Journal of Applied School Psychology, 28:133–154, 2012

ISSN: 1537-7903 print / 1537-7911 online
DOI: 10.1080/15377903.2012.669743

Routledge
Taylor & Francis Group

Diagnostic Utility of WISC-IV General Abilities Index and Cognitive Proficiency Index Difference Scores Among Children With ADHD

SARAH E. DEVENA
Arizona State University, Tempe, Arizona, USA

MARLEY W. WATKINS
Baylor University, Waco, Texas, USA

*The Wechsler Intelligence Scale for Children-Fourth Edition General Abilities Index and Cognitive Proficiency Index have been advanced as possible diagnostic markers of attention deficit hyperactivity disorder. This hypothesis was tested with a hospital sample with attention deficit hyperactivity disorder (*n = *78), a referred but nondiagnosed hospital sample (*n = *66), a school sample with attention deficit hyperactivity disorder (*n = *196), a school matched comparison sample (*n = *196), and a simulated standardization sample (*n = *2,200). On the basis of receiver operating characteristic analyses, the General Abilities Index-Cognitive Proficiency Index discrepancy method had an area under the curve of (a) .64, 95% CI [0.58, 0.71] for the hospital attention deficit hyperactivity disorder sample compared with the simulated normative sample, (b) .46, 95% CI [0.37, 0.56] for the hospital attention deficit hyperactivity disorder sample compared with the referred but nondiagnosed hospital sample, (c) .63, 95% CI [0.59, 0.67] for the school attention deficit hyperactivity disorder sample compared with the simulated sample, and (d) .50, 95% CI [0.45, 0.56] for the school attention deficit hyperactivity disorder sample compared with the matched comparison sample. These area-under-the-curve values indicate that the General Abilities Index-Cognitive Proficiency Index discrepancy method has low accuracy in identifying children with attention deficit hyperactivity disorder.*

This article is based on a thesis by the first author (S.E.D.). Preliminary results were presented at the 2010 meeting of the National Association of School Psychologists. The authors acknowledge the assistance of Michael S. Lavoie, Ph.D.

Address correspondence to Sarah E. Devena, School Psychology Program, Arizona State University, P.O. Box 871811, Tempe, AZ 85287, USA. E-mail: sarah.devena@asu.edu

KEYWORDS Attention Deficit Hyperactivity Disorder, general abilities index, cognitive proficiency index, Wechsler Intelligence Scale for Children-Fourth Edition, diagnostic utility

Attention deficit hyperactivity disorder (ADHD) is a developmental disorder distinguished by behavioral impulsivity and difficulties with goal-directed thoughts and processes (Schwean & McCrimmon, 2008). According to the Centers for Disease Control and Prevention (2005), ADHD is currently one of the most common neurobehavioral disorders of children with 3–7% of school-aged children diagnosed with the disorder (Adams, Lucas, & Barnes, 2008). ADHD can have a profound effect on academic achievement and future career success (Frazier, Youngstrom, Glutting, & Watkins, 2007) so an accurate diagnosis is crucial to ensure appropriate help for students in need and to remove the risk of misdiagnoses for nondisabled students (Skounti, Philalithis, & Galanakis, 2007).

Various methods are used to diagnose ADHD and can include the following: (a) direct observations (DuPaul, 1992), (b) structured interviews (Power & Ikeda, 1996), (c) behavior rating scales (Barkley, 1991), (d) multiple stage evaluation (DuPaul, 1992), and (e) cognitive profiles (Prifitera & Dersh, 1993). Although structured interviews and behavior rating scales are considered best practice for the identification of ADHD (American Academy of Child and Adolescent Psychiatry, 2007), analysis of cognitive profiles has also been recommended (Prifitera & Dersh, 1993) because cognitive tests measure abilities, such as working memory, which are considered to be theoretical underpinnings of ADHD (Schwean & McCrimmon, 2008). Some researchers suggest that cognitive profiles are useful in understanding the cognitive strengths and weaknesses of children that can, therefore, contribute to treatment planning (Kaufman, 1994). For example, clinicians might use processing speed interventions for children with ADHD profiles (Schwean & McCrimmon). Because cognitive test use is widespread in school assessments (Wilson & Reschly, 1996), and profiles can provide additional information for assessment (Schwean & McCrimmon), they warrant further investigation.

Of all the available cognitive tests for children, the Wechsler series is the most popular with clinicians (Kaufman & Lichtenberger, 2000) and the Wechsler Intelligence Scale for Children–Fourth Edition (WISC-IV; Wechsler, 2003a) is the most widely used measure of children's intelligence. Many clinicians believe that the Wechsler tests, beyond their popularity, are valuable instruments for the diagnostic assessment of children (Weiss, Beal, Saklofske, Alloway, & Prifitera, 2008). Accordingly, clinicians sometimes use the Wechsler tests to detect ADHD in children by examining specific score patterns that have been identified through research as markers of ADHD (Sattler, 2008).

Past research has shown three main cognitive subtest score patterns linked to ADHD. First, Kaufman (1994) found a profile of low scores on the arithmetic, coding, and digit span subtests on the Wechsler Intelligence Scale for Children–Revised (WISC-R; Wechsler, 1974) which was labeled the Freedom from Distractibility (FD) factor. With the introduction of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1991), the FD index was created which included just the arithmetic and digit span subtests. When children scored high on this FD index it was thought to indicate the ability to sustain attention and when they scored low on this FD index it was thought to indicate distractibility (Kaufman). Because of this hypothesis, low scores on the FD index were considered a possible indicator of ADHD.

Research on the WISC-III standardization sample subsequently showed that children with ADHD scored lower on the FD index subtests than on the other subtests (Wechsler, 1991). For instance, Mayes, Calhoun, and Crowell (1998) reported that 23% of children with ADHD ($n = 87$) had digit span and arithmetic as two of their three lowest scores whereas none of the non-ADHD children ($n = 32$) showed this pattern. Moreover, the FD index was significantly lower than the children's full-scale IQ (FSIQ) for the ADHD sample. Additional research with groups of children with and without ADHD found that, on average, scores of the ADHD groups on those two subtests were significantly lower than the scores for non-ADHD groups (Anastopoulos, Spisto, & Maher, 1994; Wielkiewicz, 1990).

The coding and symbol search subtests of the WISC-III were added to the two subtests of the FD index to yield the second major Wechsler score pattern associated with ADHD. This score pattern included lower scores on the symbol search, coding, arithmetic, and digit (SCAD) Span subtests; the acronym *SCAD* was coined for this profile (Kaufman, 1994). Research with the WISC-III standardization sample indicated that children with learning disabilities had lower scores on this profile (Prifitera & Dersh, 1993). Mayes et al. (1998) supported the validity of this cognitive pattern by finding the SCAD profile in the majority of their sample of children with ADHD. In their analysis, 87% of children were correctly identified as having ADHD if their SCAD scores were lower than their other core subtest scores compared with 47% in the non-ADHD group.

The third and final Wechsler score pattern associated with ADHD included lower scores on the arithmetic, coding, information, and digit span (ACID) subtests (Joschko & Rourke, 1985; Snow & Sapp, 2000). The ACID profile incorporated the Information subtest along with the original three subtests in the FD factor to enhance diagnostic accuracy. Research on clinical versus nonclinical groups indicated that the ACID profile occurred in 12% of children with ADHD compared with only 1% of children from the non-ADHD group (Prifitera & Dersh, 1993). These findings led Prifitera and Dersh (1993) to propose that the ACID profile could be useful for diagnostic

purposes. In a later study, 6% of children with ADHD also exhibited the ACID profile (Swartz, Gfeller, Hughes, & Searight, 1998). However, Swartz et al. (1998) found no significant difference between the ADHD and LD samples in the frequency of ACID profiles.

Although the FD, SCAD, and ACID profiles initially appeared to be valid markers of ADHD, there are two substantial limitations to this research. The first limitation is the focus on subtest scores. Subtest scores have relatively weak internal consistency, especially when compared with index scores, which are composites of multiple subtests that measure the same underlying cognitive construct. For example, in the WISC-IV normative sample the median internal consistency for subtests is .86, compared with .88 to .94 for the composite scores (Wechsler, 2003b). Furthermore, the stability of subtest scores is weak. For example, the median stability coefficients of WISC-IV subtest and composite scores for a small sample ($n = 43$) of elementary and middle school students across an 11-month interval were .51 and .73, respectively (Ryan, Glass, & Bartels, 2010). Likewise, the long term stability of WISC-III subtest scores among a large clinical sample was found to be considerably weaker than the composite scores derived from multiple subtests with median coefficients of .68 versus .87, respectively (Canivez & Watkins, 1998). Moreover, subtest score analysis necessitates the comparison of difference scores. However, the reliability of the difference between two scores is smaller than the reliability of the individual scores, which introduces further error into subtest comparisons (Feldt & Brennan, 1993).

The second limitation to the research supporting subtest score patterns is that researchers often use statistically significant group differences in support of the patterns. In other words, the mean subtest scores of a group of children with ADHD is compared with the mean subtest scores of a group of children without ADHD and statistically significant group differences are declared sufficient for individual diagnosis. Unfortunately, increased distributional overlap of group scores reduces the diagnostic accuracy for individuals. That is, a profile may have discriminant validity but it does not necessarily have clinical utility. As a result, discriminant validity cannot be considered strong evidence at the individual diagnostic level (Watkins, Glutting, & Youngstrom, 2005). This concept is illustrated in Figure 1 which shows a possible score distributional overlap in two hypothetical groups of children. Although, in this case, each group is distinguishable from the other, the distributional overlap illustrates the problem of diagnosing a child based on group mean differences.

In addition to these theoretical limitations, considerable empirical research indicates that subtest patterns are not accurate diagnostic indicators for individual children. For example, in an analysis of the FD profile, Gussin and Javorsky (1995) found that there were no significant differences between ADHD and non-ADHD participants. As a result, the researchers concluded that the FD profile was not a valid predictor of ADHD. Likewise, an analysis

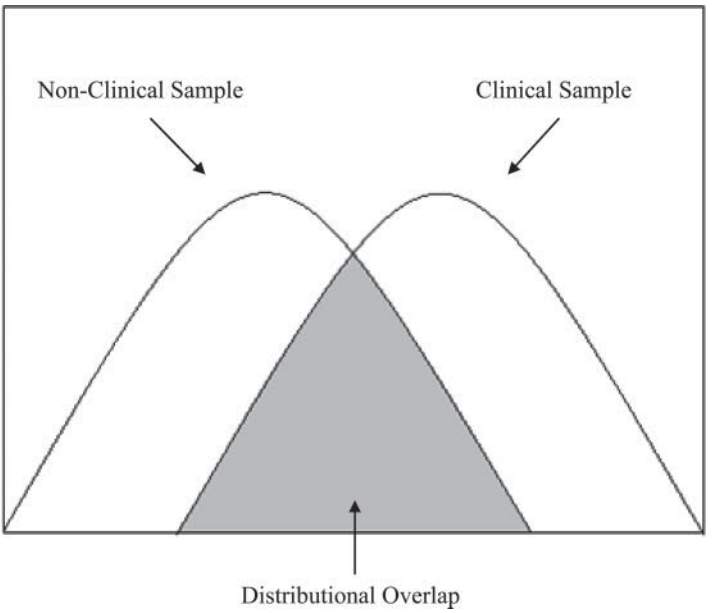

FIGURE 1 Hypothetical mean differences between attention deficit hyperactivity disorder and non–attention deficit hyperactivity disorder groups showing the distributional overlap of the groups in the shaded region.

of the diagnostic accuracy of the SCAD profile among children with disabilities revealed that a randomly selected child with a disability would exhibit a SCAD profile only 59% of the time (Watkins, Kush, & Glutting, 1997a). In addition, in a study to distinguish between children with and without learning disabilities, the ACID profile indicated that a randomly selected child with a learning disability would display an ACID profile only 60% of the time (Watkins, Kush, & Glutting, 1997b). In addition to individual studies, reviews of multiple studies also support the conclusion that subtest patterns are not accurate diagnostic indicators for individual children. For instance, Bray, Kehle, and Hintze (1998) reported that there is overwhelming evidence against using subtest analysis. Another review addressing subtest analysis indicated that subtest profiles did not show an acceptable level of accuracy for diagnostic purposes (Watkins, 2003). Consequently, Sattler (2008) concluded that subtest analysis is not appropriate for clinical diagnoses.

In recognition of the problems with subtest patterns, most current approaches for using cognitive assessments to assist in the diagnosis of ADHD have shifted focus to factor index score patterns. Because the WISC-IV has been shown to have greater sensitivity to ADHD symptoms than the WISC-III and the intended focus of this study is on current approaches, only studies based on the WISC-IV will be addressed. The WISC-IV factor index composites include processing speed (PSI), working memory (WMI), verbal

comprehension (VCI), and perceptual reasoning (PRI). According to Weiss et al. (2008), "differences among the four-factor-based WISC-IV index scores are clinically meaningful and worthy of study within the context of the complete individual" (p. 9).

Following the trend of addressing composite scores, the WISC-IV was administered to 89 children aged from 8 to 13 years who were identified as having ADHD based on the *Diagnostic and Statistical Manual for Mental Disorders* (4th edition; American Psychiatric Association, 2000) diagnostic criteria. The children were selected, on the basis of their availability, from a variety of educational and clinical settings. On average, children with ADHD performed worse on PSI and WMI indexes compared with VCI and PRI indexes (Wechsler, 2003b). The effect size for PSI was moderate (.59) and the effect sizes for VCI, WMI, and FSIQ were small (.26, .38, and .38, respectively). Wechsler indicated that this discrepancy showed that children with ADHD may have typical intelligence levels but they differ from non-ADHD children in their special abilities. However, this study had three major limitations. First, the effect sizes were only small to moderate. This reflects considerable overlap of score distributions and consequently the probability of correctly distinguishing between individuals in the two groups is only slightly higher than chance. Second, FSIQ scores were different between the two groups (children with ADHD average FSIQ was 97.6 versus children without ADHD average FSIQ of 102.7), which may have confounded the results. Third, the sample size was relatively small ($n = 89$) and did not cover the entire age range of the WISC-IV. This restricted age range makes it difficult to determine if children outside of 8 to 13 years of age would display the same score patterns.

Additional research that included 118 children with ADHD whose ages ranged from 6 to 16 years of age was conducted on the WISC-IV index scores to identify ADHD profiles (Mayes & Calhoun, 2006). The VCI and PRI scores, on average, were significantly higher than the WMI and PSI scores for the children with ADHD ($d = 1.6$ to 1.9).WMI and PSI scores were lower than the VCI and PRI scores in 88% of the ADHD cases. Furthermore, all the children with ADHD either had the WMI (55%) or PSI (45%) as their lowest index score. Based upon these results, Mayes and Calhoun concluded that, "If future studies support the enhanced distinctiveness of the low WMI and PSI and high VCI and PRI WISC-IV profile in children with ADHD, this may be diagnostically and clinically useful" (p. 490). However, there are two notable drawbacks to the methods used in this study. First, the sample only included children referred to the researchers' psychiatric clinic, which may have introduced sample or testing bias. A second drawback was that the mean standard scores for the FSIQ, VCI, and PRI in the ADHD sample were

considerably higher than the national average scores (108, 114, and 117, respectively).

Subsequently, the four WISC-IV factor indexes were collapsed into two index scores to better reflect two hypothetical underlying clinical constructs. The four WMI and PSI subtests were combined to form the Cognitive Proficiency Index (CPI; Weiss & Gable, 2007) and the six VCI and PRI subtests were merged to form the General Abilities Index (GAI; Raiford, Weiss, Rolfhus, & Coalson, 2005). The CPI is thought to correspond to how proficiently children process specific types of cognitive information, which in turn facilitates learning and problem solving. In contrast, the GAI is thought to measure intellectual functioning without the influence of working memory and processing speed.

To investigate CPI and GAI differences, clinical and nonclinical groups were selected during the WISC-IV standardization project (Weiss & Gable, 2007). By comparing children's CPI with their GAI, Weiss and Gable attempted to identify a cutoff score that would distinguish between clinical and nonclinical groups with a true positive rate (TPR) and a true negative rate (TNR) of at least 60%. The TPR is the positives that are correctly identified divided by the total positives. On the other hand, the TNR is the negatives that are correctly identified divided by all negatives. In addition, one minus the TNR gives the false positive rate (FPR), which is the negatives that are incorrectly classified divided by the total negatives (Fawcett, 2006). Of the 12 clinical groups analyzed by Weiss and Gable (2007), 4 had high enough TPR and TNR to be considered noteworthy. The learning disabilities group was identified with a TPR of 66% and a TNR of 63% when CPI was lower than GAI by at least 5 points. The closed head traumatic brain injury group was identified with a TPR of 65% and TNR of 61% when CPI was at least 4 points lower than GAI. The open head traumatic brain injury group was identified with a TPR of 67% and TNR of 62% when CPI was at least 4 points lower than GAI. Last, the Asperger's group was identified with a TPR of 68% and TNR of 63% when CPI was at least 11 points lower than GAI. Weiss and Gable (2007) concluded that CPI < GAI discrepancies alone cannot be considered diagnostic markers of most specific disorders but they are implicated in a variety of disorders. Subsequently, Weiss et al. (2008) concluded that GAI-CPI differences that occur in 10% or less of the population (which is equivalent to approximately a 16 point discrepancy) are rare and interpretable.

One problem with Weiss and Gable's (2007) study was that only 4 out of the 12 clinical groups were identified with 60% accuracy, with the highest group only reaching a TPR of 68% and TNR of 63%. This reveals a lack of accurate results for most individuals in the clinical groups. Another problem is that the analysis used the TPR and TNR to identify a specific cutoff score. TPR and TNR values would have differed if other cutoff scores had been

selected. In addition, the TPR and TNR depend on base rates (Elwood, 1993), which means that the TPR and TNR will vary depending on the population or subgroup (i.e., boys vs. girls). Overall, these problems make the analysis unsuitable for accurate estimation of the diagnostic utility of WISC-IV index profiles.

A suitable measure of diagnostic utility should not be dependent upon base rate or cutoff score (Swets, 1988). A receiver operating characteristic (ROC) analysis avoids these issues by using proportions of TPR and FPR that ignore base rates and by looking at all possible scores instead of a single cutoff score (Pintea & Moldovan, 2009). A ROC curve is drawn by plotting individual points for all possible cutoff scores. In other words, plotting the balance between the TPR and the FPR for the test while moving the cutoff score across the full range of values (Fawcett, 2006). The more accurate a test is, the farther the ROC curve will move to the upper left corner of the graph (see Figure 2). Overall, the ROC curve will allow for a complete description of diagnostic performance of a test (Pepe, 2003).

Although WISC-IV factor index scores possess theoretical coherence lacking in subtest scores and are more reliable then subtest scores, research conducted by Mayes and Calhoun (2006) as well as Wechsler (2003b) has not addressed the issue of using group averages to diagnose individuals. In addition, in the research conducted by Weiss and Gable (2007) only the TPR and TNR for one cutoff score were calculated when considering the diagnostic accuracy of CPI < GAI discrepancies. For these reasons, this

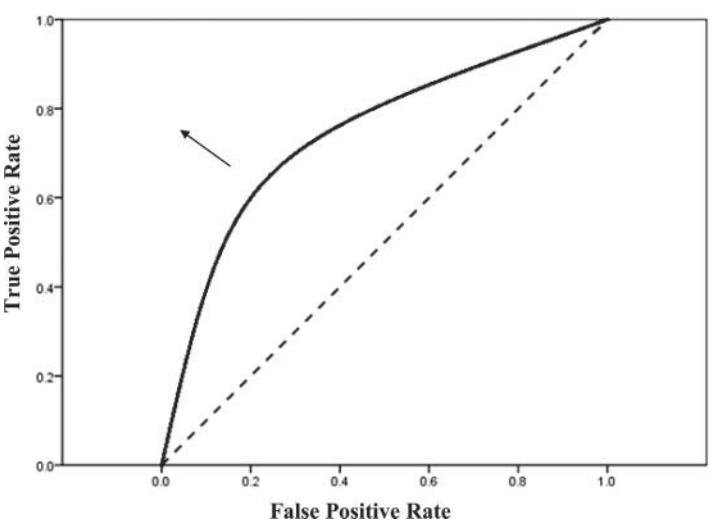

FIGURE 2 Hypothetical receiver operating characteristic curve with diagonal chance line showing that as the curve moves farther toward the left corner of the graph, the more accurate a test is.

study will apply diagnostic utility statistics, including a ROC analysis, to test the ability of WISC-IV GAI-CPI difference scores to identify children with ADHD.

METHOD

Participants

The hospital ADHD sample included 78 children (56 boys, 22 girls) aged 6 to 16 years ($M = 10.1$ years, $SD = 2.7$ years) from a major children's hospital who had received an ADHD diagnosis and who had been administered all 10 core subtests of the WISC-IV. Of the 78 children with ADHD, 21 were classified as primarily inattentive, 3 were classified as primarily hyperactive, 33 were classified as combined, and 21 were classified as not otherwise specified (NOS). Children are diagnosed as ADHD-NOS if they have the prominent aspects of inattention or hyperactivity-impulsivity but do not meet all diagnostic criteria for ADHD listed in the *Diagnostic and Statistical Manual for Mental Disorders* (4th edition; American Psychiatric Association, 2000). The WISC-IV scores for the sample were in the average range (FSIQ, $M = 91$; VCI, $M = 93$; PRI, $M = 94$; WMI, $M = 91$; PSI, $M = 90$). The referred but nondiagnosed hospital comparison sample included 66 children (29 boys, 35 girls, and 2 unreported) aged 6 to16 years ($M = 10.3$, $SD = 2.8$) from the same children's hospital who had not received a diagnosis, and who had also been administered all 10 core subtests of the WISC-IV. The WISC-IV scores for the referred but nondiagnosed hospital comparison sample were in the average range (FSIQ, $M = 98$; VCI, $M = 100$; PRI, $M = 100$; WMI, $M = 97$; PSI, $M = 93$).

The school ADHD sample included 196 children (139 boys, 57 girls) aged 6 to 16 years ($M = 10.3$ years, $SD = 2.6$ years) from two southeastern school districts who had received an ADHD diagnosis and had been administered all 10 core subtests of the WISC-IV. The WISC-IV scores for the school ADHD sample were in the average range (FSIQ, $M = 94$; VCI, $M = 96$; PRI, $M = 98$; WMI, $M = 92$; PSI, $M = 93$). The school children without ADHD matched comparison sample included 196 children (140 boys, 56 girls) aged 6 to16 years ($M = 10.1$ years, $SD = 2.5$ years) from the same school districts who had not received an ADHD diagnosis, and who had also been administered all 10 core subtests of the WISC-IV. The comparison sample was matched to the ADHD sample based on FSIQ, age, and gender for each participant. Of the 196 children from the school comparison sample, 128 were classified as learning disabled, 27 were classified as emotionally disabled, 8 were classified as autistic, 6 were classified as speech and language impaired, 1 was classified as mentally retarded, and 26 were given no diagnosis. The WISC-IV scores for the school comparison sample were in the average range (FSIQ, $M = 94$; VCI, $M = 97$; PRI, $M = 99$; WMI, $M = 92$; PSI, $M = 92$). For the nondisabled comparison group a virtual sample

was created using EQS for Windows version 6.1 with virtually identical psychometric characteristics as reported for the standardization sample from the WISC-IV (Wechsler, 2003b). The WISC-IV normative sample was requested for this analysis but was denied by the publishing company.

Instrument

The WISC-IV is an individually administered cognitive test composed of 10 mandatory subtests ($M = 10$; $SD = 3$) that form a FSIQ score and four indexes ($M = 100$; $SD = 15$) including the VCI, PRI, WMI, and PSI. The core subtests for VCI include Similarities, vocabulary, and comprehension. The core subtests for PRI include Block Design, Picture Concepts, and Matrix Reasoning. The core subtests for WMI include digit span and letter-number sequencing, whereas the core subtests for PSI include coding and symbol search (Wechsler, 2003b).

The WISC-IV was standardized on 2,200 children ages 6 years and zero months to 16 years and 11 months who were selected to be representative of children in the United States based on the 2000 census. This sample was stratified on age, sex, race, ethnicity, parent education level, and geographic region. The average internal consistency coefficients were .97 for the FSIQ, .94 for VCI, .92 for PRI, .92 for WMI, and .88 for PSI. The median internal consistency coefficients for individual subtests ranged from .79 for symbol search and Cancellation to .90 for letter-number sequencing. A sample of 243 children were administered the WISC-IV twice at intervals ranging from 13 to 63 days, which yielded a test-retest stability coefficient of .89 for FSIQ, .89 for VCI, .85 for PRI, .85 for WMI, and .79 for PSI. An exploratory factor analysis found the factor loadings of the core subtests matched the predicted factor structure of VCI, PRI, WMI, and PSI. In addition, a confirmatory factor analysis supported this same structure (Wechsler, 2003b).

Procedure

After we received institutional review board approval, we collected childrens' WISC-IV scores and diagnoses from 322 hospital files by hospital volunteers. Participants' data were collected systematically for all active referrals from the children's hospital outpatient practice that treats neurological and behavioral conditions in children. In addition, childrens' WISC-IV scores and diagnoses were collected from 3,086 special education files of two southwestern school districts by volunteers. The participant data were collected systematically for all special education referrals, but children who were not administered the WISC-IV were excluded. The data collected included demographic information, WISC-IV scores, and diagnoses. After data collection, each child's information was reviewed and excluded if he or she was missing scores from any of the 10 core subtests.

The CPI score for each child was computed by summing the four core subtest scaled scores that comprise the working memory and processing speed indexes. After this, the child's CPI standard score was found by referencing norm tables (Weiss et al., 2008). The GAI of each child was computed by summing the six core subtest scaled scores that comprise the verbal comprehension and perceptual reasoning indexes. The GAI standard score was also found by referencing norm tables (Weiss et al., 2008). The difference between the GAI and CPI scores were then calculated for each child. These computations were repeated for all children in the simulated WISC-IV standardization sample.

Analyses

The GAI-CPI difference scores were used to compute true positive and false positive rates for each case for every possible cutoff score that then formed the ROC graphs. The resulting ROC curves are graphical representation of the accuracy of the GAI and CPI difference scores. The area under the curve (AUC) quantifies the ROC curve by producing an overall index of accuracy (Fawcett, 2006). The AUC is equal to the likelihood that test results from a randomly selected pair of affected and nonaffected participants are correctly ordered (Pepe, 2003). The AUC will always fall from 0.00 to 1.00 but random guessing equals a diagonal line that has an area of 0.50 (Fawcett, 2006). According to Swets (1988), an AUC of .50 to .70 indicates low accuracy, .70 to .90 indicates moderate accuracy, and .90 to 1.00 indicates high accuracy.

The AUC can be computed with either nonparametric (Bamber, 1975; Hanley & McNeil, 1982) or parametric (Metz, 1978) methods. The parametric approach produces a smooth ROC curve based on normal distributional assumptions. The nonparametric approach does not rely on distributional assumptions and an AUC can be obtained for a small sample size (Hajian-Tilaki, Hanley, Joseph, & Collet, 1997). Nonparametric and parametric approaches usually yield similar results but "the nonparametric method yields lower area estimates than the maximum-likelihood-estimation technique. However, these differences generally were small, particularly with ROC curves derived from five or more cutoff points" (Centor & Schwartz, 1985, p. 149). Consequently, the nonparametric approach as implemented in PASW version 18 was applied so as to remove any distributional assumptions and because this approach is more appropriate with smaller samples (Hajian-Tilaki et al., 1997).

RESULTS

Descriptive statistics for the subtest, FSIQ, GAI, CPI, and difference scores for the hospital participants are included in Table 1 and descriptive statistics for

TABLE 1 Test Score Statistics for the Hospital Attention Deficit Hyperactivity Disorder, Referred But Nondiagnosed, and Simulated Standardization Samples

	Standardization		Attention deficit hyperactivity disorder		Nondiagnosed	
	M	*SD*	*M*	*SD*	*M*	*SD*
Block design	9.98	3.01	8.73**	2.73	9.27	3.44
Similarities	9.97	3.02	9.21	3.34	10.29	3.73
Digit Span	10.00	3.03	7.95**	2.98	8.80**	2.87
Picture concepts	10.02	3.02	9.17	3.28	10.71	3.34
Coding	9.98	2.97	6.99**	3.23	7.59**	3.53
Vocabulary	9.95	3.04	8.46**	2.93	9.45	3.85
Letter-number sequencing	9.96	2.95	8.37**	3.57	9.55	3.56
Matrix reasoning	10.00	3.07	9.08	3.43	9.92	3.61
Comprehension	9.94	2.99	8.78**	2.82	9.82	3.61
Symbol search	9.85	3.01	7.94**	3.34	8.42*	3.61
Full-scale IQ	99.65	15.42	90.66**	17.01	97.95	20.81
General Ability Index	100.34	15.27	93.00**	15.67	99.79	21.31
Cognitive Proficiency Index	99.34	14.86	85.74**	16.21	90.77**	16.92
Difference between General Ability Index and Cognitive Proficiency Index score	1.20	12.16	7.26**	13.40	9.02**	12.17

*$p < .01$. **$p < .004$.

the school participants are included in Table 2. The mean subtest, GAI, CPI, and FSIQ scores for the samples were slightly lower and somewhat more variable than the normative sample. Similar patterns have been found with other clinical samples (Canivez & Watkins, 1998). We conducted a one-way analysis of variance to test whether the means differed significantly between groups. A Welch approximate F test, which does not assume homogeneity of variance, was used because of unequal group sizes. The Dunnett's C test, which does not assume equal variances, was conducted to evaluate differences among the means that proved to be statistically significant (see Tables 1 and 2). However, conducting multiple tests increases the chance that at least one of them will be statistically significant by chance alone (Type 1 error). Thus, the alpha level for each individual test was set at .004 (.05 ÷ 14) to maintain the experimentwise error rate at .05.

For the hospital sample many of the subtests as well as the GAI, CPI, FSIQ were statistically significant at the .004 level. The tests that were statistically significant included block design, $F(2, 99.91) = 8.91$, $p < .001$; digit span, $F(2, 100.41) = 22.53$, $p < .001$; coding, $F(2, 98.66) = 42.76$, $p < .001$; vocabulary, $F(2, 99.07) = 10.06$, $p < .001$; letter-number sequencing, $F(2, 98.12) = 7.87$, $p < .001$; comprehension, $F(2, 99.33) = 6.29$, $p < .004$; symbol search, $F(2, 98.51) = 17.01$, $p < .001$; GAI, $F(2, 98.39) = 8.24$, $p < .001$; CPI, $F(2, 98.84) = 33.93$, $p < .001$; and the FSIQ, $F(2, 93.66) = 9.75$, $p < .001$. The Dunnett's C post hoc test indicated that digit span, coding, letter-number sequencing, symbol search, GAI, and FSIQ scores were significantly different between the ADHD and normative samples. In addition, the digit span, coding, and CPI scores were significantly different between the normative sample and both the ADHD and nondiagnosed samples.

GAI-CPI difference scores for the ADHD, nondiagnosed, and simulated samples were different at a statistically significant level $F(2, 99.47) = 20.22$, $p < .001$. The Dunnett's C test indicated that the ADHD and nondiagnosed hospital samples were both significantly different from the simulated normative sample but not significantly different from each other. The ADHD and nondiagnosed groups each had larger GAI-CPI difference scores than the simulated normative group.

For the school sample the tests that were statistically significant at the .004 level included digit span, $F(2, 292.55) = 43.30$, $p < .001$; coding, $F(2, 280.17) = 37.73$, $p < .001$; vocabulary, $F(2, 298.25) = 9.00$, $p < .001$; letter-number sequencing, $F(2, 292.43) = 26.22$, $p < .001$; comprehension, $F(2, 291.26) = 5.93$, $p < .004$; symbol search, $F(2, 283.70) = 12.28$, $p < .001$; GAI, $F(2, 295.96) = 5.86$, $p < .004$; CPI, $F(2, 295.20) = 59.88$, $p < .001$; and the FSIQ, $F(2, 295.91) = 25.49$, $p < .001$. The Dunnett's C post hoc test indicated that block design, vocabulary, letter-number sequencing, comprehension, symbol search, CPI, and FSIQ scores were significantly different between the normative sample and both the ADHD and school comparison samples. In addition, the vocabulary scores were statistically significant between the

TABLE 2 Test Score Statistics for the School Attention Deficit Hyperactivity Disorder, School Comparison, and Simulated Standardization Samples

	Standardization		Attention deficit hyperactivity disorder		School	
	M	*SD*	*M*	*SD*	*M*	*SD*
Block design	9.98	3.01	9.35^{+}	3.07	9.61	2.99
Similarities	9.97	3.02	9.46	2.92	9.65	2.94
Digit Span	10.00	3.03	8.68**	2.71	8.41**	2.79
Picture concepts	10.02	3.02	10.11	2.87	9.83	3.16
Coding	9.98	2.97	8.30**	3.01	8.42**	3.33
Vocabulary	9.95	3.04	9.39^{+}	2.46	9.24**	2.75
Letter-number sequencing	9.96	2.95	8.78**	2.62	8.95**	2.74
Matrix reasoning	10.00	3.07	*9.68*	*2.96*	9.72	2.74
Comprehension	9.94	2.99	9.33^{+}	2.71	9.49	2.79
Symbol search	9.85	3.01	9.10**	3.02	8.92**	3.07
Full-scale IQ	99.65	15.42	94.11**	13.59	94.14**	13.37
General Ability Index	100.34	15.27	97.61^{+}	12.86	97.85	13.87
Cognitive Proficiency Index	99.34	14.86	91.29**	13.31	90.96**	12.87
Difference between General Ability Index and Cognitive Proficiency Index score	1.20	12.16	6.32**	10.82	6.89**	13.65

$^{+}p < .05$. $^{*}p < .01$. $^{**}p < .004$.

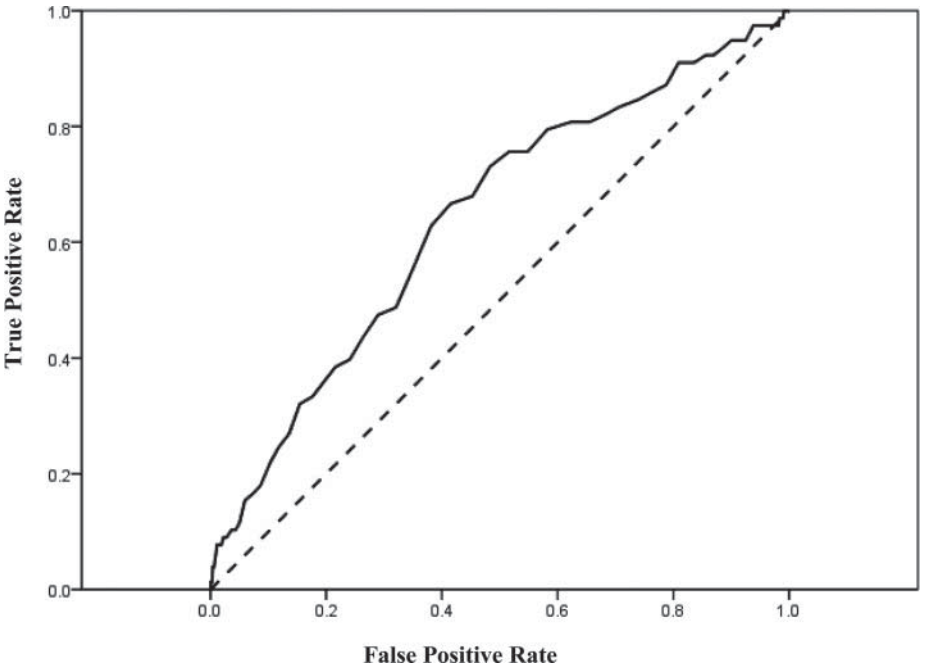

FIGURE 3 Receiver operating characteristic curve of children with attention deficit hyperactivity disorder compared with the simulated Wechsler Intelligence Scale for Children-Fourth Edition standardization sample.

normative and school comparison sample. The GAI-CPI difference scores for the ADHD, matched comparison, and simulated samples were different at a statistically significant level $F(2, 285.56) = 31.25$, $p < .001$. Similar to the hospital sample the ADHD and matched comparison groups had larger GAI-CPI difference scores than the simulated normative group.

The result of the ROC analysis comparing hospital children with ADHD to the simulated WISC-IV standardization sample is presented in Figure 3. The AUC of .64, 95% CI [0.58, 0.71] quantifies these visual results. If a child were randomly selected from the ADHD sample and another child randomly chosen from the standardization sample, the child with ADHD would have a higher GAI-CPI difference score about 64% of the time (Ruttimann, 1994).

The ROC analysis comparing hospital children with ADHD to the nondiagnosed hospital comparison sample is presented in Figure 4. The resulting AUC was .46, 95% CI [0.37, 0.56]. If one child from each sample was randomly selected, the child with ADHD could not be differentiated from the child who was referred but not diagnosed based on having a higher GAI-CPI difference score (Ruttimann, 1994).

The ROC analysis comparing school children with ADHD to the simulated WISC-IV standardization sample is presented in Figure 5. The resulting AUC of .63, 95% CI [0.59, 0.67] indicates that if a child was randomly selected

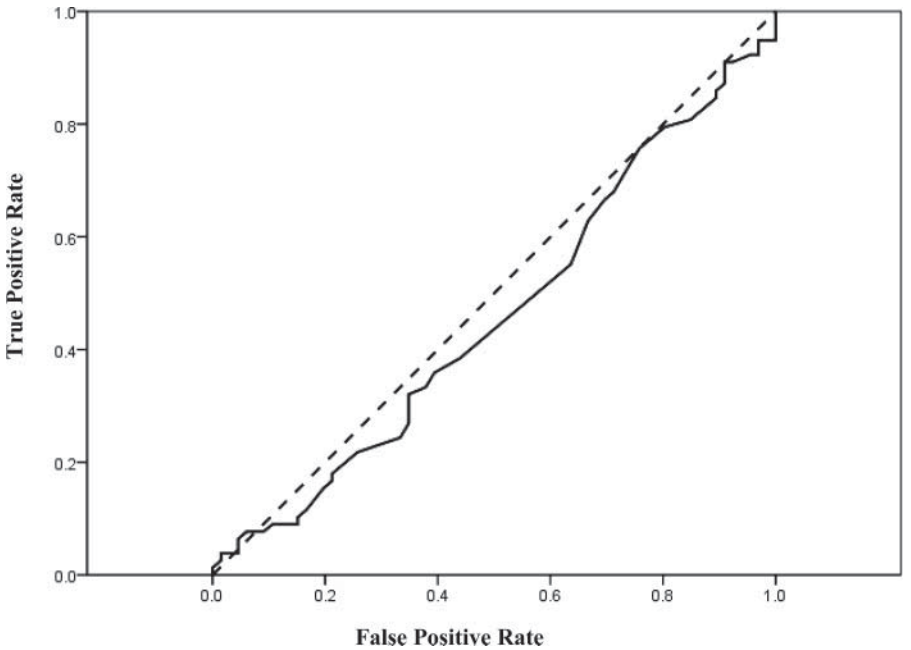

FIGURE 4 Receiver operating characteristic curve of children with attention deficit hyperactivity disorder compared with the referred but nondiagnosed hospital comparison sample.

from the ADHD school sample and another child randomly chosen from the standardization sample, the child with ADHD would have a higher GAI-CPI difference score about 63% of the time (Ruttimann, 1994).

Last, the ROC analysis comparing school children with ADHD to the matched school comparison sample is presented in Figure 6. The resulting AUC of .50, 95% CI [0.45, 0.56] indicates that the GAI-CPI discrepancy method operated at chance levels for these two groups of children. The AUC score for each of the ROC analyses indicate that the GAI-CPI discrepancy method would be classified as low accuracy (Swets, 1988).

DISCUSSION

Some researchers have hypothesized that WISC-IV GAI-CPI difference scores can be used to accurately diagnose children with ADHD. The results of this study indicated that children with ADHD and those without ADHD had significantly different group mean scores on several subtest, CPI, and GAI-CPI discrepancy scores than children in the simulated standardization sample. In contrast, children with ADHD did not perform differently, on average, than non-ADHD children. These group differences mirror past research on children with ADHD versus nonclinical children that found children with

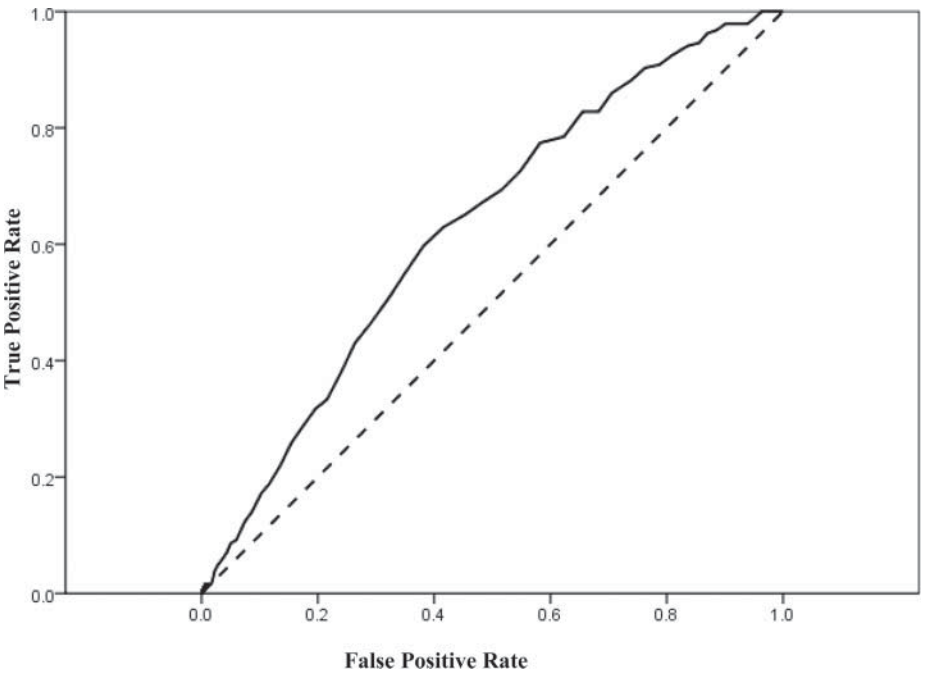

FIGURE 5 Receiver operating characteristic curve of school children with attention deficit hyperactivity disorder compared with the simulated Wechsler Intelligence Scale for Children-Fourth Edition standardization sample.

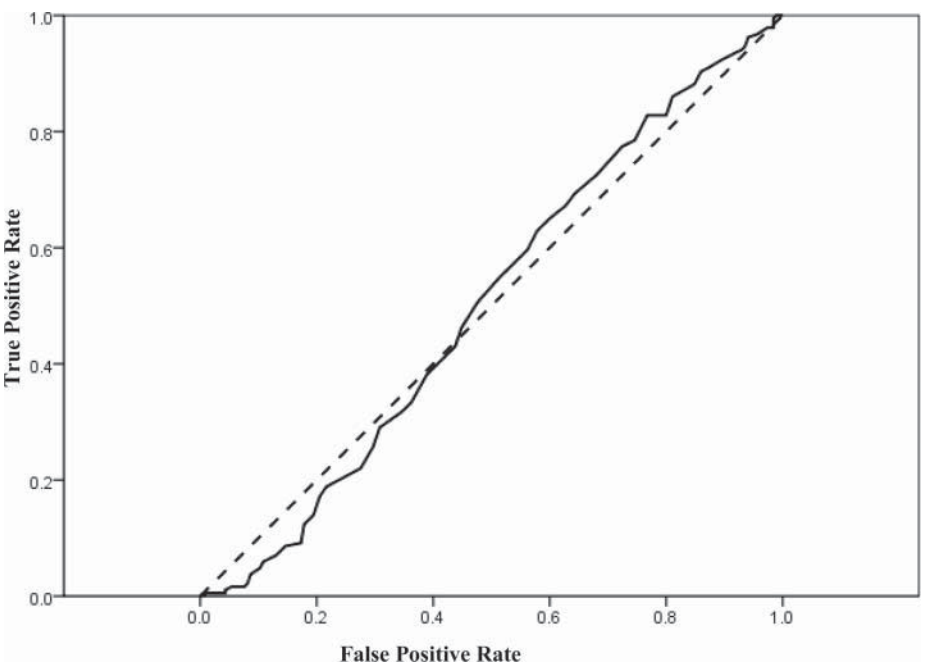

FIGURE 6 Receiver operating characteristic curve of school children with attention deficit hyperactivity disorder compared with the matched special education comparison sample.

ADHD to exhibit VCI and PRI scores higher than their PSI and WMI scores (Mayes & Calhoun, 2006; Wechsler, 2003b).

However, group mean differences on GAI-CPI discrepancy scores do not necessarily indicate clinical utility for individual children (Watkins, 2009). ROC analyses demonstrated that the GAI-CPI discrepancy method can accurately distinguish a randomly chosen child with ADHD from a randomly chosen nonclinical child 64% of the time for the hospital sample and 63% of the time for the school sample compared with 84% of the time when child behavior checklists are employed (Chen, Faraone, Biederman, & Tsuang, 1994). ROC analyses also revealed that GAI-CPI difference scores cannot distinguish between children with ADHD versus those with other clinical disorders from the same hospital and schools at greater than chance levels. Thus, using the GAI-CPI cognitive profile to distinguish children with ADHD is less accurate than the methods already used by many clinicians and considered best practice for identifying children with ADHD (American Academy of Child and Adolescent Psychiatry, 2007).

Limitations

As with all research, this study was marked by several limitations. The first limitation was the diagnoses given to participants. The examining psychologists used a variety of methods to diagnose ADHD. Although each child in these samples was given a psychological evaluation, his or her diagnosis was based on a variety of tests, interviews, behavioral checklists, and clinical judgments not necessarily consistent with the *Diagnostic and Statistical Manual for Mental Disorders* (4th edition; American Psychiatric Association, 2000) criteria. In addition, many of the children with ADHD had co-morbid diagnoses. Comorbidity, however, is a common occurrence for children with ADHD (Acosta, Arcos-Burgos, & Muenke, 2004; Faraone & Biederman, 1998). Furthermore, children included in this study had a mixture of ADHD subtypes including primarily inattentive, primarily hyperactive, combined, and NOS. Differences have been found in the cognitive processes of children with primarily hyperactive and combined types of ADHD compared with children with the primarily inattentive type of ADHD (Schwean & McCrimmon, 2008). In addition, children that are diagnosed with ADHD-NOS do not meet the necessary criteria for ADHD (American Psychiatric Association, 2000). However, when children with ADHD-NOS were removed from the hospital sample the result of the ROC analyses were almost identical.

A second limitation is that medication use of participants was not known. The effect of medication on children with ADHD has not shown to change cognitive impairments but has been shown to normalize deficits in executive functioning including working memory (Schwean & McCrimmon, 2008). As

a result, children with ADHD who were on medication may have achieved higher CPI scores than children with ADHD not on medication.

A final limitation is the generalizability of these results to other children. The samples were not collected from random hospitals and school districts. This resulted in the samples being demographically and regionally limited. In addition, simulated data were used instead of actual participants from the WISC-IV standardization sample. As a result, caution should be used when applying these study results to other groups of children.

Future Research

Future research should continue to address GAI-CPI difference scores as possible indicators of ADHD. Method of diagnosis, co-morbidity, medication usage, and ADHD subtypes should be controlled in order to allow unambiguous diagnostic utility results to emerge. Additional research should also be conducted on GAI-CPI discrepancy scores for other specialized groups of children. Specifically, groups of children with learning disabilities, traumatic brain injury, and Asperger's syndrome who have been hypothesized to have noteworthy GAI-CPI difference scores (Weiss & Gable, 2007). This research should assess GAI-CPI difference scores without depending on cutoff scores or base rates (Swets, 1988).

Implications

Although the study results should be considered preliminary because of its limitations, clinicians should be cautious about interpreting WISC-IV GAI-CPI difference scores as evidence of ADHD. GAI-CPI difference scores, although statistically significant between groups, have little to no individual diagnostic accuracy (Swets, 1988). As with past research, GAI-CPI difference scores alone should not be considered diagnostic markers of ADHD (Weiss & Gable, 2007). Unless additional research indicates that there is higher diagnostic accuracy of GAI-CPI difference scores to differentiate children with ADHD from those without ADHD this method should not be used by clinicians.

REFERENCES

Acosta, M. T., Arcos-Burgos, M., & Muenke, M. (2004). Attention deficit/hyperactivity disorder (ADHD): Complex phenotype, simple genotype? *Genetics in Medicine*, *6*, 1–15.

Adams, P. F., Lucas, J. W., & Barnes, P. M. (2008). Summary health statistics for U.S. children: National Health Interview Survey 2006. *Vital Health Statistics*, *10*, 1–104.

American Academy of Child and Adolescent Psychiatry. (2007). Practice parameter for the assessment and treatment of children and adolescents with

attention-deficit/hyperactivity disorder. *Journal of the American Academy of Child and Adolescent Psychiatry*, *46*, 894–921.

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

Anastopoulos, 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.

Bamber, D. (1975). The area above the ordinal dominance graph and the area below the receiver operating characteristic graph. *Journal of Mathematical Psychology*, *12*, 387–415.

Barkley, R. A. (1991). The ecological validity of laboratory and analogue assessment methods of ADHD symptoms. *Journal of Abnormal Child Psychology*, *19*, 149–178.

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler test: Why does it persist? *School Psychology International*, *19*, 209–220.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children–Third Edition. *Psychological Assessment*, *10*, 285–291.

Centers for Disease Control and Prevention. (2005). Mental health in the United States: Prevalence of diagnosis and medication treatment for attention-deficit hyperactivity disorder—United States, 2003. *Morbidity and Mortality Weekly Report*, *54*, 842–847.

Centor, R. M., & Schwartz, J. S. (1985). An evaluation of methods for estimating the area under the receiver operating characteristic (ROC) curve. *Medical Decision Making*, *5*, 149–156.

Chen, W. J., Faraone, S. V., Biederman, J., & Tsuang, M. T. (1994). Diagnostic accuracy of the Child Behavior Checklist Scales for attention-deficit hyperactivity disorder: A receiver-operating characteristic analysis. *Journal of Consulting and Clinical Psychology*, *62*, 1017–1025.

DuPaul, G. J. (1992). How to assess attention-deficit hyperactivity disorder within school settings. *School Psychology Quarterly*, *7*, 60–74.

Elwood, R. W. (1993). Psychological tests and clinical discriminations: Beginning to address the base rate problem. *Clinical Psychology Review*, *13*, 409–419.

Faraone, S. V., & Biederman, J. (1998). Neurobiology of attention-deficit hyperactivity disorder. *Biological Psychiatry*, *44*, 951–958.

Fawcett, T. (2006). An introduction to ROC analysis. *Pattern Recognition Letters*, *2*, 861–874.

Feldt, L. S., & Brennan, R. L. (1993). *Reliability.* In R. L. Linn (Ed.), *Educational measurement (3rd ed.*, pp. 105–146). Phoenix, AZ: Oryx Press.

Frazier, T. W., Youngstrom, E. A., Glutting, J. J., & Watkins, M. W. (2007). ADHD and achievement: Meta-analysis of the child, adolescent, and adult literatures and a concomitant study with college students. *Journal of Learning Disabilities*, *40*, 49–65.

Gussin, B., & Javorsky, J. (1995). The utility of the WISC–III freedom from distractibility in the diagnosis of youth with attention deficit hyperactivity disorder in a psychiatric sample. *Diagnostique*, *21*, 29–40.

Hajian-Tilaki, K. O., Hanley, J. A., Joseph, L., & Collet, J. P. (1997). A comparison of parametric and nonparametric approaches to ROC analysis of quantitative diagnostic tests. *Decision Making*, *17*, 94–102.

Hanley, J. A., & McNeil, B. J. (1982). The meaning and use of the area under an ROC curve. *Radiology*, *143*, 29–36.

Joschko, M., & Rourke, B. P. (1985). Neuropsychological subtypes of learning-disabled children who exhibit the ACID pattern on the WISC. In B. P. Rourke (Ed.), *Neuropsychology of learning disabilities* (pp. 65–88). New York, NY: Guilford Press.

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

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

Mayes, S. D., & Calhoun, S. L. (2006). WISC-IV and WISC-III profiles in children with ADHD. *Journal of Attention Disorders*, *9*, 486–493.

Mayes, S. D., Calhoun, S. L., & Crowell, E. W. (1998). 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.

Metz, C. E. (1978). Basic principles of ROC analysis. *Seminars in Nuclear Medicine*, *8*, 283–298.

Pepe, M. S. (2003). *Statistical evaluation of medical tests for classification and prediction*. New York, NY: Oxford University Press.

Pintea, S., & Moldovan, R. (2009). The receiver-operating characteristics (ROC) analysis: Fundamentals and applications in clinical psychology. *Journal of Cognitive & Behavioral Psychotherapies*, *9*, 49–66.

Power, T. J., & Ikeda, M. J. (1996). The clinical utility of behavior rating scales: Comments on the diagnostic assessment of ADHD. *Journal of School Psychology*, *34*, 379–385.

Prifitera, A., & Dersh, J. (1993). Base rates of WISC-III diagnostic subtest patterns among normal, learning-disabled, and ADHD samples. *Journal of Psychoeducational Assessment, WISC-III Monograph*, 43–55.

Raiford, S. E., Weiss, L. G., Rolfhus, E. L., & Coalson, D. (2005). *Wechsler Intelligence Scale for Children-Fourth Edition, General Ability Index* (Technical Report No. 4). San Antonio, TX: Harcourt Assessment.

Ruttimann, U. E. (1994). Statistical approaches to development and validation of predictive instruments. *Critical Care Clinics*, *10*, 19–35.

Ryan, J. J., Glass, L. A., & Bartels, J. M. (2010). Stability of the WISC-IV in a sample of elementary and middle school children. *Applied Neuropsychology*, *17*, 68–72.

Sattler, J. M. (2008). *Assessment of children: Cognitive foundations* (5th ed.). San Diego, CA: Jerome M. Sattler.

Schwean, V. L., & McCrimmon, A. (2008). Attention-deficit/hyperactivity disorder: Using the WISC-IV to inform intervention planning. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (pp. 193–215). San Diego, CA: Academic Press.

Skounti, M., Philalithis, A., & Galanakis, E. (2007). Variations in prevalence of attention deficit hyperactivity disorder worldwide. *European Journal of Pediatrics*, *166*, 117–123.

Snow, J. B., & Sapp, G. L. (2000). WISC-III subtest patterns of ADHD and normal samples. *Psychological Reports*, *87*, 759–765.

Swartz, C. L., Gfeller, J. D., Hughes, H. M., & Searight, H. R. (1998). The prevalence of WISC–III profiles in children with attention deficit hyperactivity disorder and learning disabilities. *Archives of Clinical Neuropsychology*, *13*, 85.

Swets, J. A. (1988). Measuring the accuracy of diagnostic systems. *Science*, *240*, 1285–1293.

Watkins, M. W. (2003). IQ subtest analysis: Clinical acumen or clinical illusion. *Scientific Review of Mental Health Practice*, *2*, 118–141.

Watkins, M. W. (2009). Errors in diagnostic decision making and clinical judgment. In T. B. Gutkin & C. R. Reynolds (Eds.), *Handbook of school psychology* (4th ed., pp. 210–229). New York, NY: Wiley.

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (2005). Issues in subtest profile analysis. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 251–268). New York, NY: Guilford Press.

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

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997b). Discriminant and predictive validity of the WISC–III ACID profile among children with learning disabilities. *Psychology in the Schools*, *34*, 309–319.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (1991). *Wechsler Intelligence Scale for Children-Third Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children-Fourth Edition*. San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003b). *WISC-IV technical and interpretive manual*. San Antonio, TX: Psychological Corporation.

Weiss, L. G., Beal, A. L., Saklofske, D. H., Alloway, T. P., & Prifitera, A. (2008). Interpretation and intervention with the WISC-IV in the clinical assessment context. In A. Prifitera, D. H. Saklofske, & L. G. Weiss (Eds.), *WISC-IV clinical assessment and intervention* (pp. 3–66). San Diego, CA: Academic Press.

Weiss, L. G., & Gabel, A. D. (2007). *Using the cognitive proficiency index in psychoeducational assessment* (Technical Report No. 6). San Antonio, TX: Harcourt Assessment.

Wielkiewicz, R. M. (1990). Interpreting low scores on the WISC-R third factor: It's more than distractibility. *Psychological Assessment: A Journal of Consulting and Clinical Psychology*, *2*, 91–97.

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