Copyright © by LDW 2005

Diagnostic Validity of Wechsler Subtest Scatter

Marley W. Watkins[1]
Pennsylvania State University

Cognitive subtest scatter has often been considered to be diagnostically significant. The current study tested the diagnostic validity of four separate operationalizations of WISC-III subtest scatter: (a) range of verbal, performance, and full-scale subtests; (b) variance of verbal, performance, and full-scale subtests; (c) number of subtests deviating by ±3 points from verbal, performance, and full-scale average; and (d) Mahalanobis distance of full-scale subtests. The WISC-III normative sample was compared to 1,592 students with learning disabilities (LD). Scatter indices were strongly correlated with each other (i.e., r = .81 to .95). Receiver operating characteristic (ROC) curve analyses revealed that using any of the subtest scatter indices to diagnose LD resulted in correct decisions only 50% to 55% of the time. Chance would afford similar levels of accuracy.

Keywords: WISC-III; Wechsler Intelligence Scale, Subtest Analysis, Diagnostic Decision-Making, Learning Disabilities

It has long been speculated that cognitive subtest scatter or variability is a pathognomonic sign. Early researchers hypothesized that subtest scatter would predict scholastic potential or membership in exceptional groups (Harris & Shakow, 1937). In fact, Binet suggested that scattered passes and failures on the Binet-Simon scale were diagnostically significant (Matarazzo, 1985), and Wechsler (1941) conjectured that subtests discrepant from the mean of the Wechsler Bellevue scale might be useful in differential diagnosis. Similar propositions have been proffered for each successive revision of the Wechsler and Stanford-Binet scales (Watkins, 2003).

Across time, subtest scatter has been quantified in four ways. The first method was the range (i.e., the difference between an examinee's highest and lowest subtest scaled scores). Variance, the second quantification of subtest scatter, was thought to be more useful than range because it utilizes information from all subtests, rather than just the highest and lowest (Wilcox, 1996). Labeled the profile variability index (PVI) by Plake, Reynolds, and Gutkin (1981), this scatter index is computed by applying the sample variance formula to the subtest scores of an individual examinee. Third, Mahalanobis distance (D^2), a multivariate method, has been recommended for its statistical superiority over univariate measures such as range and variance (Huba, 1985). Given that subtest scores are correlated, overall subtest scatter will be lower than if the subtests were uncorrelated (Burgess, 1991). This is corrected by D^2, which takes into account the means and variances of each subtest and the correlations and covariance between subtests. Finally, researchers have looked at the number of

[1] Address correspondence to Marley W. Watkins, Ph.D., Pennsylvania State University, Department of Educational and School Psychology and Special Education, 125 CEDAR Building, University Park, PA 16802. Email: mww10@psu.edu

subtests differing by ±3 points from the individual examinee's own mean score (Schinka, Vanderploeg, & Curtiss, 1997). While range and variance are normative metrics, this final quantification of subtest scatter is an ipsative measure, comparing an examinee's subtest variability with that examinee's own mean performance.

Research on subtest scatter with previous Wechsler tests has been unproductive (Sattler, 2001). For example, based on their qualitative analysis of decades of IQ subtest scatter research, Kramer, Henning-Stout, Ullman, and Schellenberg (1987) found no evidence that subtest scatter uniquely identified any diagnostic group. A quantitative combination of results from 94 studies (N = 9,372) also demonstrated that subtest scatter failed to uniquely distinguish children with learning disabilities (LD) (Kavale & Forness, 1984). Similarly, a narrative review of 70 years of research on subtest scatter arrived at pessimistic conclusions about its diagnostic accuracy (Zimmerman & Woo-Sam, 1985).

More recently, Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) range and PVI indices were unable to exhibit adequate diagnostic accuracy for students with LD (Daley & Nagle, 1996; Kline, Snyder, Guilmette, & Castellanos, 1993; Watkins, 1999). The number of WISC-III subtests deviating from an individual examinee's mean score also failed to demonstrate accurate discrimination of students with and without LD (Watkins & Worrell, 2000). Research on the D^2 operationalization of scatter has not been conducted with modern IQ tests. One study of its utility among a sample of 200 healthy U.K. adults with the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981) found distributions very similar to those expected from a normal population (Crawford & Allan, 1994).

Despite the negative research evidence, tables of subtest scatter range are included in both WISC-III and Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997) manuals. Accompanying these tables is the comment that subtest scatter is "frequently considered as diagnostically significant" (Wechsler, 1991, p. 177). Schinka, Vanderploeg, and Curtiss (1997) provided additional tables of subtest scatter, and contemporary texts encourage the formation of diagnostic hypotheses based on subtest scatter (Kellerman & Burry, 1997). Additionally, subtest scatter continues to be identified as a diagnostic indicator of learning disabilities (Nielsen, 2002). Thus, clinical interest in subtest scatter remains high. The present study was conducted to evaluate the diagnostic accuracy of several operationalizations of subtest scatter among a large national sample of students with LD.

Method

Instruments

The WISC-III is an individually administered measure of intellectual functioning designed to assess children from ages 6 years, 0 months to 16 years, 11 months. It consists of 13 individual subtests (M = 10, SD = 3), 10 standard and 3 supplementary, that combine to yield three composite scores: Verbal (VIQ), Performance (PIQ), and Full-Scale (FSIQ) IQ (M = 100, SD = 15). In addition, the WISC-III provides four factor-based index scores: Verbal Comprehension (VC), Perceptual Organization (PO), Freedom from Distractibility (FD), and Processing Speed (PS) (M = 100, SD = 15). Full details of the WISC-III and its standardization are presented in Wechsler (1991). Additional reliability and validity data are provided by Sattler (2001) as well as Zimmerman and Woo-Sam (1997).

Academic achievement was measured by 62 tests or combinations of tests. The Woodcock-Johnson Tests of Achievement and the Wechsler Individual Achievement Test were applied in around 85% of the cases. Achievement in reading and math was first determined by averaging the reading achievement scores (i.e., basic reading skills and reading comprehension subtests) and math achievement scores (i.e., math computation and math reasoning subtests). If only one achievement score was provided in reading or math, that score was used.

Procedure

Requests to contribute to an investigation of the WISC-III were mailed to 9,227 school psychologists throughout the United States. Invitees all worked in school settings and were members of the National Association of School Psychologists. The school psychologists were asked to report anonymous data from their five most recent evaluations that resulted in special education placement under the following categories: learning disability (LD), emotional disability, or mental retardation. Responses were received from 492 school psychologists from 47 states. Respondents worked in rural (31.1%), urban (21.7%), suburban (36.4%), and mixed (5.9%) school districts and ranged in years of experience from 1 year to 37 years ($M = 12.26$; $SD = 8.35$). Forty-three percent of the respondents held a master's degree, 17% possessed a doctoral degree, and 39% were trained at the specialist level.

Participants

Anonymous scores were reported on 2,356 students; however, 28 could not be retained due to insufficient or invalid data. The remaining 2,328 students included 1,611 identified as LD, 265 identified as emotionally disturbed, 258 identified as mentally retarded, and 194 with other or multiple diagnoses. Identification of students in these special education categories was accomplished according to diagnostic criteria used in the local setting.

Students with learning disabilities. Scores on the 10 mandatory subtests (Picture Arrangement, Block Design, Object Assembly, Picture Completion, Coding, Information, Similarities, Vocabulary, Comprehension, and Arithmetic) were necessary to compute scatter indices. Consequently, WISC-III data from 1,592 students with a local diagnosis of LD (1,093 male and 499 female) who had all 10 subtest scores on record were included in the study. No other selection criteria (e.g., VIQ-PIQ differences, low IQ scores.) were used.

Based upon local diagnostic criteria, students were classified as having an LD in reading alone ($n = 252$); math alone ($n = 171$); written expression alone ($n = 285$); reading and math ($n = 82$); reading and written expression ($n = 457$); math and written expression ($n = 89$); reading, math, and written expression ($n = 233$); and unspecified ($n = 23$). Students ranged in age from 6 to 16 years ($Mdn = 9$) and grade from kindergarten to 11 ($Mdn = 4$). Students' ethnic background was 74.6% White, 8.4% Hispanic, 12.6% Black, 1.8% Native American, .8% Asian/Pacific, .7% other, and 1.1% unspecified. Students represented 47 states and were enrolled in rural (31.5%), urban (19.8%), suburban (36.9%), mixed (6.9%), and unspecified (4.8%) school districts. Level of parental education, as reported by responding school psychologists, included primary education (4.8%), high-school education (43.1%), some college education (15.4%), college education (10.1%), graduate education (3.5%), and unspecified (23.1%).

Students with specific learning disability in reading. Given the imprecision of LD diagnosis in practice (Kavale, Fuchs, & Scruggs, 1994), a subsample of students with specific reading disabilities was selected from the larger group of 1,592 students with a LD. These students were identified by the following criteria: (a) diagnosed as having a LD in reading or both reading and writing by the local multidisciplinary team; (b) not diagnosed as having an LD in math by the local multidisciplinary team; (c) exhibiting a significant discrepancy between expected and obtained reading achievement based on regression analysis of FSIQ and reading test scores with a 95% confidence interval (Reynolds, 1984-85); and (d) displaying nonsignificant discrepancy between expected and obtained math achievement based on regression analysis of FSIQ and math test scores. Based upon these criteria, 600 students were identified as having a specific reading disability.

Students with low reading achievement. Given the recent emphasis on low achievement and response to treatment in the definition of LD (Fuchs, Fuchs, McMaster, & Otaiba, 2003), a low reading achievement subsample was also identified. Students who were determined to be LD by local multidisciplinary evaluation teams and who scored ≤ 85 in reading achievement were included in this subsample. Following these methods, 846 students were selected.

Students without disabilities. The WISC-III normative sample served as the contrast group for this investigation (Wechsler, 1991). It comprised 2,200 children, 100 males and 100 females in each of 11 age groups, ranging from 6 years, 0 months to 16 years, 11 months. The standardization sample was stratified by age, gender, race/ethnicity, geographic location, and parent education according to the 1988 U.S. Census.

Analyses

Subtest scatter. First, the total range was calculated for each child. This was based on the highest subtest score minus the lowest. Also, the range within the five verbal and five performance subtests was calculated. Thus, both the highest verbal subtest score minus the lowest verbal subtest score and the highest performance subtest score minus the lowest performance subtest score were computed. Second, the PVI for total, verbal, and performance subtests was computed for each child. Third, the number of subtests deviating by ±3 points from the total, verbal, and performance mean of each child was calculated. Finally, the D^2 for all 10 subtests was computed for each child, based on the means and covariances of the WISC-III standardization sample. When the means and covariance of a Wechsler normative sample are used to represent the population, the D^2 for an individual examinee represents the probability that the individual's subtest scores came from the standardization sample (Burgess, 1991). D^2 for verbal and performance dimensions were not calculated because only five subtests were included in each area.

Diagnostic validity. Most research on subtest scatter has relied on a classical validity paradigm (Wiggins, 1988). That is, the mean scatter difference between a group of students with LD and a group without LD is tested statistically. If significant group differences are found, then scatter is deemed diagnostically useful. However, this methodology is faulty in several respects. For example, it confuses a priori with posterior odds, and it does not take into account the overlap in score distributions between the two groups (McFall & Treat, 1999). As noted by Weiner (2003), "mean differences between two groups obtained in nomothetic research, even when statistically significant by usual standards, rarely have sufficient predictive power to support reliable inferences in idiographic appraisals, this is, in deciding whether a particular person should be classified as belonging to one group or another" (p. 336).

In contrast, diagnostic validity directly assesses the idiographic aspect of making individual diagnostic decisions. The most popular diagnostic validity statistics include sensitivity, specificity, positive predictive power, and negative predictive power (Hsu, 2002). Although these indices provide valuable information about a diagnostic test, they have several limitations. Most important, each is influenced by the prevalence of the disability and the cutoff value used on a given test (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000). Thus, if prevalence rates or cutoff values vary, then sensitivity, specificity, and predictive values might also change (Meehl & Rosen, 1955). This limitation is important because the two groups compared in this study have relatively similar prevalence rates. That is, there was a large number of students with LD. In actual practice, the number of children with LD would be much smaller than those without LD (U.S. Department of Education, 2001); thus, relative prevalence rates would be quite different than those suggested by the present analysis.

To ameliorate these limitations, receiver operating characteristic (ROC) curve methods were applied. The ROC is independent of prevalence rates and cutoff values (McFall & Treat, 1999). Essentially, a ROC is a graph of the percentage of true positive decisions against the percentage of false positive decisions across all possible cutoff values. ROC analysis involves calculating true positive and false positive rates across an entire range of cutoff scores, plotting the resulting pairs of true positive and false positive rates to form a curve, and calculating the area under the curve (AUC), which provides an overall accuracy index of the test (Henderson, 1993). Given that scatter indices are not likely to be normally distributed, AUC were calculated using a nonparametric formula (Hanley & McNeil, 1982).

AUC values can range from 0.5 to 1.0. An AUC value of 0.5 signifies that the true positive rates and false positive rates are equal across all possible cutoff scores and that no discrimination exists (McFall & Treat, 1999; Swets, 1988). In that case, the ROC curve lies on the main diagonal of the graph and the diagnostic system is functioning at the level of chance. In contrast, an AUC value of 1.0 denotes perfect discrimination. Accordingly, AUC values of 0.5 to 0.7 indicate low test accuracy, 0.7 to 0.9 signify moderate test accuracy, and 0.9 to 1.0 represent high test accuracy (Swets, 1988). In this study, the AUC can be interpreted as the probability of correctly identifying the student with LD if two students are drawn at random, one from the sample of students with LD and one from the sample without LD.

Results

As expected, the students with LD performed lower than the normative sample on WISC-III IQ scores (Table 1). Statistically significant differences were found between groups on several scatter indices, but only the range for all 10 subtests and the PVI for all 10 subtests were statistically significantly different from the normative sample (to control for multiple tests, individual indices were tested at $p \leq .005$) across all three clinical groups. However, standardized mean difference effect sizes for these significant tests were small (Kraemer et al., 2003), averaging .17.

Table 1

Means and Standard Deviations of WISC-III Index Scores and Scatter Indices for Normative and Clinical Samples

	Sample							
	Normative		All LD		Specific LD		Low Achievement	
Index	Mean	*SD*	Mean	*SD*	Mean	*SD*	Mean	*SD*
VIQ	100.0	15.0	93.4*	13.1	94.4*	13.7	88.4*	11.9
PIQ	100.0	15.0	97.1*	14.2	99.7	13.4	94.3*	12.8
FSIQ	100.0	15.0	94.5*	12.5	96.4*	12.7	90.3*	10.9
Reading	–	–	84.5	12.5	78.4	9.2	75.6	7.7
Math	–	–	89.8	12.2	94.8	10.0	86.4	10.9
No. subtests ≥ 3 points	1.2	1.2	1.3	1.2	1.3	1.1	1.3	1.1
VIQ range	4.6	2.0	4.9*	2.0	4.7	1.9	4.8*	1.9
PIQ range	6.0	2.4	6.0	2.4	6.0	2.4	5.8	2.3
FSIQ range	7.5	2.3	7.9*	2.3	7.8*	2.3	7.8*	2.3
VIQ PVI	4.1	3.3	4.5*	3.5	4.3	3.1	4.3	3.1
PIQ PVI	6.8	5.1	6.7	5.0	6.8	5.0	6.4	4.7
FSIQ PVI	6.0	3.3	6.8*	3.8	6.6*	3.9	6.6*	3.7
D^2	10.0	4.9	10.8*	5.1	10.4	4.9	10.8*	5.0
N	2,200		1,592		600		846	

Note. VIQ = Verbal IQ, PIQ = Performance IQ, FSIQ = Full-Scale IQ, PVI = Profile Variability Index, D^2 = Mahalanobis distance.

* $p < .005$.

Range and variance measures of scatter were strongly related. Across samples, for example, the range and variance quantifications of scatter were correlated at .91 to .95, and the D^2 index was correlated .81 to .88 with range and PVI indices, respectively. Given these relationships, it was anticipated that all scatter indices would produce similar diagnostic validity results.

As hypothesized, diagnostic validity results were consistent: None of the subtest scatter indices was able to identify students with LD at a rate appreciably better than chance (see Table 2). Based on the categorization system of Swets (1988), all had low diagnostic accuracy. For a graphic illustration, the ROC curve for the normative group-clinical group comparison with the highest AUC value, the PVI of all 10 subtests for the 1,592 students with LD, is displayed in Figure 1. As illustrated, diagnostic accuracy remained near the line of chance discrimination across all possible cut scores.

Discussion

Cognitive subtest scatter is often considered clinically significant, perhaps even a sign of brain dysfunction (Mitrushina et al., 1994) or LD (McLean, Reynolds, & Kaufman, 1990). Pursuant to this hypothesis, subtest scatter has been quantified in a variety of ways. The current study tested the diagnostic validity of four separate operationalizations of WISC-III subtest scatter among 1,592 students with LD. Scatter indices were strongly correlated with each other (i.e., $r = .81$ to .95). Although some statistically significant mean group differences emerged, scatter indices did not accurately identify children with LD. That is, using any of the subtest scatter indices to diagnose LD resulted in correct decisions only 50% to 55% of the time. Chance would afford similar accuracy. Thus, contrary to the WISC-III manual, subtest scatter was not "diagnostically significant" (Wechsler, 1991, p. 177) for LD.

Table 2

AUC Statistics for Normative Group-Clinical Group Scatter Comparisons

	Sample		
Scatter Index	All LD	Specific LD	Low Achievement
No. subtests ≥ 3 points	.52	.51	.51
VIQ range	.54	.52	.53
PIQ range	.50	.50	.50
FSIQ range	.54	.53	.53
VIQ PVI	.54	.52	.53
PIQ PVI	.50	.50	.50
FSIQ PVI	.55	.53	.54
D^2	.55	.52	.55

Note. VIQ = Verbal IQ, PIQ = Performance IQ, FSIQ = Full-Scale IQ, PVI = Profile Variability Index, D^2 = Mahalanobis distance.

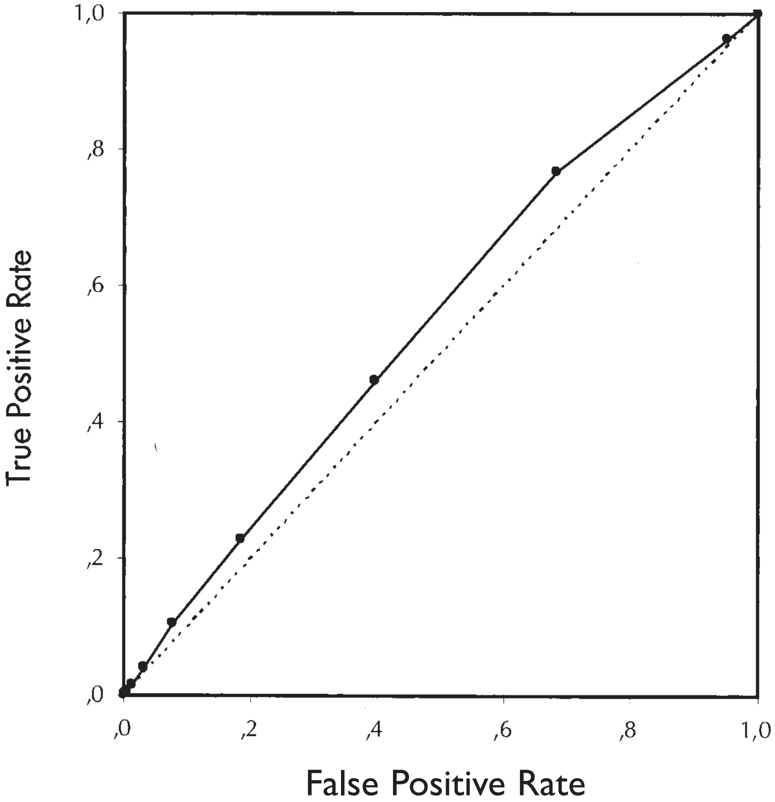

Figure 1. Receiver operating characteristic (ROC) curve (solid line) for the profile variability index across 10 WISC-III subtests for the 1,592 students with LD compared to the 2,200 students in the normative sample (AUC = .55) with chance represented by the dotted diagonal line.

Some authors acknowledge that subtest scatter is an inaccurate diagnostic indicator, yet advocate its use for developing hypotheses about an examinee's cognitive strengths and weaknesses (Sattler, 2001). For example, Kaufman and Lichtenberger (1998) asserted that, "through research knowledge, theoretical sophistication, and clinical ability examiners must generate hypotheses about an individual's assets and deficits and then confirm or deny these hypotheses by exploring multiple sources of evidence" (p. 192). However, to generate hypotheses about a child's functioning from subtest scatter would be the equivalent of relying on the flip of a coin or the roll of a die. Accordingly, this unsound practice should be abjured.

Marley W. Watkins *is Professor-in-Charge of Graduate Programs in School Psychology at The Pennsylvania State University and a Diplomate of the American Board of Professional Psychology. He is a graduate of the University of Nebraska-Lincoln school psychology program and was a practicing school psychologist for 15 years before entering academe.*

References

Burgess, A. (1991). Profile analysis of the Wechsler intelligence scales: A new index of subtest scatter. *British Journal of Clinical Psychology, 30,* 257–263.

Crawford, J. R., & Allan, K. M. (1994). The Mahalanobis distance index of WAIS-R subtest scatter: Psychometric properties in a healthy UK sample. *British Journal of Clinical Psychology, 33,* 65–69.

Daley, C. E., & Nagle, R. J. (1996). Relevance of WISC-III indicators for assessment of learning disabilities. *Journal of Psychoeducational Assessment, 14,* 320–333.

Fuchs, D., Fuchs, L. S., McMaster, K. N., & Otaiba, S. A. (2003). Identifying children at risk for reading failure: Curriculum-based measurement and the dual-discrepancy approach. In H. L. Swanson, K. R. Harris, & S. Graham (Eds.), *Handbook of learning disabilities* (pp. 431–449). New York: Guilford.

Hanley, J. D., & McNeil, B. J. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology, 143,* 29–36.

Harris, A. J., & Shakow, D. (1937). The clinical significance of numerical measures of scatter on the Stanford-Binet. *Psychological Bulletin, 34,* 134–150.

Hsu, L. M. (2002). Diagnostic validity statistics and the MCMI-III. *Psychological Assessment, 14,* 410–422.

Huba, G. J. (1985). How unusual is a profile of test scores? *Journal of Psychoeducational Assessment, 4,* 321–325.

Kaufman, A. S., & Lichtenberger, E. O. (1998). Intellectual assessment. In A. S. Bellack & M. Hersen (Eds.), *Comprehensive clinical psychology: Assessment* (Vol. 4, pp. 187–238). New York: Elsevier.

Kavale, K. A., & Forness, S. R. (1984). A meta-analysis of the validity of Wechsler scale profiles and recategorizations: Patterns or parodies? *Learning Disability Quarterly, 7,* 136–156.

Kavale, K. A., Fuchs, D., & Scruggs, T. E. (1994). Setting the record straight on learning disability and low achievement: Implications for policymaking. *Learning Disabilities Research & Practice, 9,* 70–77.

Kellerman, H., & Burry, A. (1997). *Handbook of psychodiagnostic testing: Analysis of personality in the psychological report* (3rd ed.). Boston: Allyn and Bacon.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1993). External validity of the profile variability index for the K-ABC, Stanford-Binet, and the WISC-R: Another cul-de-sac. *Journal of Learning Disabilities, 26,* 557–567.

Kraemer, H. C., Morgan, G. A., Leech, N. L., Gliner, J. A., Vaske, J. J., & Harmon, R. J. (2003). Measures of clinical significance. *Journal of the American Academy of Child and Adolescent Psychiatry, 42,* 1524–1529.

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

Matarazzo, J. D. (1985). Psychological assessment of intelligence. In H. I. Kaplan & B. J. Sadock (Eds.), *Comprehensive textbook of psychiatry* (Volume 1, 4th ed., pp. 502–513). Baltimore: Williams & Wilkins.

McFall, R. M., & Treat, T.A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215–241.

McLean, J. E., Reynolds, C. R., & Kaufman, A. S. (1990). WAIS-R subtest scatter using the profile variability index. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 289–292.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Mitrushina, M., Drebing, C., Satz, P., Gorp, W. V., Chervinsky, A., & Uchiyama, C. (1994). WAIS-R intersubtest scatter in patients with dementia of Alzheimer's type. *Journal of Clinical Psychology, 50,* 753–758.

Nielsen, M. E. (2002). Gifted students with learning disabilities: Recommendations for identification and programming. *Exceptionality, 10,* 93–111.

Plake, B. S., Reynolds, C. R., & Gutkin, T. B. (1981). A technique for the comparison of profile variability between independent groups. *Journal of Clinical Psychology, 37,* 142–146.

Reynolds, C. R. (1984–85). Critical measurement issues in learning disabilities. *Journal of Special Education, 18,* 451–476.

Sattler, J. M. (2001). *Assessment of children: Cognitive applications* (4th ed.). San Diego: Jerome M. Sattler.

Schinka, J. A., Vanderploeg, R. D., & Curtiss, G. (1997). WISC-III subtest scatter as a function of highest subtest scaled score. *Psychological Assessment, 9,* 83–88.

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

Swets, J. A., Dawes, R. M., & Monahan, J. (2000). Psychological science can improve diagnostic decisions. *Psychological Science in the Public Interest, 1,* 1–26.

U.S. Department of Education. (2001). *Twenty-first annual report to congress on the implementation of the Individuals with Disabilities Education Act.* Jessup, MD: Author.

Watkins, M. W. (1999). Diagnostic utility of WISC-III subtest variability among students with learning disabilities. *Canadian Journal of School Psychology, 15,* 11–20.

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

Watkins, M. W., & Worrell, F. C. (2000). Diagnostic utility of the number of WISC-III subtests deviating from mean performance among students with learning disabilities. *Psychology in the Schools, 37,* 303–309.

Wechsler, D. (1941). *The measurement of adult intelligence* (2nd ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-Revised.* New York: Psychological Corporation.

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

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale-Third Edition.* San Antonio, TX: Psychological Corporation.

Weiner, I. B. (2003). Prediction and postdiction in clinical decision making. *Clinical Psychology: Science and Practice, 10*, 335–338.

Wiggins, J. S. (1988). *Personality and prediction: Principles of personality assessment.* Malabar, FL: Krieger.

Wilcox, R. R. (1996). *Statistics of the social sciences.* New York: Academic Press.

Zimmerman, I. L., & Woo-Sam, J. M. (1985). Clinical applications. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 873–898). New York: Wiley.

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

Author's Note

This study was made possible by a research award from the Society for the Study of School Psychology. Standardization data of the Wechsler Intelligence Scale for Children-Third Edition copyright © 1990 by The Psychological Corporation. Used by permission. All rights reserved.

Received July 20, 2004
Revised October 16, 2004
Accepted October 17, 2004