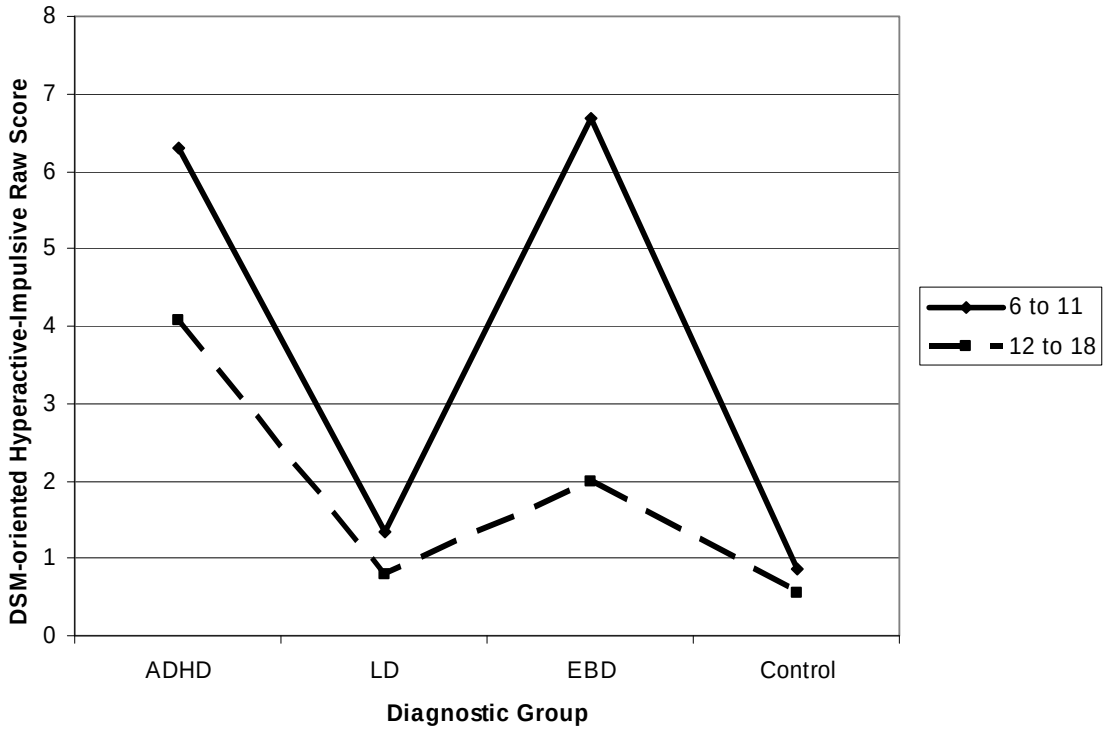


Figure 15.

DSM-oriented Hyperactivity-Impulsivity Raw Score by Age Group and Diagnostic Group.

Note. N = 412.

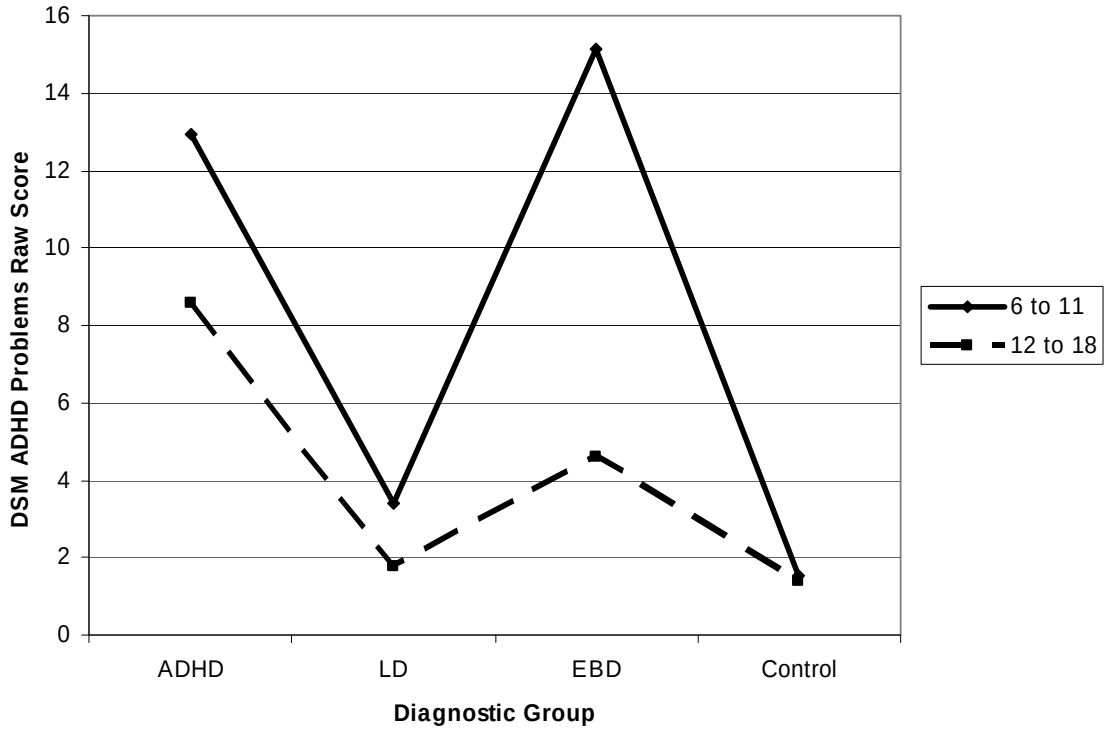


Figure 16.

DSM-oriented Attention Deficit/Hyperactivity Problems Total Raw Scores by Age Group and Diagnostic Group.

Note. N = 412.

REFERENCES

Achenbach, T.M. (2005). Advancing assessment of children and adolescents: Commentary on evidence-based assessment of child and adolescent disorders. *Journal of Clinical Child and Adolescent Psychology, 34,* 541-547.

Achenbach, T.M., & Rescorla, L.A. (2001). *Manual for the ASEBA School-Age Forms & Profiles*. Burlington, VT: University of Vermont, Research Center for Children, Youth, and Families.

Achenbach, T.M. (1991). *Manual for the Child Behavior Checklist/4-18 and 1991 Profile*. Burlington, VT: University of Vermont, Department of Psychiatry.

Achenbach, T.M., McConaughy, S.H., & Howell, C. (1987). Child/adolescent behavioral and emotional problems: Implications of cross-informant correlations for situational specificity. *Psychological Bulletin, 101,* 213-232.

Adams, W., & Sheslow, D. (1990). *Wide range assessment of memory and learning*. Wilmington, Delaware: Wide Range.

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

Anastopolous, A. D., Spisto, M. A., & Maher, M. C. (1994). The WISC-III Freedom from Distractibility factor: Its utility in identifying children with attention deficit hyperactivity disorder. *Psychological Assessment, 6,* 368-371.

Anderson, V., & Stanley, G. (1992). Ability profiles of learning disabled children. *Australian Psychologist, 27,* 48-51.

Andreou, G., Agapitou, P., & Karapetsas, A. (2005). Verbal skills in children with ADHD. *European Journal of Special Needs Education, 20,* 231-238.

Barkley, R.A. (1996). Clinical use of the third factor—proceed with caution. The *ADHD Report, 4*, 6-8.

Barkley, R.A., DuPaul, G.J., & McMurray, M.B. (1990). Comprehensive evaluation of attention deficit disorder with and without hyperactivity as defined by research criteria. *Journal of Consulting and Clinical Psychology, 58,* 775-789.

Bridgett, D.J., & Walker, M.E. (2006). Intellectual functioning in Adults with ADHD: A meta-analytic examination of full scale IQ differences between adults with and without ADHD. *Psychological Assessment, 18,* 1-14.

Calhoun, S.L, & Mayes, S.D. (2005). Processing speed in children with clinical disorders. *Psychology in the Schools, 42*, 333-343.

Canivez, G.L. (1996). Validity and diagnostic efficiency of the K-BIT in reevaluation students with learning disabilities. *Journal of Psychoeducational Assessment, 14,* 4-19.

Carter, B. D., Zelko, F. A., Oas, P. T., & Waltonen, S. (1990). A comparison of ADD/H children and clinical controls on the Kaufman Assessment Battery for Children (K-ABC). *Journal of Psychoeducational Assessment, 8*, 155-164.

Cohen, N.J. (1997). *Children's Memory Scale*. San Antonio, TX: The Psychological Corporation.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Lawrence Earlbaum Associates.

Connery, S.L., Katz, D., Kaufman, A.S., & Kaufman, N.L. (1996). Correlations between two short cognitive tests and a WISC-III short form using a sample of adolescent inpatients. *Psychological Reports, 78*, 1373-1378.

Culbertson, J.L., & Edmonds, J.E. (1996). Learning disabilities. In R.L. Adams, O.A. Parsons, J.L. Culbertson, & S.J. Nixon (Eds.), *Neuropsychology for clinical practice: Etiology, assessment, and treatment of common neurological disorders*. Washington, DC: American Psychological Association.

Daleiden, E., Drabman, R. S., & Benton, J. (2002). The Guide to the Assessment of Test Session Behavior: Validity in relation to cognitive testing and parent-reported behavior problems in a clinical sample. *Journal of Clinical Child & Adolescent Psychology, 31*, 263-271.

Demaray, M. K., Schaefer, K., & Delong, L. K. (2003). Attention-Deficit/Hyperactivity Disorder (ADHD): A national survey of training and current assessment practices in the Schools. *Psychology in the Schools, 40*, 583-597.

Doll, B., & Boren, R. (1993). Performance of severely language-impaired students on the WISC-II, language scales and academic achievement measures. *Journal of Psychoeducational Assessment, 81*, 77-86.

Edwards, M.C. (2005). Agreeing about disagreements: A personal reflection on Achenbach, McConaughy, and Howell (1987). *Clinical Child Psychology and Psychiatry, 10*, 440-445.

Elliot, C.D. (1990). *DAS administration and scoring manual*. San Antonio, TX: The Psychological Corporation.

Faraone, S. V., Biederman, J., Lehman, B. K., Spencer, T., & et al. (1998). Intellectual performance and school failure in children with attention deficit hyperactivity disorder and in their siblings. *Journal of Abnormal Psychology, 102*, 616-623.

Filippatou, D.N., & Livaniou, E.A. (2005). Comorbidity and WISC-III profiles of Greek children with attention deficit hyperactivity disorder, learning disability, and language disorders. *Psychological Reports, 97*, 485-504.

Ford, L., Floyd, R.G., Keith, T.Z., Fields, C., & Schrank, F.A. (2003). Using the Woodcock-Johnson III Tests of Cognitive Ability with students with attention deficit/hyperactivity disorder. In F.A. Schrank & D.P. Flanagan (Eds.), *Woodcock-Johnson III Clinical Use and Interpretation* (pp. 319-344). San Diego: CA: Academic Press.

Frisby, C.L., & Osterlind, S.J. (2006). A descriptive analysis of Test Session Observation Checklist Ratings from the Woodcock Johnson III standardization sample. *Journal of Psychoeducational Assessment, 24*, 342-357.

Gathercole, S.E., Alloway, T.P., Willis, C. & Adams, A. (2006). Working memory in children with reading disabilities. *Journal of Experimental Child Psychology, 93*, 265-281.

Glutting, J. J., & McDermott, P. A. (1998). Generality of test session observations to kindergarteners' classroom behavior. *Journal of Abnormal Child Psychology, 16*, 527-537.

Glutting, J. J., McDermott, P. A., Watkins, M. M., Kush, J. C., & Konold, T.R. (1997). The base rate problem and its consequences for interpreting children's ability

profiles. *School Psychology Review, 26*, 176-188.

Glutting, J. J., Oakland, T., & Konold, T. R. (1994). Criterion-related bias with the Guide to the Assessment of Test-Session Behavior for the WISC-III and WIAT: Possible race/ethnicity, gender, and SES effects. *Journal of School Psychology, 32*, 355-369.

Glutting, J. J., Robins, P. M., & de Lancey, E. (1997). Discriminant validity of test observations for children with attention deficit/hyperactivity. *Journal of School Psychology, 35*, 391-401.

Glutting, J. J., Youngstrom, E. A., Oakland, T., & Watkins, M. W. (1996). Situational specificity and generality of test behaviors for samples of normal and referred children. *School Psychology Review, 25*, 94-107.

Glutting, J.J., & Oakland, T. (1993). *Manual for the Guide to the Assessment of Test Session Behavior*. San Antonio: TX: Psychological Corporation.

Gordon, M., DiNiro, D., Mettelman, B.B., & Tallmadge, J. (1989). Observations of test behavior, quantitative scores, and teacher ratings. *Journal of Psychoeducational Assessment, 7*, 141-147.

Groth-Marnat, G. (1997). *Handbook of Psychological Assessment* (3rd ed.). New York: Wiley.

Hale, J.B., Fiorello, C.A., Kavanagh, J.A., Holdnack, J.A., & Aloe, A.M. (2007). Is the demise of IQ interpretation justified? A response to Special Issue authors. *Applied Neuropsychology, 14*, 37-51.

Hale, R.L., & Saxe, J.E. (1983). Profile analysis of the Wechsler Intelligence Scale for Children-Revised. *Journal of Psychoeducational Assessment, 1*, 155-162.

Hintze, J.M (2005). Psychometrics of direct observation. *School Psychology Review, 34*, 507-519.

Hosp, J.L., & Reschly, D.J. (2002). Regional differences in school psychology practice. *School Psychology Review, 31*, 11-29.

Individuals with Disabilities Education Improvement Act of 2004. 20, U.S.C. Section 1400, H.R. 135.

Javorsky, J. (1993). The relationship between the Kaufman Brief Intelligence Test and the WISC-III in a clinical sample. *Diagnostique, 19*, 377-385.

Kaplan, C. (1993). Reliability and validity of test session behavior observations: Putting the horse before the cart. *Journal of Psychoeducational Assessment, 11,* 314-322.

Kaufman, A.S. (1994). *Intelligent testing with the WISC-III*. New York: Wiley.

Kaufman, A.S., & Kaufman, N.L. (1983). *Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A.S., & Lichtenberger (2000). *Essentials of WISC-III and WPPSI-R assessment*. New York: Wiley.

King, C., & Young, R.D. (1982). Attention deficits with and without hyperactivity: Teacher and peer perceptions. *Journal of Abnormal Child Psychology, 10*, 483-496.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). External validity of the profile variability index for the K-ABC, Stanford-Binet, and WISC--R: Another

cul-de-sac. *Journal of Learning Disabilities, 26*, 557-567.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1993). Relative usefulness of elevation, variability, and shape information from WISC--R, K-ABC, and Fourth Edition Stanford-Binet profiles in predicting achievement. *Psychological Assessment, 4*, 426-432.

Konold, T. R., Glutting, J. J., Oakland, T., & O'Donnell, L. (1995). Congruence of test-behavior dimensions among child groups that vary in gender, race-ethnicity, and SES. *Journal of Psychoeducational Assessment, 13*, 111-119.

Kramer, J. J., Henning-Stout, M., Ullman, D. P., & Schellenberg, R. P. (1987). The viability of scatter analysis on the WISC--R and the SBIS: Examining a vestige. *Journal of Psychoeducational Assessment, 5*, 37-47.

Krane, E., & Tannock, R. (1992). WISC-III third factor indexes learning problems but not attention deficit/hyperactivity disorder. *Journal of Attention Disorders, 5*, 69-78.

Landau, S., & Swerdlik, M.E. (2005). What you see is what you get: A commentary on school-based direct observation systems. *School Psychology Review, 34*, 529-536.

Lavin, C. (1996). Scores on the WISC-III and Woodcock-Johnson Tests of Achievement-Revised for a sample of children with emotional handicaps. *Psychological Reports, 79*, 1291-1295.

Lipsitt, P. D., Buka, S. L., & Lipsitt, L. P. (1990). Early intelligence scores and subsequent delinquency: A prospective study. *American Journal of Family Therapy, 18*, 197-208.

Mahone, E., Miller, T. L., Koth, C. W., Mostofsky, S. H., Goldberg, M. C., & Denckla, M. B. (2003). Differences between WISC-R and WISV-III performance scale among children with ADHD. *Psychology in the Schools, 40*, 331-340.

Mash, E.J., & Hunsley, J. (2005). Evidence-based assessment of child and adolescent disorders: Issues and challenges. *Journal of Child and Adolescent Psychology, 34*, 362-379.

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (1998a). WISC-III Freedom from Distractibility as a measure of attention in children with and without attention deficit hyperactivity disorder. *Journal of Attention Disorders, 2*, 217-227.

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (1998b). WISC-III profiles for children with and without learning disabilities. *Psychology in the Schools, 35*, 309-316.

McConaughy, S.H. (2005). Direct observational assessment during test sessions and child clinical interviews. *School Psychology Review, 34*, 490-506.

McConaughy, S.H., & Achenbach, T.M. (2004). *Manual for the Test Observation Form for Ages 2-18*. Burlington, VT: University of Vermont, Research Center for Children, Youth & Families.

McConaughy, S.H., Mattison, R.E., & Peterson, R. (1994). Behavioral/emotional problems of children with serious emotional disturbances and learning disabilities. *School Psychology Review, 23*, 81-98.

McDermott, P. A., & Glutting, J. J. (1997). Informing stylistic learning behavior, disposition, and achievement through ability subtests: or More illusions of meaning? *School Psychology Review, 26*, 163-175.

McDermott, P. A., Fantuzzo, J. W., & Glutting, J. J. (1990). Just say no to subtest analysis: A critique on Wechsler theory and practice. *Journal of Psychoeducational Assessment, 8,* 290-302.

McDermott, P.A., Glutting, J.J., Jones, J.N., Watkins, M.W., & Kush, J. (1989). Core profile types in the WISC-R national sample: Structure, membership, and applications. *Psychological Assessment, 1*, 292-299.

McInnes, A., Humphries, T., Hogg-Johnson, S., & Tannock, R. (2003). Listening comprehension and working memory are impaired in attention-deficit hyperactivity disorder irrespective of language impairment. *Journal of Abnormal Child Psychology, 31,* 427-443.

McKevitt, B.C., & Elliott, S.N. (2005). Observations and ratings of preschool children's social behavior: Issues of representativeness and validity. *Psychology in the Schools, 42*, 13-26.

McMahon, R.J., & Frick, P.J. (2005). Evidence-based assessment of conduct problems in children and adolescents. *Journal of Child and Adolescent Psychology, 34,* 477-505.

Mealer, C., Morgan, S., & Luscomb, R. (1996). Cognitive functioning of ADHD and non-ADHD boys on the WISC-III and WRAML: An analysis within a memory model. *Journal of Attention Disorders, 1*, 133-145.

Milich, R., Baletine, A.C., & Lynam, D.R. (2001). ADHD combined type and ADHD predominantly inattentive types are distinct and unrelated disorders. *Clinical Psychology: Science and Practice, 8*, 463-487.

Mleko, A.L., & Burns, T.G. (2005). Stanford-Binet Intelligence Scales [Test Review]. *Applied Neuropsychology, 12*, 179-180.

Naglieri, J. A., Goldstein, S., Iseman, J. S., & Schwebach, A. (2003). Performance of children with attention deficit hyperactivity disorder and anxiety/depression on the WISC-III and Cognitive Assessment System (CAS). *Journal of Psychoeducational Assessment, 21*, 32-42.

Naglieri, J.A., & Das, J.P. (1997). *Cognitive Assessment System*. Itasca, IL: Riverside Publishing.

Oakland, T., & Glutting, J. J. (1990). Examiner observations of children's WISC--R test-related behaviors: Possible socioeconomic status, race, and gender effects. *Psychological Assessment, 2*, 86-90.

Oakland, T., Broom, J., & Glutting, J. (2000). Use of Freedom from Distractibility and Processing Speed to assess children's test-taking behaviors. *Journal of School Psychology, 38*, 469-475.

Ottem, E. (2002). Do the Wechsler scales underestimate the difference between verbal and performance abilities in children with language-related disorders? *Scandinavian Journal of Psychology, 43*, 291-298.

Pelham, W.E., Fabiano, G.A., & Massetti, G.M (2005). Evidence-based assessment of Attention Deficit Hyperactivity Disorder in children and adolescents. *Journal of Child and Adolescent Psychology, 34*, 449-476.

Prewett, P.N., & Matavich, M.A. (1993). A comparison of referred students' performance on the WISC-III and the Stanford-Binet Intelligence Scale: Fourth Edition. *Journal of Psychoeducational Assessment, 81*, 142-148.

Preiss, M., & Lenka, F. (2006). Depressive symptoms, academic achievement, and intelligence. *Studia Psychologica, 48*, 57-67.

Prifitera, A., & Dersh, J. (1993). Base rates of WISC-III diagnostic subtest patterns among normal, learning-disabled, and ADHD samples. In B.A. Bracken & R.S. McCallum (Eds.), *Wechsler Intelligence Scale for Children: Third Edition* (pp 43-55). Brandon, VT: Clinical Psychology Publishing Co.

Reid, D.K., Hresko, W.P., & Swanson, H.L. (1996). *Cognitive approaches to learning disabilities*. Austin, TX: Pro-Ed.

Riccio, C. A., & Hynd, G. W. (2000). Measurable biological substrates to verbal-performance differences in Wechsler scores. *School Psychology Quarterly, 15*, 386-399.

The Riverside Publishing Company (2004). *Frequently asked Questions about the Stanford-Binet Intelligence Scales, Fifth Edition*. No longer available from original source; retrieved July, 2007 from http://www.assess.nelson.com/pdf/sb5-faq.pdf.

Roid, G.H. (2003a). *Stanford-Binet Intelligence Scales, Fifth Edition*. Itasca, IL: Riverside Publishing.

Roid, G.H. (2003b). *Stanford-Binet Intelligence Scales, Fifth Edition, Interpretive manual: Expanded guide for the interpretation of SB5 test results*. Itasca, IL: Riverside Publishing.

Rose, J. C., Lincoln, A. J., & Allen, M. H. (1992). Ability profiles of developmental language disordered and learning disabled children: A comparative analysis. *Developmental Neuropsychology, 8,* 413-426.

Sakoda, J.M., Cohen, B.H., & Beall, G. (1954). Test of significance for a series of statistical tests. *Psychological Bulletin, 51*, 172-175.

Saklofske, D.H., Schwean, V.L., & O'Donnell, L. (1995). WIAT performance of children with ADHD. *Canadian Journal of School Psychology, 12,* 55-59.

Saklofske, D.H., Schwean, V.L., Yackulic, R.A., & Quinn, D. (1994). WISC-III and SB:FE performance of children with attention deficit hyperactivity disorder. *Canadian Journal of School Psychology, 10*, 167-171.

Sattler, J.M. (1998). *Assessment of children* (3rd ed.). San Diego, CA: Author.

Sattler, J.M. (2001). *Assessment of Children: Cognitive Applications* (4th ed.). La Mesa, CA: Author.

Schuck, S.E.B., & Crinella, F.M. (2005). Why children with ADHD do not have low IQs. *Journal of Learning Disabilities, 38,* 262-280.

Schwean, V.L., & Saklofske, D.H. (2005). Assessment of attention deficit hyperactivity disorder with the WISC-IV. In A. Prifitera, D. Saklofske, & L. Weiss (Eds.), *WISC-IV clinical use and interpretation: Scientist-practitioner perspectives* (pp. 235-280). San Diego, CA: Elsevier Academic Press.

Schwean, V.L., Saklofske, D.H., Yackulic, R.A., & Quinn, D. (1993). WISC-III performance of ADHD children. In Bracken, B.A. & McCallum, R. S. (Eds.), *Wechsler Intelligence Scale for Children: Third Edition* (pp 56-70). Brandon, VT: Clinical Psychology Publishing Co.

Semrud-Clikeman, M., Hynd, G. W., Lorys, A. R., & Lahey, B. B. (1998). Differential diagnosis of children with ADHD and ADHD/with co-occurring conduct disorder. *School Psychology International, 14*, 361-370.

Shapiro, E. G., Hughes, S. J., August, G. J., & Bloomquist, M. L. (1993). Processing of emotional information in children with attention-deficit hyperactivity disorder. *Developmental Neuropsychology, 9,* 207-224.

Shapiro, E.S., & Heick, P. (2004). School psychologist assessment practices in the evaluation of students referred for social/behavioral/emotional problems. *Psychology in the Schools, 41*, 551-561.

Silverman, W.K., & Ollendick, T.H. (2005). Evidence-based assessment of anxiety and its disorders in children and adolescents. *Journal of Clinical Child and Adolescent Psychology, 34,* 380-411.

Slate, J.R., & Jones, C.H. (1995). Relationship of the WISC-III and WISC-R for students with specific learning disabilities and mental retardation. *Diagnostique, 21,* 9-17.

Stevens, J., Quittner, A. L., Zuckerman, J. B., & Moore, S. (2002). Behavioral inhibition, self-regulation of motivation, and working memory in children with attention deficit hyperactivity disorder. *Developmental Neuropsychology, 21,* 117-140.

Swanson, H.L. (2005). Working memory, intelligence, and learning disabilities. In O.

Wilhelm & R. Engle (Eds.), *Handbook of understanding and measuring intelligence*. Thousand Oaks, CA: Sage Publications.

Teicher, M. H., Ito, Y., Glod, C. A., & Barber, N. I. (1996). Objective measurement of hyperactivity and attentional problems in ADHD. *Journal of the American Academy of Child & Adolescent Psychiatry, 35*, 334-342.

Terman, L.M., & Merrill, M.A. (1960). *Stanford-Binet Intelligence Scale*. Boston: Houghton-Mifflin.

The Psychological Corporation. (1992). *Wechsler Individual Achievement Test manual*. San Antonio, TX: The Psychological Corporation.

Thorndike, R.L., Hagen, E.P., & Sattler, J.M. (1986). *Stanford-Binet Intelligence Scale: Fourth Edition*. Boston: Houghton Mifflin.

Vile Junod, R. E., DuPaul, G.J., Jitendra, A.K., Volpe, R.J., & Cleary, K.S. (2006). Classroom observations of students with and without ADHD: Differences across types of engagement. *Journal of School Psychology, 44*, 87-104.

Volpe, R.J., & McConaughy, S.H. (2005). Systematic direct observational assessment of student behavior: Its use and interpretation in multiple settings. *School Psychology Review, 34*, 451-453.

Wallbrown, F. H., Vance, H. B., & Blaha, J. (1979). Developing remedial hypotheses from ability profiles. *Journal of Learning Disabilities, 12*, 557-561.

Ward, S.B., Ward, T.J., Hart, C. V., Young, D.L, & Mollner, N.R. (1995). The incidence and utility of the ACID, ACIDS, and SCAD profiles in a referred population. *Psychology in the Schools, 32*, 267-276.

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997). Prevalence and diagnostic utility of the WISC-III SCAD profile among children with disabilities. *School Psychology Quarterly, 12*, 235-248.

Watkins, M.W., & Glutting, J.J. (2000). Incremental validity of WISC-III profile elevation, scatter, and shape information for predicting reading and math achievement. *Psychological Assessment, 12*, 402-408.

Wechsler, D. (1974). *Manual for the Wechsler Preschool and Primary Scale of Intelligence (WPPSI)*. New York: The Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children – Third Edition (WISC-III)*. San Antonio, TX: The Psychological Corporation.

Wilson, M.S., & Reschly, D.J. (1996). Assessment in school psychology training and practice. *School Psychology Review, 25*, 9-23.

Woodcock, R.W., & Johnson, M.B. (1989). *Woodcock-Johnson Psycho-Educational Battery-Revised*. Itasca, IL: Riverside Publishing.

Woodcock, R.W., McGrew, K., & Mather, N. (2001). *Woodcock-Johnson-III*. Itasca, IL: Riverside Publishing.

Zimmerman, I.L., & Woo-Sam, J.M. (1997). Review of the criterion-related validity of the WISC-III: The first five years. *Perceptual and Motor Skills, 85*, 531-546.

APPENDIX A

Organization of the SB5 by Factor and Domain

Factors	Domains: Verbal	Domains: Nonverbal
Fluid Reasoning	Early Reasoning (Levels 2-3) Verbal Absurdities (Level 4) Verbal Analogies (Levels 5-6)	Object Series/Matrices (routing)
Knowledge	Vocabulary (routing)	Procedural Knowledge (Levels 2-3) Picture Absurdities (Levels 4-6)
Quantitative Reasoning	Verbal Quantitative Reasoning	Nonverbal Quantitative Reasoning
Visual-Spatial Processing	Position and Direction	Form Board (Levels 1-2) Form Patterns (Levels 3-6)
Working Memory	Delayed Response (Level 1) Block Span (Levels 2-6)	Delayed Response (Level 1) Block Span (Levels 2-6)