(2005). D. P. Flanagan & P. L. Harrison (Eds.). Contemporary intellectual assessment: Theories, tests, and issues (2nd ed.; pp. 251-268). New York: Guilford.

12

Issues in Subtest Profile Analysis

MARLEY W. WATKINS
JOSEPH J. GLUTTING
ERIC A. YOUNGSTROM

More than 250 million standardized tests are administered to public school students each year in the United States (Salvia & Ysseldyke, 1998). Although much school-based testing is accomplished in groups, a substantial number of standardized tests are also employed in individual evaluations. For example, millions of children served in special education programs have participated in individual psychoeducational evaluations (U.S. Department of Education, 2001). Individual evaluations to determine special education eligibility often include a standardized measure of intellectual functioning. Of the available individual intelligence tests, the Wechsler scales are the most frequently used (Kaufman & Lichtenberger, 2000). Among school-age children, the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949), Wechsler Intelligence Scale for Children—Revised (WISC-R; Wechsler, 1974), and Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) have been the most popular (Kamphaus, Petoskey, & Rowe, 2000; Oakland & Hu, 1992) for decades.

Typically, interpretation of individual intelligence tests is based on a hierarchical, top-down model that first considers the global IQ scores. Next, to extract more information from the test, distinct patterns or profiles of subtest scores are analyzed. Finally, individual subtests are considered in isolation. This practice of interpreting the pattern of subtest scores attained by children on individual measures of intelligence is known as *subtest profile analysis*. Based on these principles, elaborate subtest interpretation systems (Kaufman, 1994; Sattler, 2001) have achieved wide popularity in school psychology training and practice (Alfonso, Oakland, LaRocca, & Spanakos, 2000; Groth-Marnat, 1997; Kaufman, 1994; Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). For example, approximately 74% of school psychology training programs place moderate to great emphasis on the use of subtest scores in their individual cognitive assessment courses (Alfonso et al., 2000). Among a sample of school psychologists, Pfeiffer and colleagues (2000) found that almost 70% reported profile analysis to be a useful feature of the WISC-III, and 29% reported that they derived specific value from individual WISC-III subtests.

PURPOSES

Profile analysis has typically been applied for two major purposes: (1) diagnostically discriminating average and exceptional children, and (2) identifying specific cognitive strengths and weaknesses. Wechsler (1958)

himself may have encouraged the process of diagnostically interpreting children's subtest profiles when he advanced the hypothesis that childhood schizophrenia could be identified by high scores on the Picture Completion and Object Assembly subtests and low scores on the Picture Arrangement and Digit Symbol subtests. Eventually, more than 75 patterns of subtest variation were identified for the Wechsler series alone (McDermott, Fantuzzo, & Glutting, 1990). Currently, many different subtest profiles are purportedly diagnostic of learning and emotional problems (e.g., Kaufman, 1994).

To identify cognitive strengths and weaknesses, specific patterns of subtest scores are presumed to substantially invalidate global intelligence indices (Groth-Marnat, 1997), so that subtest patterns, rather than IQ composites, become the focus of interpretation. Subtests that are significantly higher or lower than a child's own average are considered relative strengths or weaknesses, respectively. Next, hypotheses concerning the underlying causes of significant subtest variations are identified by locating abilities thought to be shared by two or more subtests. Extensive lists of the abilities presumed to underlie each subtest are provided by Kaufman (1994), Sattler (2001), and Kaufman and Lichtenberger (2000). Finally, these hypotheses are used to generate educational and psychological interventions and remedial suggestions (Zeidner, 2001).

BRIEF HISTORICAL REVIEW OF PROFILE ANALYSIS RESEARCH

Scatter

Attempts to analyze IQ subtest variations date back to the inception of standardized intelligence testing (Zachary, 1990). For example, Binet suggested that the passes and failures of psychotic or alcoholic examinees would exhibit more *scatter* across age levels on his scale than would other examinees (see Matarazzo, 1985). Early researchers hypothesized that subtest scatter would predict scholastic potential or membership in exceptional groups (Harris & Shakow, 1937). Uneven subtest scores were assumed to be signs of pathology or of greater potential than indicated by averaged IQ composites. Hundreds of studies were conducted to test these hypotheses. Based on their 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, and they opined that "we regard scatter analysis as inefficient and inappropriate" (p. 45). A narrative review of 70 years of research on subtest scatter also arrived at pessimistic conclusions concerning its diagnostic accuracy (Zimmerman & Woo-Sam, 1985). Although isolated studies found abnormal scatter within clinical groups, differences tended to disappear when adequate comparison samples were used. Zimmerman and Woo-Sam (1985) observed that "extensive scatter proved to be both typical and 'normal,' and thus of limited use as a diagnostic feature" (p. 878).

Similar negative results were reported in a longitudinal study of the WISC-R. Moffitt and Silva (1987) concluded that perinatal, neurological, and health problems did not cause extreme Verbal IQ–Performance IQ (VIQ-PIQ) discrepancies, and found that neither behavior problems nor motor problems were significantly related to VIQ-PIQ scores. Furthermore, VIQ-PIQ score discrepancies were unreliable across time. That is, the majority of children with extreme VIQ-PIQ score discrepancies did not maintain such a large difference when tested with the WISC-R 2 years later. Thus "VIQ-PIQ discrepancies are of doubtful diagnostic value" (Moffitt & Silva, 1987, p. 773).

The quantitative combination of results from 94 studies (N = 9,372) also demonstrated that subtest scatter and scatter between VIQ and PIQ failed to uniquely identify children with learning disabilities (LDs) (Kavale & Forness, 1984). For example, the average VIQ-PIQ difference for children with LDs was only 3.5 points—a difference found in 79% of the normal population. In sum, subtest scatter was determined to be of "little value in LD diagnosis" (Kavale & Forness, 1984, p. 139).

Subtest Profiles

Given the popularity of the Wechsler scales, Wechsler subtest profiles have been the source of much research. For example, the validity of using WISC and WISC-R subtests for diagnosing LDs was the focus of a meta-

analysis by Kavale and Forness (1984). This quantitative summary of 94 studies revealed that "the differential diagnosis of LD using the WISC, although intuitively appealing, appears to be unwarranted" because "regardless of the manner in which WISC subtests were grouped and regrouped, no recategorization, profile, pattern, or factor cluster emerged as a 'clinically' significant indicator of LD" (Kavale & Forness, 1984, p. 150).

Taking another approach, Mueller, Dennis, and Short (1986) statistically clustered the WISC-R subtest data of 119 samples of nonexceptional and exceptional children (N = 13,746) to determine whether profiles would emerge that were diagnostically characteristic of various disabilities. Results indicated that WISC-R subtest profiles were typically marked by general intellectual level but could not reliably distinguish among diagnostic groups. Like Kavale and Forness (1984), Mueller and colleagues concluded that Wechsler subtest profiles were not helpful in differentiating among children with emotional and learning impairments, and they recommended that IQ tests be used only to estimate global intellectual functioning.

The poor diagnostic accuracy of subtest profiles has generalized across tests and cultures. For example, Rispens and colleagues (1997) analyzed the ability of subtest profiles from the Dutch version of the WISC-R to distinguish among 511 children with conduct disorder, mood disorder, anxiety disorder, attention deficit disorder, and other psychiatric disorders. Rispens and colleagues found that subtest patterns did not significantly differ across the various groups and concluded that "WISC profiles . . . cannot contribute to differential diagnosis" (p. 1593).

Subtest profiles have also been found to be unexceptional markers when empirically generated subtest profiles serve as a normative standard against which subtest profiles obtained from clinical groups are compared. If subtest profiles are markers of disability, then unique profiles should be found in the sample with disabilities. On the other hand, if subtest profiles are not distinctive of disability, profiles from the sample without disabilities should replicate for the sample with disabilities. To test this hypothesis, Watkins and Kush (1994) applied normative WISC-R subtest profiles (McDermott, Glutting, Jones, Watkins, & Kush, 1989) to 1,222 students with LDs, emotional handicaps, and mental retardation. They found that 96% of the children with disabilities displayed subtest profiles that were similar to those of the WISC-R standardization sample. No statistical or logical patterns could be detected in the subtest scores of the 4% of students with disabilities who exhibited profiles dissimilar to those of the standardization sample.

Another normative comparison applied the Wechsler Adult Intelligence Scale—Revised (WAIS-R) standardization sample core profiles to 161 adults with brain damage and found that 82% exhibited typical or normal subtest profiles (Ryan & Bohac, 1994). Patients with unique profiles did not differ on the basis of age, education, or organic etiology, so the atypical profiles did not contribute any diagnostic information. The WAIS-R core profiles were also applied to 194 college students with LDs (Maller & McDermott, 1997). Almost 94% of these students were found to have normatively typical subtest profiles. Unique profiles were disparate and not indicative of subtypes of LDs.

Historical Review: Summary

Subtest scatter had been determined to be clinically ineffectual as early as 1937 (Harris & Shakow, 1937). Cautions concerning the accuracy of subtest profiles were frequent (McNemar, 1964; Simensen & Sutherland, 1974). By 1983, Frank was able to say that "in spite of the fact that the Wechsler looked like it would be ideal for a comparative study of the intellectual/cognitive behavior of various psychopathological types, 40 years of research has failed to support that idea" (p. 79). A cumulative body of research evidence has shown that neither subtest scatter nor subtest profiles demonstrate acceptable accuracy in discriminating among diagnostic groups.

METHODOLOGICAL, STATISTICAL, AND PSYCHOMETRIC PROBLEMS WITH SUBTEST PROFILE ANALYSIS

Fundamental methodological, statistical, and psychometric problems cause subtest analysis results to be more illusory than real

and to represent more of a shared professional myth (Faust, 1990) than clinically astute detective work (Kaufman, 1994). Although a large number of problems have been identified (Glutting, Watkins, & Youngstrom, 2003; Watkins, 2003), this chapter concentrates on four major issues: *reliability*, *ipsative measurement*, *group mean differences*, and *inverse probabilities*.

Reliability

The weak reliability of subtest scores has been repeatedly demonstrated. For example, none of the WISC-III subtests reached the internal-consistency reliability criterion of ≥ .90 recommended by Salvia and Ysseldyke (1998) for making decisions about individuals. Similar results were found on the Wechsler Adult Intelligence Scale—Third Edition (WAIS-III; Wechsler, 1997), where the median internal consistency of the subtests was .85, and the Wechsler Intelligence Scale for Children—Fourth Edition (WISC-IV; Wechsler, 2003), where the median internal consistency of the subtests was .86.

The stability of subtest scores across time has also been inadequate for individual decisions. After administering the Dutch version of the WISC-R to a group of children three times at 6-month intervals, Neyens and Aldenkamp (1996) concluded that, "subtest scores show only fair to good, or even poor stability, which suggests that changes in subtest scores, as stated by many researchers, should not [be taken] into account when evaluating the cognitive development of children of normal intelligence" (p. 168). More recently, Smith, Smith, Bramlett, and Hicks (1999) tested a small sample of rural school children with the WISC-III and again 3 years later. The median stability coefficient for the subtests was .59, compared to .83 for the composite IQ scores. Using a large national sample of students twice tested for special education eligibility ($N = 667$), Canivez and Watkins (1998) found a median subtest stability coefficient of .68 and a median stability coefficient of .87 for the IQ composites. Even short-term stability coefficients have been inadequate for individual decisions. For example, the median short-term stability of the Stanford–Binet Intelligence Scales, Fifth Edition (Roid, 2003) subtests was .825, in contrast to .90 for the Full Scale IQ (FSIQ) score, and the median WISC-IV short-term stability coefficient was .78 for subtests, compared to .89 for the FSIQ.

However, these published reliability coefficients are overestimates of the typical reliability of subtest scores, for at least two reasons. First, they were calculated by test companies or researchers who paid close attention to standardization procedures and double-checked scoring accuracy (Feldt & Brennan, 1993; Thorndike, 1997). In clinical practice, administration, clerical, and scoring errors are ubiquitous (Belk, LoBello, Ray, & Zachar, 2002). Second, internal-consistency and stability reliability coefficients do not take into account all types of measurement error. For example, Schmidt, Le, and Ilies (2003) found that the reliability of a measure of general mental ability was overestimated by about 7% when transient measurement error was ignored.

Even worse, subtest profile analysis involves the comparison of multiple difference scores. The reliability of the difference between two scores is lower than the reliability of either score alone (Feldt & Brennan, 1993). The increased error generated by the use of difference scores makes even the best subtest-to-subtest comparison unstable (e.g., the reliability of the WISC-III Block Design and Vocabulary subtests is .87, but the reliability of their difference is .76).

Ipsative Measurement

Many subtest interpretative systems move away from *normative* measurement and instead rely on *ipsative* measurement principles (Cattell, 1944). That is, subtest scores are subtracted from mean composite scores for an individual. Thereby, the scores are transformed into person-relative metrics and away from their original population-relative metric (McDermott et al., 1990). Ipsative measurement is concerned with how a child's subtest scores relate to his or her personalized, average performance, and discounts the influence of global intelligence (Jensen, 2002). For example, two hypothetical students' normative (population-relative) and ipsative (person-relative) scores are displayed in Table 12.1. These two students have identical ipsative scores, but their normative scores are very different.

TABLE 12.1. Normative (Population-Relative) and Ipsative (Person-Relative) Scores for Two Hypothetical Examinees

	Student A		Student B	
Subtest	Norm. score	Ipsative score	Norm. score	Ipsative score
A	3	−4	10	−4
B	7	0	14	0
C	11	+4	18	+4
Mean	7	0	14	0

The ipsative perspective holds intuitive appeal, because it seems to isolate and amplify aspects of cognitive ability. Nevertheless, transformation of the score metric from normative to ipsative is psychometrically problematic. For example, McDermott and colleagues (1990) demonstrated that the ipsatization of WISC-R scores produced a loss of almost 60% of that test's reliable variance. McDermott and Glutting (1997) replicated those results with the Differential Ability Scales (DAS; Elliott, 1990) and WISC-III. They found that, on average, ipsative scores lost two-thirds to three-fourths of the reliable information provided by normative scores.

This information loss has been concretely demonstrated by analyzing the stability of ipsative subtest profiles. The temporal stability of subtest profiles among 303 randomly selected children tested twice as part of WISC-R validation studies, and among 189 children twice tested for special education eligibility, was computed (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Classificatory stability of relative cognitive strengths and weaknesses identified by subtest elevations and depressions was near chance levels for both groups of children. Livingston, Jennings, Reynolds, and Gray (2003) also found that the stability of WISC-R subtest profiles across a 3-year test–retest interval was too low for clinical use.

Similar temporal instability of subtest patterns was found for 579 students who were twice tested for special education eligibility with the WISC-III across a 2.8-year interval (Watkins & Canivez, 2004). Based on 66 subtest composites, subtest-based strengths and weaknesses replicated across test–retest occasions at chance levels (median kappa = .02). Because subtest-based cognitive strengths and weaknesses are unreliable, recommendations based on them will also be unreliable.

Both practical and theoretical analyses suggest that the mathematical properties of ipsative methods are profoundly different from those of familiar normative methods (Hicks, 1970; McDermott et al., 1990, 1992). Thus ipsative subtest scores cannot be interpreted as if they possessed the psychometric properties of normative scores. Based on this evidence, Jensen (2002) concluded that "the ipsative use of test batteries like the Wechsler scales, the Kaufman scales, and the Stanford–Binet IV is nearly worthless" (p. 11).

Group Mean Differences

Identification of IQ subtest profiles has generally been based upon statistically significant group differences. That is, a group of children with a particular disorder is identified, and their mean subtest score is compared to the mean subtest score of a group of children without the disorder. Statistically significant subtest score differences between the two groups are subsequently interpreted as evidence that the profile is diagnostically accurate for individuals. However, differences in group mean scores may not support individual interpretation. Nor does statistical significance of group differences equate to individual discrimination. As noted by Elwood (1993), "significance alone does *not* reflect the size of the group differences nor does it imply the test can discriminate subjects with sufficient accuracy for clinical use" (p. 409; original emphasis).

This situation illustrates reliance on classical validity methods instead of the more appropriate clinical utility approach (Wiggins, 1988). Average group subtest score differences indicate that *groups* can be discriminated. This classical validity approach cannot be uncritically extended to conclude that mean group differences are distinctive enough to differentiate among *individuals*. Figures 12.1 and 12.2 illustrate this dilemma. They display hypothetical score distributions of children from nondisabled and disabled populations. Group mean differences are clearly discernible in both, but the overlap between distributions in Figure 12.2 makes it difficult to determine group mem-

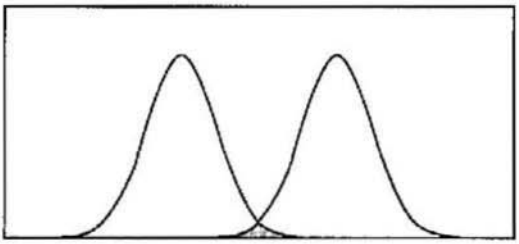

FIGURE 12.1. Hypothetical test score distributions from nondisabled (left) and disabled (right) populations that overlap (shaded region) only at the extremes.

bership accurately for those individuals in the shaded region. Although real score distributions are more similar to Figure 12.2, many researchers and clinicians act as if Figure 12.1 describes the relative score distributions.

There are four possible outcomes when one is using test scores to diagnose a disability: *true positive*, *true negative*, *false positive*, and *false negative*. Two outcomes are correct (true positive and true negative), and two are incorrect (false positive and false negative). True positives are children with disabilities who are correctly identified as such by the test. False positives are children identified by the test as having a disability who do not actually have one. In contrast, false negatives are children with disabilities who are not identified by the test as having disabilities. A test with a low false-negative rate has high sensitivity, and a test with a low false-positive rate has high specificity.

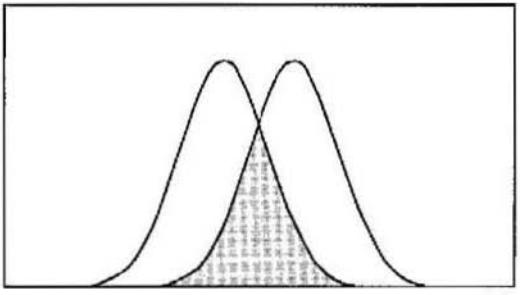

FIGURE 12.2. Hypothetical test score distributions from nondisabled (left) and disabled (right) populations that show considerable overlap (shaded region).

The relative proportion of correct and incorrect diagnostic decisions depends on the cut score used. For example, a cut score at the mean of the normal distribution in Figure 12.2 produces both a high true-positive and a high false-positive rate. That is, it correctly identifies those who are disabled, but it makes many mistakes for those who are not disabled. In contrast, a cut score at the mean of the disabled distribution makes few false-positive errors but many false-negative errors. Figure 12.2 graphically displays the tradeoffs between sensitivity and specificity that are always encountered when test scores are used to differentiate groups (Zarin & Earls, 1993).

Beyond cut scores, the accuracy of diagnostic decisions is dependent on the base rate or prevalence of the particular disability in the population being assessed. Very rare disabilities are difficult for a test to identify accurately (Meehl & Rosen, 1955). This issue is relevant for psychological practice and research, because many disabilities are by definition unusual or rare.

Although sensitivity and specificity are both desirable attributes of a diagnostic test, they are dependent on the cut score and prevalence rate. Thus neither provides a unique measure of diagnostic accuracy (McFall & Treat, 1999). In contrast, by systematically using all possible cut scores of a diagnostic test and graphing true-positive against false-positive decision rates, one can determine the full range of that test's diagnostic utility. Designated the *receiver operating characteristic* (ROC), this procedure was originally applied more than 50 years ago to determine how well an electronics receiver was able to distinguish signal from noise (Dawson-Saunders & Trapp, 1990). Because they are not confounded by cut scores or prevalence rates, ROC methods were subsequently widely adopted in the physical (Swets, 1988), medical (Dawson-Saunders & Trapp, 1990), and psychological (Swets, 1996) sciences. More recently, ROC methods were strongly endorsed for evaluating the accuracy of psychological assessments (McFall & Treat, 1999; Swets, Dawes, & Monahan, 2000).

As illustrated in Figure 12.3, the main diagonal of a ROC curve represents chance diagnostic accuracy. The more accurate the

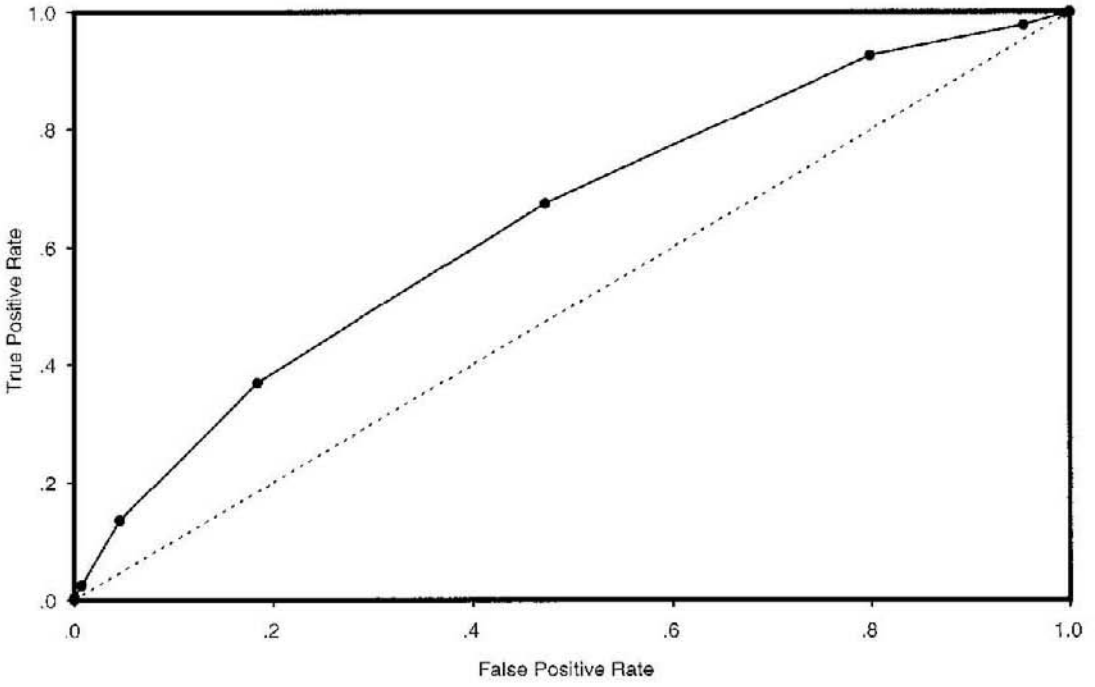

FIGURE 12.3. ROC curve for the WISC-III LDI for 445 students with LDs and 2,200 students without disabilities (AUC = .64). Data from Watkins, Kush, and Schaefer (2002); see this article for a full report of this study.

test, the further the ROC curve will deviate toward the upper left of the graph. This can be quantified by calculating the *area under the curve* (AUC), which provides an overall accuracy index of the test (Henderson, 1993). AUC values can range from .5 to 1.0. An AUC value of .5 signifies that the test has chance discrimination. In contrast, an AUC value of 1.0 denotes perfect discrimination. AUC values of .5 to .7 indicate low test accuracy, .7 to .9 indicate moderate test accuracy, and .9 to 1.0 indicate high test accuracy (Swets, 1988). More practically, the AUC can be interpreted in terms of two students: one drawn at random from the nondisabled population, and one randomly selected from the disabled population. The AUC is the probability that the test will correctly identify the student from the disabled population.

Clinical utility statistics (e.g., sensitivity, specificity, ROC) must be considered when one is evaluating the accuracy of test scores used to differentiate individuals. Group separation is necessary, but not sufficient, for accurate decisions about individuals.

Inverse Probabilities

Another fatal flaw in much subtest profile research is the use of inverse probabilities. Although related to the group–individual problem, it is particularly pernicious in clinical practice. Specifically, it is generally not understood that the probability of a particular score on a diagnostic test, given membership in a diagnostic group, is different from the probability of membership in a diagnostic group, given a particular score on a diagnostic test (McFall & Treat, 1999). For example, the probability of being a chronic smoker, given a diagnosis of lung cancer, is about .99, but the probability of having lung cancer, given chronic smoking, is only around .10 (Gambrill, 1990). That is, 99% of patients with lung cancer are chronic smokers, but only 10% of smokers develop lung cancer.

This quandary can be illustrated with a hypothetical subtest analysis example. A small group of children with LDs are located, and their WISC subtest scores are analyzed. It is found that many of these children exhibit a profile marked by relatively depressed scores on four specific subtests. Thus the probability of this subtest profile is high, given that a child has an LD. However, clinical use of subtest profiles is predicated on a different probability—namely, determining the probability that a referred child has an LD, given the subtest profile. Inverse probabilities systematically overestimate prospective accuracy (Dawes, 1993).

RECENT PROFILE ANALYSIS RESEARCH

Diagnosis

Scatter

Tables of subtest scatter have been included in the WISC-III, WAIS-III, and WISC-IV manuals. Accompanying these tables is the comment that subtest scatter has "frequently been considered diagnostically significant" (Wechsler, 2003, p. 46). Thus clinical interest in subtest scatter has remained high.

It is often asserted that the presence of marked subtest variability reduces the predictive validity of the IQs (Kaufman, 1994). When this hypothesis was tested with the WAIS-III, it was found to be incorrect for predicting memory indices (Ryan, Kreiner, & Burton, 2002). However, Fiorello, Hale, McGrath, Ryan, and Quinn (2002) applied regression commonality analysis to WISC-III factor scores used to predict achievement. They concluded that a general intelligence factor (*g*) did not exist for about 80% of their sample, and they suggested that factor variability justified profile analysis. The scientific status of *g* is too complex to explicate for this chapter, but the results of Fiorello and colleagues run counter to massive amounts of existing research (Jensen, 1998). When multiple regression is applied, there is no universally accepted definition of predictor importance (Azen & Budescu, 2003) and there is no consensus on which method is best employed to explain the relative importance of multiple correlated predictors (Whittaker, Fouladi, & Williams, 2002). More critically, Pedhazur (1997) reported that communality analysis is not "a valid approach for ascertaining relative importance of variables" (p. 275).

To test the predictive validity and diagnostic utility of WISC-III subtest scatter, two samples of students were analyzed: the 1,118 students in the WISC-III/Wechsler Individual Achievement Test (WIAT; Wechsler, 1992) linking sample, and the 1,302 students with LDs from 46 states from Smith and Watkins (2004). For predictive validity, the FSIQ of the WISC-III/WIAT sample was entered as a predictor in multiple regression, followed by the absolute amount of subtest scatter among the 10 mandatory subtests. The FSIQ was substantially predictive of WIAT reading achievement (R = .70), but the addition of subtest scatter did not improve prediction (R = .70). In terms of diagnostic utility, subtest scatter of the WISC-III/WIAT sample was compared to the subtest scatter of the sample with LDs. Figure 12.4 shows that the discrimination was at chance levels (AUC = .51).

The value of WISC-III subtest scatter as a diagnostic indicator was also analyzed by Daley and Nagle (1996) among 308 children with LDs. They found that "subtest scatter and Verbal–Performance discrepancies do not appear to hold any special utility in the diagnosis of learning disabilities" (p. 331). These negative results were replicated in several other studies (Dumont & Willis, 1995; Greenway & Milne, 1999; Iverson, Turner, & Green, 1999). More definitive research was conducted by Watkins (1999), using the WISC-III standardization sample as a normative comparison group. Subtest variability as quantified by range and variance exhibited no diagnostic utility in distinguishing 684 children with LDs from the 2,200 children of the WISC-III standardization sample. Likewise, the number of subtests deviating from examinees' VIQ, PIQ, and FSIQ by ±3 points exhibited no diagnostic utility in distinguishing the children of the WISC-III standardization sample from 684 children with LDs (Watkins & Worrell, 2000). Based upon these results, Watkins and Worrell concluded that "using subtest variability as an indicator of learning disabilities would constitute a case of acting in opposition to scientific evidence" (2000, p. 308).

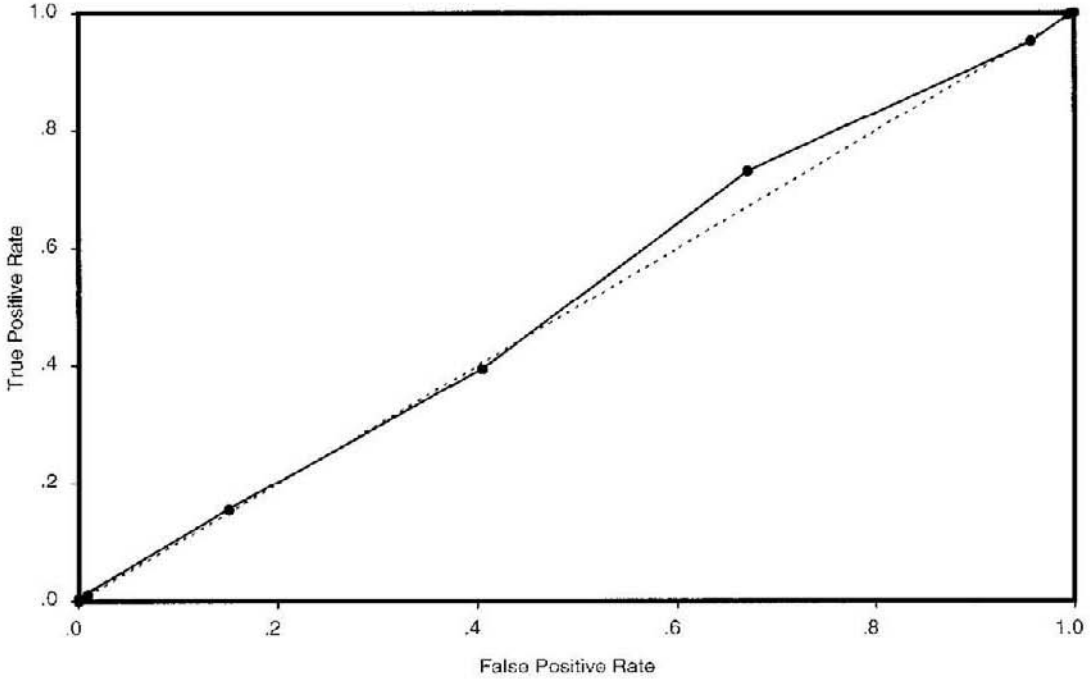

FIGURE 12.4. ROC analysis of 10-subtest scatter between 1,118 students in the WIAT norm sample and 1,302 students with LDs (AUC = .51).

Subtest Profiles

ACID Profile

The ACID subtest profile is probably the most venerable. Based on the Arithmetic, Coding, Information, and Digit Span subtests, it has been applied to the WISC, WISC-R, and WISC-III. Most recently, Prifitera and Dersh (1993) compared percentages of children with WISC-III ACID profiles from samples with LDs and attention-deficit/hyperactivity disorder (ADHD) to percentages showing the ACID profile in the WISC-III standardization sample. Their findings uncovered a greater incidence of ACID profiles in clinical samples, with approximately 5% of the children with LDs and 12% of those with ADHD showing the ACID profile, while such a configuration occurred in only 1% of the cases from the WISC-III standardization sample. Based upon these data, Prifitera and Dersh (1993) concluded that ACID profiles "are useful for diagnostic purposes" because "the presence of a pattern or patterns would suggest strongly that the disorder is present" (pp. 50–51).

Ward, Ward, Hatt, Young, and Mollner (1995) investigated the prevalence of the WISC-III ACID profile among children with LDs (N = 382) and found a prevalence rate of 4.7% (vs. the expected rate of 1%). Obtaining similar ACID results from a sample of children with LDs (N = 165), Daley and Nagle (1996) suggested practitioners "investigate the possibility of a learning disability" (p. 330) when confronted by an ACID profile.

The studies described above relied on the group mean difference approach to incorrectly infer diagnostic utility. Watkins, Kush, and Glutting (1997a) evaluated the diagnostic utility of the WISC-III ACID profile among children with LDs. As in previous research, ACID profiles were more prevalent among children with LDs (N = 612) than among nondisabled children (N = 2,158). However, when ACID profiles were used to classify students into disabled and nondis-

abled groups, they operated with considerable error. At best, only 51% of the children identified by a positive ACID profile were previously diagnosed as having LDs. These data indicated that a randomly selected child with an LD had a more severe ACID profile than a randomly selected child without an LD about 60% of the time (AUC = .60). Although marginally better than chance, the degree of accuracy was quite low (cf. classificatory criteria presented by Swets, 1988).

SCAD Profile

Whereas the ACID profile has a long history, the SCAD profile (based on the Symbol Search, Coding, Arithmetic, and Digit Span subtests) seems to have been first studied by Prifitera and Dersh in 1993. Kaufman (1994) opined that Arithmetic, Coding, and Digit Span have "been quite effective at identifying exceptional groups from normal ones, and . . . are like a land mine that explodes on a diversity of abnormal populations but leaves most normal samples unscathed" (p. 213). Kaufman concluded that the SCAD profile is "an important piece of evidence for diagnosing a possible abnormality" (p. 221); it "won't identify the type of exceptionality, but [the profile is] likely to be valuable for making a presence–absence decision and helping to pinpoint specific areas of deficiency" (p. 214).

Prifitera and Dersh (1993) found the SCAD profile to be more common within a sample of children with LDs (n = 99) and another sample of children with ADHD (n = 65) than within the WISC-III standardization sample. Using this imbalance of prevalence rates as guidance, Prifitera and Dersh suggested that the subtest configuration would be "useful in the diagnosis of LD and ADHD" (p. 53).

These claims were tested by Watkins, Kush, and Glutting (1997b) with children who were enrolled in learning and emotional disability programs (N = 365). When these children were compared to the WISC-III standardization sample via diagnostic utility statistics, an AUC of .59 was generated. Thus, contrary to Kaufman's (1994) assertion, SCAD subtest scores were not found to be important evidence for diagnosing exceptionalities.

Learning Disability Index

The Learning Disability Index (LDI; Lawson & Inglis, 1984) is of particular interest, because it was hypothesized to relate to specific neuropsychological deficits of students with LDs. Lawson and Inglis (1984) conjectured that WISC-R subtests are sensitive to the presence of LDs in direct proportion to their verbal saturation, which is in turn related to left-hemisphere dysfunction. This theoretical link between test and brain functioning is important, because contemporary definitions of LDs specify an endogenous etiology related to "central nervous system dysfunction" (Hammill, 1990, p. 82).

Comparisons of groups of students with and without LDs found significantly higher mean LDI scores among students with LDs than among students in regular education (Bellemare, Inglis, & Lawson, 1986; Clampit & Silver, 1990). Statistically significant LDI differences between groups were subsequently interpreted as evidence that the LDI is diagnostically effective. For example, Kaufman (1990) concluded that the LDI taps a sequential–simultaneous processing dimension and is "quite valuable for distinguishing learning-disabled children from normal children" (p. 354).

However, mean differences between groups are insufficient for individual diagnostic decisions. Consequently, more appropriate diagnostic utility statistics were applied to the WISC-III LDI scores of 2,053 students previously diagnosed with LDs and 2,200 students without LDs. Subsamples of students with specific reading (n = 445) and math (n = 168) disabilities permitted more refined assessment of the efficacy of the LDI. ROC analyses revealed that the LDI resulted in a correct diagnostic decision only 55–64% of the time (Watkins, Kush, & Schaefer, 2002). See Figure 12.3 for an illustrative ROC curve for students with specific reading disabilities (n = 445) compared to the 2,200 students in the normative sample. These results demonstrated that the LDI is an invalid diagnostic indicator of LDs.

Diagnosis: Summary

When properly analyzed, data have consistently failed to find subtest scatter or subtest

profiles to be diagnostically accurate. Based on their review of IQ subtest scatter research, Kline, Snyder, Guilmette, and Castellanos (1992) suggested that psychologists "have pursued scatter analysis . . . with little success. It is time to move on" (p. 11). That suggestion was reiterated by McGrew and Knopik (1996), who remarked that "considering the years of study attributed to the concept of scatter and the lack of an empirical foundation, it is recommended that future research efforts be directed elsewhere" (p. 362). Clearly, there is no scientific support for the use of subtest scatter to inform diagnosis or prediction.

Likewise, results have been consistent in indicating that subtest profiles offer little diagnostic advantage. Limited space prohibits reviews of research regarding other subtest profiles, but diagnostic utility results have been unfailingly negative (Glutting, McDermott, Watkins, Kush, & Konold, 1997; Glutting, McGrath, Kamphaus, & McDermott, 1992; Gussin & Javorsky, 1995; Lowman, Schwanz, & Kamphaus, 1996; Oh, Glutting, & McDermott, 1999; Piedmont, Sokolove, & Fleming, 1989; Smith & Watkins, 2004; Teeter & Korducki, 1998; Watkins, 1996). Sattler (2001) also concluded that subtest profiles "cannot be used for classification purposes or to arrive at a diagnostic label" (p. 299). Similar cautions regarding the use of subtests for diagnostic decisions have been offered by Kaufman and Lichtenberger (2000) and Kamphaus (2001). Unmistakably, abundant scientific evidence and expert consensus recommend against the use of subtest profiles for the diagnosis of childhood learning and behavior disorders.

Specific Cognitive Strengths and Weaknesses

Although there is general agreement that IQ subtest-based diagnosis should be avoided, the use of subtest profiles to identify specific cognitive strengths and weaknesses is frequently recommended. As articulated by Kaufman and Lichtenberger (1998), the examiner "must generate hypotheses about an individual's assets and deficits" (p. 192). Next, the examiner must "confirm or deny these hypotheses by exploring multiple sources of evidence" (p. 192). Finally, "well-validated hypotheses must then be translated into meaningful, practical recommendations" (p. 192) concerning interventions, instructional strategies, and remediation activities (Groth-Marnat, 1997).

Subtest interpretation systems provide hundreds of hypotheses to consider when IQ subtest patterns are obtained (Kaufman, 1994; Sattler, 2001). The enormous number of hypotheses makes it impossible to review the entire scientific literature. However, subtest profiles are often assumed to be related to a wide variety of learning and behavioral variables (Kaufman, 1994; Sattler, 2001).

Academic Achievement

One way to test the utility and validity of subtest scores is to decompose profiles into their elemental components. The unique, incremental predictive validity of each component can then be analyzed separately to determine what aspect or aspects, if any, of the subtest profile are related to academic performance. To this end, Cronbach and Gleser (1953) cautioned that subtest profiles contain only three types of information: *elevation*, *scatter*, and *shape*. Elevation information is represented by a person's aggregate performance (i.e., mean normative score) across subtests. Profile scatter is defined by how widely scores in a profile diverge from its mean. Scatter is typically operationalized by the standard deviation of the subtest scores in a profile. Finally, shape information reflects where "ups and downs" occur in a profile. Even if two profiles have the same elevation and scatter, their high and low points may be different.

Watkins and Glutting (2000) tested the incremental validity of WISC-III subtest profile level, scatter, and shape in forecasting academic performance. WISC-III subtest profiles were decomposed into their three fundamental elements and sequentially regressed onto reading and mathematics achievement scores for nonexceptional (n = 1,118) and exceptional (n = 538) children. Profile elevation was statistically and practically significant for both nonexceptional (R = .72 to .75) and exceptional (R = .36 to .61) children. Profile scatter did not aid in the prediction of achievement. Profile shape accounted for an

additional 5–8% of the variance in achievement measures: One pattern of relatively high verbal scores positively predicted both reading and mathematics achievement, and a pattern of relatively low scores on the WISC-III Arithmetic subtest was negatively related to mathematics. Beyond these two somewhat intuitive patterns, profile shape information had inconsequential incremental validity for both nonexceptional and exceptional children. In other words, it was the averaged, norm-referenced information (i.e., elevation) contained in subtest profiles that best predicted achievement. This information is essentially redundant to that conveyed by global intelligence scores.

Similar results have been obtained by other researchers with a variety of intelligence tests (e.g., the WISC-R, Stanford–Binet, Woodcock–Johnson, etc.) (Glutting, Youngstrom, Ward, Ward, & Hale, 1997; Hale & Raymond, 1981; Hale & Saxe, 1983; Kahana, Youngstrom, & Glutting, 2002; Kline et al., 1992; McDermott & Glutting, 1997; McGrew & Knopik, 1996; Ree & Carretta, 1997; Youngstrom, Kogos, & Glutting, 1999). Interestingly, research has consistently demonstrated that general mental ability also accounts for the vast majority of the predictable variance in job learning among adults (Ree & Carretta, 2002; Schmidt, 2002). Given these results, Kline and colleagues (1992) concluded that "the most useful information from IQ-type tests is the overall elevation of the child's profile. Profile shape information adds relatively little unique information, and therefore examiners should not overinterpret particular patterns of scores" (p. 431). Thorndike (1986) similarly concluded that 80–90% of the predictable variance in scholastic performance is accounted for by general ability, with only 10–20% accounted for by all other scores in IQ tests. From these findings, it has been concluded that subtest scatter and shape offer minimal assistance for generating hypotheses about children's academic performance.

Learning Behaviors

Teacher ratings of child learning behaviors, as operationalized by the Learning Behaviors Scale (LBS; McDermott, Green, Francis, & Stott, 1996), reflect four relatively independent subareas: competence motivation, attitude toward learning, attention/persistence, and strategy/flexibility. The DAS and LBS were conormed with a nationally representative sample of 1,250 children. When DAS and LBS scores were compared, DAS global ability accounted for 8.2% of learning behavior, and DAS subtests only increased the explained variance by 1.7%.

Test Session Behaviors

It is widely assumed that astute test session observation and clinical insight allow psychologists to draw valid inferences regarding an examinee's propensities and behaviors outside the testing situation (Sparrow & Davis, 2000). That is, a quiet child during testing is assumed to be retiring in other situations, an active child is inferred to be energetic in the classroom, and so on. However, a synthesis of research on test session behaviors found that the average correlation between test session behaviors and conduct in other environments was only .18 (Glutting, Youngstrom, Oakland, & Watkins, 1996).

Among the normative sample of the Guide to the Assessment of Test Session Behavior (Glutting & Oakland, 1993), a standardized test session observation scale, there was little differential variability across WISC-III subtests (Oakland, Broom, & Glutting, 2000). In fact, global ability accounted for 9.2% of test session behaviors, and the addition of WISC-III subtests only explained another 3.2% (McDermott & Glutting, 1997).

Behavioral Adjustment

Subtest profiles are commonly assumed to reflect dispositions that allow inferences about school behavior and adjustment. To test this assumption, a nationally representative sample of 1,200 children was administered the DAS, and their teachers independently provided standardized ratings of school and classroom behaviors on the Adjustment Scales for Children and Adolescents (ASCA; McDermott, Marston, & Stott, 1993). The ASCA provides measures of six core syndromes: attention-deficit hyperactivity, solitary aggressive (provocative), solitary

aggressive (impulsive), oppositional defiant, diffident, and avoidant. Following the method of Kaufman (1994), the 5% of children with the most unusual DAS subtest profiles were identified and then matched on the basis of age, race, gender, parent education levels, and overall IQs to an equal number of comparison group children without unusual subtest profiles. There were no significant differences between these two groups on the six ASCA behavioral scales. Nor were there significant differences on academic tests. Thus academic and behavioral problems were not related to unusual DAS subtest profiles (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998).

This negative result was replicated by an analysis of the relationships between IQs from the WISC-III and classroom behaviors measured by the ASCA (Glutting et al., 1996). Four of the 56 correlations between the WISC-III and ASCA reached statistical significance at $p < .05$, but this ratio barely exceeded the expected chance rate of three significant correlations. The average coefficient was also low (average $r = -.04$, with a 95% confidence interval of −.20 to +.13) and indicated that as much as 99% of the score variation was unique to each instrument. These meager results were not surprising: "11 previous investigations showed an average relationship of −.19 between children's scores on individually administered IQ tests and their home and school behavior" (Glutting et al., 1996, p. 103).

Specific Strengths and Weaknesses: Summary

Most hypotheses generated from subtest variation "are either untested by science or unsupported by scientific findings" (Kamphaus, 1998, p. 46). The evidence that exists regarding relationships between subtest profiles and socially important academic and behavioral outcomes is, at best, weak: Subtest profile information contributes 2–8% variance beyond general ability to the prediction of achievement, and 2–3% to the prediction of learning behaviors and test session behaviors. Hypothesized relationships between subtest profiles and measures of psychosocial behavior have persistently failed to achieve statistical or clinical significance. Thus "neither subtest patterns nor profiles of IQ have been systematically found to be related to personality variables" (Zeidner & Matthews, 2000, p. 585). After reviewing the research on subtest analysis, Hale and Green (1995) concluded that "knowledge of a child's subtest profile does not appreciably help the clinician in predicting either academic achievement levels or behavioral difficulties" (p. 98). Thus hypotheses derived from subtest analysis "are based on the clinician's acumen and not on any sound research base" (Kamphaus, 2001, p. 598).

GENERAL CONCLUSIONS

Many researchers have found the popularity of IQ subtest profile analysis to greatly outstrip its meager scientific support (Braden, 1997; Bray, Kehle, & Hintze, 1998; Gresham & Witt, 1997; Kamphaus, 2001; Reynolds & Kamphaus, 2003; Watkins, 2000, 2003). Although subtest profile analysis has not demonstrated adequate reliability, diagnostic utility, or treatment validity, it continues to be endorsed by assessment specialists and applied widely in training and practice.

Apparently, subtest profile interpretation flourishes due to its intuitive appeal (Bracken, McCallum, & Crain, 1993) and clinical tradition (Shaw, Swerdlik, & Laurent, 1993). Subtest-based interpretation systems are often justified by prescientific arguments without consideration for the professional obligations of psychologists (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), the limitations of clinical judgment (Garb, 2003), or the demands of scientific reasoning (Gibbs, 2003). Scientific psychological practice cannot be sustained by clinical conjectures, personal anecdotes, and unverifiable beliefs that have consistently failed empirical validation. Given the paucity of empirical support, subtest profile analysis can best be described as reliance on clinical delusions, illusions, myths, or folklore. Consequently, psychologists should eschew interpretation of cognitive subtest profiles.

REFERENCES

Alfonso, V. C., Oakland, T. D., LaRocca, R., & Spanakos, A. (2000). The course on individual cognitive assessment. *School Psychology Review, 29,* 52–64.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (1999). *Standards for educational and psychological testing.* Washington, DC: American Educational Research Association.

Azen, R., & Budescu, D. V. (2003). The dominance analysis approach for comparing predictors in multiple regression. *Psychological Methods, 8,* 129–148.

Belk, M. S., LoBello, S. G., Ray, G. E., & Zachar, P. (2002). WISC-III administration, clerical, and scoring errors made by student examiners. *Journal of Psychoeducational Assessment, 20,* 290–300.

Bellemare, F. G., Inglis, J., & Lawson, J. S. (1986). Learning disability indices derived from a principal components analysis of the WISC-R: A study of learning disabled and normal boys. *Canadian Journal of Behavioral Science, 18,* 86–91.

Bracken, B. A., McCallum, R. S., & Crain, R. M. (1993). WISC-III subtest composite reliabilities and specificities: Interpretive aids. *Journal of Psychoeducational Assessment Monograph Series, Advances in Psychological Assessment: Wechsler Intelligence Scale for Children—Third Edition,* pp. 22–34.

Braden, J. P. (1997). The practical impact of intellectual assessment issues. *School Psychology Review, 26,* 242–248.

Bray, M. A., Kehle, T. J., & Hintze, J. M. (1998). Profile analysis with the Wechsler scales: 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.

Cattell, R. B. (1944). Psychological measurement: Normative, ipsative, interactive. *Psychological Review, 51,* 292–303.

Clampit, M. K., & Silver, S. J. (1990). Demographic characteristics and mean profiles of learning disability index subsets of the standardization sample of the Wechsler Intelligence Scale for Children—Revised. *Journal of Learning Disabilities, 23,* 263–264.

Cronbach, L. J., & Gleser, G. C. (1953). Assessing similarity between profiles. *Psychological Bulletin, 50,* 456–473.

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

Dawes, R. M. (1993). Prediction of the future versus an understanding of the past: A basic asymmetry. *American Journal of Psychology, 106,* 1–24.

Dawson-Saunders, B., & Trapp, R. G. (1990). *Basic and clinical biostatistics.* Norwalk, CT: Appleton & Lange.

Dumont, R., & Willis, J. O. (1995). Intrasubtest scatter on the WISC-III for various clinical samples vs. the standardization sample: An examination of WISC folklore. *Journal of Psychoeducational Assessment, 13,* 271–285.

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook.* San Antonio, TX: Psychological Corporation.

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

Faust, D. (1990). Data integration in legal evaluations: Can clinicians deliver on their premises? *Behavioral Sciences and the Law, 7,* 469–483.

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

Fiorello, C. A., Hale, J. B., McGrath, M., Ryan, K., & Quinn, S. (2002). IQ interpretation for children with flat and variable test profiles. *Learning and Individual Differences, 13,* 115–125.

Frank, G. (1983). *The Wechsler enterprise: An assessment of the development, structure, and use of the Wechsler tests of intelligence.* New York: Pergamon Press.

Gambrill, E. (1990). *Critical thinking in clinical practice: Improving the accuracy of judgments and decisions about clients.* San Francisco: Jossey-Bass.

Garb, H. N. (2003). Clinical judgment and mechanical prediction. In I. B. Weiner (Series Ed.) & J. R. Graham & J. A. Naglieri (Vol. Eds.), *Handbook of psychology: Vol. 10. Assessment psychology* (pp. 27–42). New York: Wiley.

Gibbs, L. E. (2003). *Evidence-based practice for the helping professions: A practical guide with integrated multimedia.* Pacific Grove, CA: Brooks/Cole.

Glutting, J. J., McDermott, P. A., Konold, T. R., Snelbaker, A. J., & Watkins, M. W. (1998). More ups and downs of subtest analysis: Criterion validity of the DAS with an unselected cohort. *School Psychology Review, 27,* 599–612.

Glutting, J. J., McDermott, P. A., Watkins, M. W., Kush, J. C., & Konold, T. R. (1997). The base rate problem and its consequences for interpreting children's ability profiles. *School Psychology Review, 26,* 176–188.

Glutting, J. J., McGrath, E. A., Kamphaus, R. W., & McDermott, P. A. (1992). Taxonomy and validity of subtest profiles on the Kaufman Assessment Battery for Children. *Journal of Special Education, 26,* 85–115.

Glutting, J. J., & Oakland, T. (1993). *Guide to the assessment of test-session behavior.* San Antonio, TX: Psychological Corporation.

Glutting, J. J., Watkins, M. W., & Youngstrom, E. A. (2003). Multifactored and cross-battery ability assessments: Are they worth the effort? In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment of children: Intelligence, aptitude, and achievement* (2nd ed., pp. 343–374). New York: Guilford Press.

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

Glutting, J. J., Youngstrom, E. A., Ward, T., Ward, S., & Hale, R. L. (1997). Incremental efficacy of WISC-III factor scores in predicting achievement: What do they tell us? *Psychological Assessment, 9,* 295–301.

Greenway, P., & Milne, L. (1999). Relationship between psychopathology, learning disabilities, or both and WISC-III subtest scatter in adolescents. *Psychology in the Schools, 36,* 103–108.

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future directions. *School Psychology Quarterly, 12,* 249–267.

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

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.

Hale, R. L., & Green, E. A. (1995). Intellectual evaluation. In L. A. Heiden & M. Hersen (Eds.), *Introduction to clinical psychology* (pp. 79–100). New York: Plenum Press.

Hale, R. L., & Raymond, M. R. (1981). Wechsler Intelligence Scale for Children—Revised patterns of strengths and weaknesses as predictors of the intelligence achievement relationship. *Diagnostique, 7,* 35–42.

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

Hammill, D. D. (1990). On defining learning disabilities: An emerging consensus. *Journal of Learning Disabilities, 23,* 74–84.

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

Henderson, A. R. (1993). Assessing test accuracy and its clinical consequences: A primer for receiver operating characteristic curve analysis. *Annals of Clinical Biochemistry, 30,* 521–539.

Hicks, L. E. (1970). Some properties of ipsative, normative, and forced-choice normative measures. *Psychological Bulletin, 74,* 167–184.

Iverson, G. L., Turner, R. A., & Green, P. (1999). Predictive validity of WAIS-R VIQ–PIQ splits in persons with major depression. *Journal of Clinical Psychology, 55,* 519–524.

Jensen, A. R. (1998). *The g factor: The science of mental ability.* Westport, CT: Praeger.

Jensen, A. R. (2002). SASP interviews: Arthur R. Jensen. *SASP News, 2*(4), 8–19.

Kahana, S. Y., Youngstrom, E. A., & Glutting, J. J. (2002). Factor and subtest discrepancies on the Differential Ability Scales: Examining prevalence and validity in predicting academic achievement. *Assessment, 9,* 82–93.

Kamphaus, R. W. (1998). Intelligence test interpretation: Acting in the absence of evidence. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 39–57). San Diego, CA: Academic Press.

Kamphaus, R. W. (2001). *Clinical assessment of child and adolescent intelligence* (2nd ed.). Boston: Allyn & Bacon.

Kamphaus, R. W., Petoskey, M. D., & Rowe, E. W. (2000). Current trends in psychological testing of children. *Professional Psychology: Research and Practice, 31,* 155–164.

Kaufman, A. S. (1990). *Assessing adolescent and adult intelligence.* Boston: Allyn & Bacon.

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

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

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

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.

Kline, R. B., Snyder, J., Guilmette, S., & Castellanos, M. (1992). Relative usefulness of elevation, variability, and shape information from WISC-R, K-ABC, and Fourth Edition Stanford–Binet profiles in predicting achievement. *Psychological Assessment, 4,* 426–432.

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.

Lawson, J. S., & Inglis, J. (1984). The psychometric assessment of children with learning disabilities: An index derived from a principal components analysis of the WISC-R. *Journal of Learning Disabilities, 17,* 517–522.

Livingston, R. B., Jennings, E., Reynolds, C. R., & Gray, R. M. (2003). Multivariate analyses of the profile stability of intelligence tests: High for IQs, low to very low for subtest analyses. *Archives of Clinical Neuropsychology, 18,* 487–507.

Lowman, M. G., Schwanz, K. A., & Kamphaus, R. W. (1996). WISC-III third factor: Critical measurement issues. *Canadian Journal of School Psychology, 12,* 15–22.

Maller, S. J., & McDermott, P. A. (1997). WAIS-R profile analysis for college students with learning disabilities. *School Psychology Review, 26,* 575–585.

Matarazzo, J. D. (1985). Psychological assessment of intelligence. In H. I. Kaplan & B. J. Sadock (Eds.),

Comprehensive textbook of psychiatry (Vol. 1, 4th ed., pp. 502–513). Baltimore: Williams & Wilkins.

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

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, R. A. (1992). Illusion of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504–526.

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

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

McDermott, P. A., Green, L. F., Francis, J. M., & Stott, D. H. (1996). *Learning Behaviors Scale.* Philadelphia: Edumetric & Clinical Science.

McDermott, P. A., Marston, N. C., & Stott, D. H. (1993). *Adjustment Scales for Children and Adolescents.* Philadelphia: Edumetric & Clinical Science.

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.

McGrew, K. S., & Knopik, S. N. (1996). The relationship between intra-cognitive scatter on the Woodcock–Johnson Psycho-Educational Battery—Revised and school achievement. *Journal of School Psychology, 34,* 351–364.

McNemar, Q. (1964). Lost: Our intelligence? Why? *American Psychologist, 19,* 871–882.

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

Moffitt, T. E., & Silva, P. A. (1987). WISC-R Verbal and Performance IQ discrepancy in an unselected cohort: Clinical significance and longitudinal stability. *Journal of Consulting and Clinical Psychology, 55,* 768–774.

Mueller, H. H., Dennis, S. S., & Short, R. H. (1986). A meta-exploration of WISC-R factor score profiles as a function of diagnosis and intellectual level. *Canadian Journal of School Psychology, 2,* 21–43.

Neyens, L. G. J., & Aldenkamp, A. P. (1996). Stability of cognitive measures in children of average ability. *Child Neuropsychology, 2,* 161–170.

Oakland, T., Broom, J., & Glutting, J. (2000). Use of freedom from distractibility and processing speed to assess children's test-taking behaviors. *Journal of School Psychology, 38,* 469–475.

Oakland, T., & Hu, S. (1992). The top 10 tests used with children and youth worldwide. *Bulletin of the International Test Commission, 19,* 99–120.

Oh, H.-J., Glutting, J. J., & McDermott, P. A. (1999). An epidemological-cohort study of DAS processing speed factor: How well does it identify concurrent achievement and behavior problems? *Journal of Psychoeducational Assessment, 17,* 362–375.

Pedhazur, E. J. (1997). *Multiple regression in behavioral research: Explanation and prediction* (3rd ed.). Fort Worth, TX: Harcourt Brace.

Pfeiffer, S. I., Reddy, L. A., Kletzel, J. E., Schmelzer, E. R., & Boyer, L. M. (2000). The practitioner's view of IQ testing and profile analysis. *School Psychology Quarterly, 15,* 376–385.

Piedmont, R. L., Sokolove, R. L., & Fleming, M. Z. (1989). An examination of some diagnostic strategies involving the Wechsler intelligence scales. *Psychological Assessment, 1,* 181–185.

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,* pp. 43–55.

Ree, M. J., & Carretta, T. R. (1997). What makes an aptitude test valid? In R. F. Dillon (Ed.), *Handbook on testing* (pp. 65–81). Westport, CT: Greenwood Press.

Ree, M. J., & Carretta, T. R. (2002). g2K. *Human Performance, 15,* 3–23.

Reynolds, C. R., & Kamphaus, R. W. (2003). *Reynolds Intellectual Assessment Scales and the Reynolds Intellectual Screening Test professional manual.* Lutz, FL: Psychological Assessment Resources.

Rispens, J., Swaab, H., van den Oord, E. J. C. G., Cohen-Kettenis, P., van Engeland, H., & van Yperen, T. (1997). WISC profiles in child psychiatric diagnosis: Sense or nonsense? *Journal of the American Academy of Child and Adolescent Psychiatry, 36,* 1587–1594.

Roid, G. H. (2003). *Stanford–Binet Intelligence Scales, Fifth Edition: Technical manual.* Itasca, IL: Riverside.

Ryan, J. J., & Bohac, D. L. (1994). Neurodiagnostic implications of unique profiles of the Wechsler Adult Intelligence Scale—Revised. *Psychological Assessment, 6,* 360–363.

Ryan, J. J., Kreiner, D. S., & Burton, D. B. (2002). Does high scatter affect the predictive validity of WAIS-III IQs? *Applied Neuropsychology, 9,* 173–178.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (7th ed.). Boston: Houghton Mifflin.

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

Schmidt, F. L. (2002). The role of general cognitive ability and job performance: Why there cannot be a debate. *Human Performance, 15,* 187–210.

Schmidt, F. L., Le, H., & Ilies, R. (2003). Beyond alpha: An empirical examination of the effects of different sources of measurement error on reliability estimates for measures of individual differences constructs. *Psychological Methods, 8,* 206–224.

Shaw, S. R., Swerdlik, M. E., & Laurent, J. (1993). Review of the WISC-III. In B. Bracken & R. S. McCallum (Eds.), *Advances in psychoeducational assessment* (pp. 151–160). Brandon, VT: Clinical Psychology.

Simensen, R. J., & Sutherland, J. (1974). Psychological assessment of brain damage: The Wechsler scales. *Academic Therapy, 10,* 69–81.

Smith, C. B., & Watkins, M. W. (2004). Diagnostic utility of the Bannatyne WISC-III pattern. *Learning Disabilities Research and Practice, 19,* 49–56.

Smith, T., Smith, B. L., Bramlett, R. K., & Hicks, N. (1999, April). *WISC-III stability over a three year period.* Paper presented at the annual meeting of the National Association of School Psychologists, Las Vegas, NV.

Sparrow, S. S., & Davis, S. M. (2000). Recent advances in the assessment of intelligence and cognition. *Journal of Child Psychology and Psychiatry, 41,* 117–131.

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

Swets, J. A. (1996). *Signal detection theory and RIC analysis in psychology and diagnosis: Collected papers.* Mahwah, NJ: Erlbaum.

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

Teeter, P. A., & Korducki, R. (1998). Assessment of emotionally disturbed children with the WISC-III. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 119–138). San Diego, CA: Academic Press.

Thorndike, R. L. (1986). The role of general ability in prediction. *Journal of Vocational Behavior, 29,* 332–339.

Thorndike, R. M. (1997). *Measurement and evaluation in psychology and education* (6th ed.). Upper Saddle River, NJ: Merrill.

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.

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

Watkins, M. W. (1996). Diagnostic utility of the WISC-III developmental index as a predictor of learning disabilities. *Journal of Learning Disabilities, 29,* 305–312.

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. (2000). Cognitive profile analysis: A shared professional myth. *School Psychology Quarterly, 15,* 465–479.

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

Watkins, M. W., & Canivez, G. L. (2004). Temporal stability of WISC-III subtest composite: Strengths and weaknesses. *Psychological Assessment, 16,* 133–138.

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

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review, 23,* 640–651.

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

Watkins, M. W., Kush, J. C., & Glutting, J. J. (1997b). 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., & Schaefer, B. A. (2002). Diagnostic utility of the Learning Disability Index. *Journal of Learning Disabilities, 35,* 98–103.

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. (1949). *Wechsler Intelligence Scale for Children.* New York: Psychological Corporation.

Wechsler, D. (1958). *The measurement and appraisal of adult intelligence* (4th ed.). Baltimore: Williams & Wilkins.

Wechsler, D. (1974). *Wechsler Intelligence Scale for Children—Revised.* New York: Psychological Corporation.

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

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

Wechsler, D. (1997). *Wechsler Adult Intelligence Scale—Third Edition.* San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children—Fourth Edition: Technical and interpretive manual.* San Antonio, TX: Psychological Corporation.

Whittaker, T. A., Fouladi, R. T., & Williams, N. J. (2002). Determining predictor importance in multiple regression under varied correlational and distributional conditions. *Journal of Modern Applied Statistical Methods, 1,* 354–366.

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

Youngstrom, E. A., Kogos, J. L., & Glutting, J. J. (1999). Incremental efficacy of Differential Ability Scales factor scores in predicting individual achievement criteria. *School Psychology Quarterly, 14,* 26–39.

Zachary, R. A. (1990). Wechsler's intelligence scales: Theoretical and practical considerations. *Journal of Psychoeducational Assessment, 8,* 276–289.

Zarin, D. A., & Earls, F. (1993). Diagnostic decision making in psychiatry. *American Journal of Psychiatry, 150,* 197–206.

Zeidner, M. (2001). Invited foreword and introduction. In J. J. W. Andrews, D. H. Saklofske, & H. L. Janzen (Eds.), *Handbook of psychoeducational assessment: Ability, achievement, and behavior in children.* San Diego, CA: Academic Press.

Zeidner, M., & Matthews, G. (2000). Intelligence and personality. In R. J. Sternberg (Ed.), *Handbook of intelligence* (pp. 581–610). New York: Cambridge University 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.