Diagnostic Utility of WISC-IV

General Abilities Index and Cognitive Proficiency Index

Difference Scores among Children with ADHD

by

Sarah Elaine Devena

A Thesis Presented in Partial Fulfillment
of the Requirements for the Degree
Master of Arts

Approved November 2010 by the
Graduate Supervisory Committee:

Marley Watkins, Chair
Amanda Sullivan
David Wodrich

ARIZONA STATE UNIVERSITY

December 2010

ABSTRACT

The Wechsler Intelligence Scale for Children-Fourth Edition (WISC-IV)

General Abilities Index (GAI) and Cognitive Proficiency Index (CPI) have been

advanced as possible diagnostic markers of Attention-Deficit Hyperactivity

Disorder (ADHD). Diagnostic utility statistics were used to test the ability of

GAI-CPI difference scores to identify children with ADHD. Participants included

an ADHD sample ($n = 78$), a referred but non-diagnosed hospital sample ($n = 66$),

and a simulated sample with virtually identical psychometric characteristics as the

WISC-IV 2,200 child standardization sample. Receiver Operating Characteristic

(ROC) analyses were computed to determine the utility of GAI-CPI difference

scores to identify children with ADHD. The GAI-CPI discrepancy method had an

AUC of .64, 95% CI [0.58, 0.71] for the ADHD sample compared to the

simulated normative sample and an AUC of .46, 95% CI [0.37, 0.56] for the

ADHD sample compared to the referred but non-diagnosed hospital sample.

These AUC scores indicate that the GAI-CPI discrepancy method has low

accuracy.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Chapter 1

Introduction

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 (2005), ADHD is currently one of the most

common neurobehavioral disorders of children in the United States. Over the last

decade, ADHD diagnoses have increased an average of 3% per year and currently

3% to 7% of school-aged children have been diagnosed with the disorder (Adams,

Lucas, & Barnes, 2008). ADHD can have a profound effect upon 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 non-disabled students

(Skounti, Philalithis, & Galanakis, 2007).

Various methods are used to diagnose ADHD and can include: (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, 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 WISC-IV, beyond its popularity, is a valuable instrument for the diagnostic assessment of children (Weiss, Beal, Saklofske, Alloway, & Prifitera, 2008).

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, Kauffman (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). With the introduction of the Wechsler Intelligence Scale for Children–Third Edition (WISC-III; Wechsler, 1991) the freedom from distractibility (FD) profile was modified to consist of just the Arithmetic and Digit Span subtests to match the

factor structure of the WISC-III. When children scored high on this FD profile it was thought to indicate the ability to sustain attention and when they scored low on this FD profile it was thought to indicate distractibility (Kauffman). Because of this hypothesis, low scores on the FD profile 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 profile 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 profile was significantly lower than the childrens' 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 profile 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 Span subtests. Subsequently, the term SCAD was coined for this profile (Kauffman, 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)

3

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 to 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 profile to enhance diagnostic accuracy. Research on clinical vs. non-clinical groups indicated that the ACID profile occurred in 12% of children with ADHD compared to only 1% of children from the non-ADHD group (Prifitera & Dersh, 1993). These findings led Prifitera and Dersh 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. found no significant difference between the ADHD and LD samples in the frequency of ACID profiles.

Although the FD, SCAD, and ACID profiles appeared to be valid markers of ADHD in these studies, there are two substantial limitations to this research. The first limitation is the focus on subtest scores. Subtest scores have relatively weak reliability, especially when compared to 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

4

subtests is .86, compared to .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 vs. .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 to 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 discriminate validity but it does not necessarily have clinical utility. As a result, discriminate 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

5

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 mean differences.
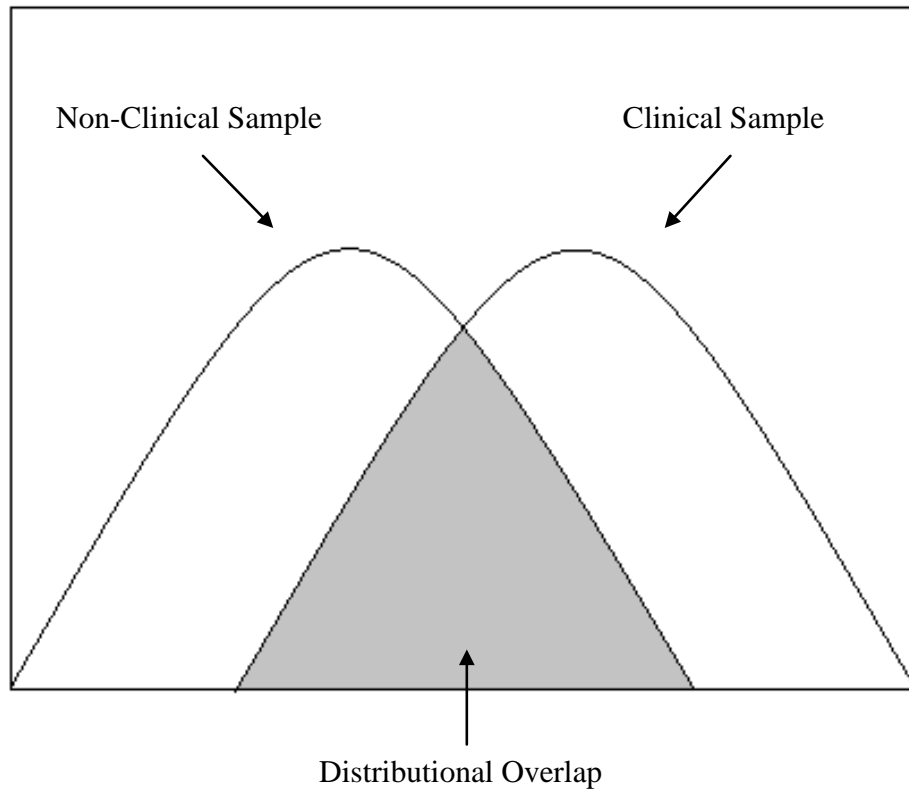
Figure 1. Hypothetical mean differences between ADHD and non-ADHD groups
showing the distributional overlap of the groups in the shaded region.

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 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). Additionally, 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).

7

The current practice of using composite scores makes research into the diagnostic utility of these score patterns of vital interest to clinicians. Following this trend, 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, fourth edition, text revision (DSM-IV-TR; American Psychiatric Association, 2000) diagnostic criteria. The children were selected, based upon their availability, from a variety of educational and clinical settings. On average, children with ADHD performed worse on PSI and WMI indexes compared to 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 vs. 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). Both 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

9

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 non-clinical groups

were selected during the WISC-IV standardization project (Weiss & Gable,

2007). By comparing childrens' CPI to their GAI, Weiss and Gable wanted to

identify a cut score that would distinguish between clinical and non-clinical

groups with a true positive rate (TPR) and a true negative rate (TNR) of at least

60%. There are four possible outcomes when applying cut scores to categorize

individual cases: (a) the child has ADHD and is classified as such, which is a true

positive, (b) the child has ADHD and is classified as not having ADHD, which is

a false negative, (c) the child does not have ADHD and is classified as such,

which is a true negative, and (d) the child does not have ADHD and is classified

as having ADHD, which is a false positive. From these statistics a TPR (correctly

classified positives divided by the total positives) and false positive rate (FPR;

incorrectly classified negatives divided by the total negatives) can be calculated

(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.

Finally, the Aspergers 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 from these results 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 cut score. TPR and TNR values would have differed if other cut scores had been selected. Additionally, the TPR and TNR are dependent upon base rates (Elwood, 1993), which means that the TPR and TNR will vary dependent upon 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 cut score (Swets, 1988). To avoid this issue, a Receiver Operating Characteristic (ROC) analysis could be conducted. ROC is a procedure used to measure the accuracy of tests that are used to discriminate between groups (Pintea & Moldovan, 2009). A ROC curve is drawn by plotting individual points for all possible cut scores. In other words, plotting the balance between the TPR and the

FPR for the test while moving the cut 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).
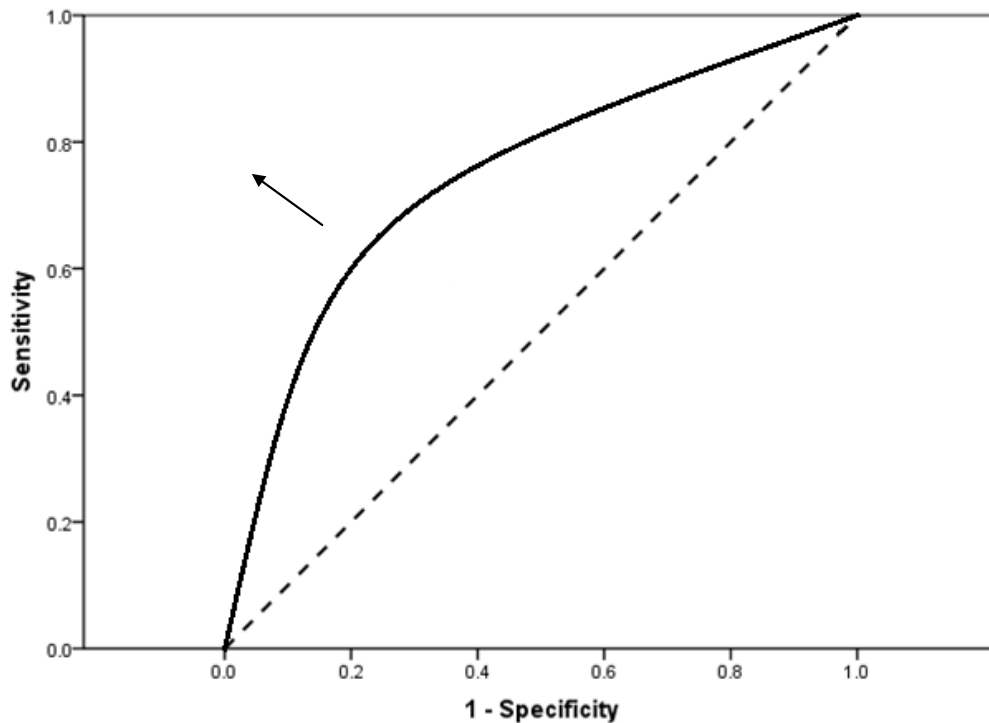
Figure 2. Hypothetical Receiver Operating Characteristic (ROC) curve with diagonal chance line showing that as the ROC curve moves farther towards the left corner of the graph the more accurate a test is.

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.

Additionally, in the research conducted by Weiss and Gable (2007) only the TPR and TNR for one cut score were calculated when considering the diagnostic accuracy of CPI < GAI discrepancies. For these reasons, this 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. The first analysis will test the difference between children with ADHD and a non-clinical WISC-IV simulated standardization sample. The second analysis will test the difference between children with ADHD and a referred but non-diagnosed comparison sample.

Chapter 2

Method

Participants

The ADHD sample included 78 children (56 males, 22 females) aged 6 to 16 years ($M = 10.1$, $SD = 2.7$) 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). Of the participants, secondary diagnoses included 11 children with Aspergers disorder, 4 with an anxiety disorder, 4 with obsessive compulsive disorder, 8 with oppositional defiant disorder, 9 with a depressive disorder, 4 with autism, 3 with a speech or language impairment, 4 with a mood disorder, 4 with bipolar disorder, 8 with a learning disability, 9 with mixed neuropsychological deficits, and 12 with various additional medical conditions. 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 non-diagnosed hospital comparison sample included 66 children (29 males, 35 females, and 2 unreported) aged 6 to 16 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 non-diagnosed hospital comparison sample were in the average range (FSIQ $M = 98$, VCI $M = 100$, PRI $M = 100$, WMI $M = 97$, PSI $M = 93$).

For the non-disabled 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 Coding, whereas the core subtests for PSI include Letter-Number Sequencing 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. Additionally, a confirmatory factor analysis supported this same structure (Wechsler, 2003b).

Procedure

Following IRB approval, childrens' WISC-IV scores and diagnoses were collected from 322 hospital files by two hospital volunteers. The participant data were collected systematically for all active referrals from the children's hospital outpatient practice that treats both neurological and behavioral conditions in children. Information was not collected from the files if the participant was not administered the WISC-IV. The data collected included demographic information, WISC-IV scores, achievement scores, and the child's primary, secondary, and tertiary diagnosis. After data collection, each child's information was reviewed and excluded if he or she was missing scores from 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. Following 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.). The difference between the

16

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

Initially, the means and standard deviations were computed as well as the statistical significance for all subtest and composite scores. Subsequently, the GAI-CPI difference scores were used to compute true positive and false positive rates for each case for every possible cut score that then formed the two ROC graphs. The smallest cutoff value was the minimum difference score minus one and the largest cutoff value was the maximum difference score plus one. 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 non-affected 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, so the classifier should never be less than that (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. For ADHD diagnostic utility of the WISC-IV, an AUC of at least .84 is desired because this is the lower end AUC score found when using the Child Behavior Checklist (Chen, Faraone, Biederman, & Tsuang, 1994) to diagnose ADHD. If the AUC does not

17

reach .84, clinicians should be encouraged to use behavior checklists instead of GAI-CPI discrepancy scores.

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.).

To conduct a power analysis, the Wilcoxon-Mann-Whitney statistic was referenced. The Wilcoxon-Mann-Whitney statistic is used to test if two data samples come from the same distribution (Wu & Flach, 2005). There is an underlying equivalence of the AUC to the Wilcoxon test and the Mann-Whitney U to the Wilcoxon rank-sum test (Centor, 1985). This relationship makes the Wilcoxon-Mann-Whitney statistic equivalent to the AUC (Wu & Flach).

Based upon the association between Wilcoxon-Mann-Whitney and AUC, this alternative statistic was analyzed to determine the appropriate sample size.

According to Cohen (1988), a rule of thumb for the behavioral sciences is that small, medium, and large effect sizes could be represented by standardized mean differences of 0.2, 0.5, and 0.8, respectively. For comparison of children with ADHD to the simulated non-disabled sample, results of the Wilcoxon-Mann-Whitney power analysis with an allocation ratio of 29:1 showed that for a relatively small effect size of $d = .20$, 169 and 4,858 participants would be needed for the first and second sample, respectively. For a moderate effect size of $d = .50$, 27 and 779 participants would be needed for the first and second sample respectively. For a much larger effect size of $d = .80$, 11 and 305 participants would be needed for the first and second sample, respectively. Because most of the effect sizes found in the subtest analysis literature range from small to moderate, an effect size of $d = .30$ will be estimated for this study. When $\alpha = .05$, power $(1 - \beta) = .80$, and $d = .30$, $n = 74$ children with ADHD and $n = 2,160$ children from the simulated WISC-IV standardization sample were needed. For comparison of children with ADHD to the referred but non-diagnosed hospital comparison sample results of the Wilcoxon-Mann-Whitney power analysis with an allocation ratio of 1 showed that when $\alpha = .05$, power $(1 - \beta) = .80$, and $d = .45$, $n = 65$ children with ADHD and $n = 65$ children with no diagnosis were needed.

Cohen's d is intended for scores of two populations being compared that are continuous and normally distributed. The AUC, on the other hand, is equal to the probability that a score drawn from one sample is higher than that drawn from a second sample (Rice & Harris, 2005). According to Rice and Harris's comparison table, when $d = .31$ then the researcher has the power to detect an

AUC of .587 or higher. Additionally when $d = .453$ then the researcher has the

power to detect an AUC of .626 or higher.

Chapter 3

Results

Descriptive statistics for the subtest, FSIQ, GAI, CPI, and difference

scores for each group of participants are included in Table 1. The mean subtest,

GAI, CPI, and FSIQ scores for the ADHD and the non-diagnosed hospital

samples were slightly lower and somewhat more variable than the normative

sample. Similar patterns have been found with other clinical samples (Canivez &

Watkins, 1998). A one-way analysis of variance was conducted to test if 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 Table 1). 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.

Table 1

Mean, Standard Deviation, and Statistical Significance of WISC-IV Scores for the

ADHD, Referred but Non-Diagnosed, and Simulated Standardization Samples

	Standardization		ADHD		Clinical	
	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	9.96	2.95	8.37**	3.57	9.55	3.56
Matrix R.	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
FSIQ	99.65	15.42	90.66**	17.01	97.95	20.81
GAI	100.34	15.27	93.00**	15.67	99.79	21.31
CPI	99.34	14.86	85.74**	16.21	90.77**	16.92
GAI-CPI	1.20	12.16	7.26**	13.40	9.02**	12.17

Note. WISC-IV = Wechsler Intelligence Scale for Children-Fourth Edition,

ADHD = Attention-Deficit Hyperactivity Disorder, Letter-Number = Letter-

Number Sequencing, Matrix R. = Matrix Reasoning, FSIQ = Full Scale IQ, GAI =

General Ability Index, CPI = Cognitive Proficiency Index, GAI-CPI = GAI-CPI

difference scores.

[+] $p < .05$. * $p < .01$. ** $p < .004$.

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 Block Design, Vocabulary, Letter-Number Sequencing, Comprehension, Symbol Search, GAI, and FSIQ scores were significantly different between the ADHD and normative samples. Additionally, the Digit Span, Coding, and CPI scores were significantly different between the normative sample and both the ADHD and non-diagnosed samples.

GAI-CPI difference scores for the ADHD, non-diagnosed, 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 non-diagnosed hospital samples were both significantly different from the simulated normative sample but not significantly different from each other. The ADHD and non-diagnosed groups each had larger GAI-CPI difference scores than the simulated normative group.

The result of the ROC analysis comparing 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. The AUC score indicates

that the GAI-CPI discrepancy method, although better than chance, would be

classified as low accuracy (Swets, 1988). That is, if a child was 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

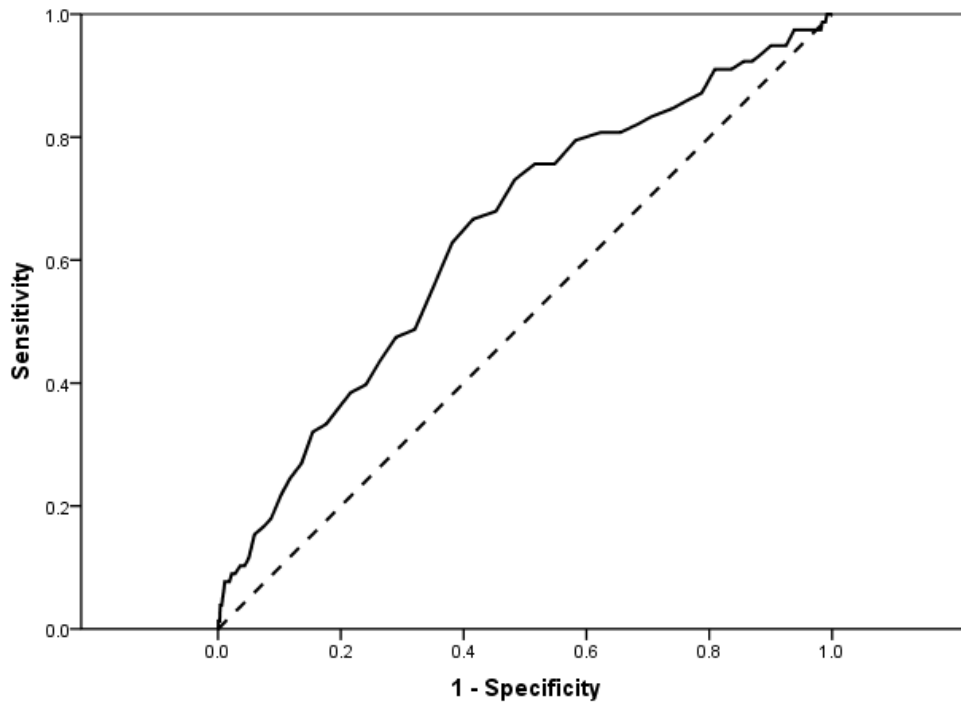difference score about 64% of the time (Ruttimann, 1994).

Figure 3. Receiver Operating Characteristic curve of children with ADHD
compared to the simulated WISC-IV standardization sample.

The ROC analysis comparing children with ADHD to the non-diagnosed

hospital comparison sample is presented in Figure 4. The resulting AUC was .46,

95% CI [0.37, 0.56]. This AUC shows that the GAI-CPI discrepancy method is

below chance levels for these two groups of children, which indicates low test

accuracy (Swets, 1988). In other words 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
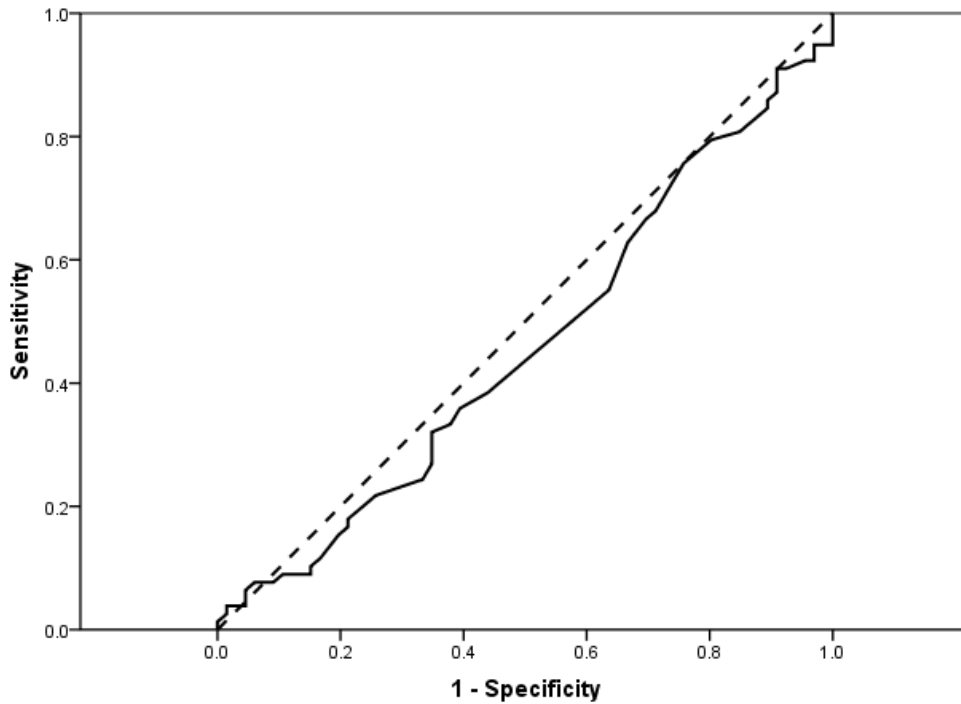
difference score (Ruttimann, 1994)

Figure 4. Receiver Operating Characteristic curve of children with ADHD
compared to the referred but non-diagnosed hospital comparison sample.

Chapter 4

Discussion

The research question was if WISC-IV GAI-CPI difference scores can be used to accurately diagnose children with ADHD. The results indicated that group mean difference scores were found between ADHD, non-diagnosed hospital, and normative groups. Children with ADHD and those in the non-diagnosed hospital sample 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-diagnosed but referred children in this study. These group differences mirror past research on children with ADHD versus non-clinical children that found children with ADHD to exhibit VCI and PRI scores higher than their PSI and WMI scores (Mayes & Calhoun, 2006; Wechsler, 2003b).

Group mean differences on GAI-CPI discrepancy scores do not necessarily indicate clinical utility for individual children. Statistical significance alone is not sufficient for diagnosing individuals due to the distributional overlap of scores (Watkins, 2009). To determine the clinical utility of GAI-CPI difference scores, a ROC analysis was used to gain a more accurate representation of the diagnostic performance of the test (Pepe, 2003). The ROC and AUC analyses showed that there was low diagnostic utility when comparing children with ADHD to the simulated standardization sample. The GAI-CPI discrepancy method can accurately distinguish a randomly chosen child with ADHD from a

26

non-clinical child 64% of the time compared to 84% of the time when child behavior checklists are employed (Chen et al., 1994). 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).

Secondly, there was no diagnostic utility when comparing children with ADHD and those in the referred but non-diagnosed group from the same hospital. The GAI-CPI difference scores of randomly selected children could not be used to differentiate between a child with ADHD and a non-diagnosed child at greater than chance levels. These results are consistent with past research by Weiss and Gable (2007) who found a lack of noteworthy GAI-CPI difference scores for children with ADHD. Specifically, in matched clinical and non-clinical groups children with ADHD could not be identified with TPR and TNR of at least 60% (Weiss & Gable).

Limitations

The first limitation to this study is the diagnoses given to participants. The hospital psychologists used a variety of methods to diagnose ADHD. Although each child in the hospital sample 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 DSM-IV-TR (2000) criteria. Additionally, many of the children with ADHD had co-morbid diagnoses. Co-morbidity, 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 to children with the primarily inattentive type of ADHD (Schwean & McCrimmon, 2008). Additionally, 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 the result of the ROC analysis comparing children with ADHD to the simulated WISC-IV standardization sample had an AUC of .65, 95% CI [0.57, 0.72]. Furthermore, the ROC analysis comparing children with ADHD to the non-diagnosed hospital comparison sample had an AUC of .47, 95% CI [0.36, 0.57]. These results are similar to the analyses that included the ADHD-NOS sample.

A further 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 sample was collected from a specific hospital instead of selected randomly. This procedure resulted in the sample being demographically and regionally

28

limited. Additionally, 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 being dependent upon cut scores or base rates (Swets, 1988).

Implications

Although the study results should be considered preliminary due to the 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 low 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: American Psychiatric Association.

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. (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. (1985). A visicalc program for estimating the area under a Receiver Operating Characteristic (ROC) curve. *Medical Decision Making, 5,* 139-148.

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.

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

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.

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.

Rice, M. E., & Harris, G. T. (2005). Comparing effect sizes in follow-up studies: ROC area, Cohen's d, and r. *Law and Human Behavior, 29,* 615-620.

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). NY: Guilford.

34

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 Assessmen*t: *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.

Wu, S., & Flach, P. (2005, August). *A scored AUC metric for classifier evaluation and selection*. Paper session presented at the meeting ICML, Bonn, Germany.

APPENDIX A

IRB APPROVAL LETTER

To:	**Marley Watkins**
	EDUC - I.
From:	**Mark Roosa, Chair**
	Soc Beh IRB
Date:	03/25/2009
Committee Action:	**Exemption Granted**
IRB Action Date:	03/25/2009
IRB Protocol #:	0903003827
Study Title:	Psychometric Properties of the WISC-IV Among Arizona Students

The above-referenced protocol is considered exempt after review by the Institutional Review Board pursuant to Federal regulations, 45 CFR Part 46.101(b)(4) .

This part of the federal regulations requires that the information be recorded by investigators in such a manner that subjects cannot be identified, directly or through identifiers linked to the subjects.   It is necessary that the information obtained not be such that if disclosed outside the research, it could reasonably place the subjects at risk of criminal or civil liability, or be damaging to the subjects' financial standing, employability, or reputation.

You should retain a copy of this letter for your records.