IQ Subtest Analysis
Clinical Acumen or Clinical Illusion?

Marley W. Watkins
Pennsylvania State University

Subtest analysis is pervasive in psychological training and practice. That is, the speculation that the variability or profile of an individual's scaled scores across the subtests of an intelligence test have meaning beyond that provided by global IQ measures. A review of subtest analysis research revealed that neither subtest scatter nor subtest profiles demonstrate acceptable accuracy in discriminating among diagnostic groups. The evidence that exists regarding relations between subtest profiles and socially important academic and psychosocial outcomes is, at best, weak: subtest profile information contributes 2% to 8% variance beyond general ability to the prediction of achievement and 2% to 3% to the prediction of learning behaviors and test-session behaviors. Hypothesized relationships between subtest profiles and other psychosocial behaviors persistently fail to achieve statistical or clinical significance. Methodological problems in research and practice that cause subtest analysis results to be more illusory than real and to represent more of a shared professional myth than clinically astute detective work are explicated.

Weiner (1989) encouraged psychologists to "(a) know what their tests can do and (b) act accordingly" (p. 829). This admonition is in accord with ethical codes (APA, 1992) and congruent with professional testing standards (AERA, APA, & NCME, 1999). It is especially important to know what IQ tests can do because intellectual assessment is a common responsibility of psychologists (Sparrow & Davis, 2000) and involves high-stakes decisions for examinees (Gresham & Witt, 1997).

There is considerable evidence supporting interpretation of global IQ indices (Jensen, 1998; Kubiszyn et al., 2000; Neisser et al., 1996). Generalization of this well-founded practice to interpretation of individual subtest patterns or profiles naturally evolved (Kehle, Clark, & Jenson, 1993). Psychologists speculated that the variability (scatter) or profile (shape) of an individual's scaled scores across the subtests of an intelligence battery might be a sign of neurological dysfunction (Drebing, Satz, Van Gorp, Chervinsky, & Uchiyama, 1994), learning disability (McLean, Reynolds, & Kaufman, 1990), or emotional disability (Drummond, 2000). Even if not used for diagnosis, subtests might identify specific cognitive strengths and weaknesses (Zeidner, 2001). Following this logic, a high degree of subtest variability or specific patterns of subtest scores were presumed to substantially invalidate global intelligence indices (Groth-Marnat, 1997) so that subtests, rather than IQ composites, became the focus of interpretation. Psychologists believed that such a multidimensional view of intelligence would provide greater insight into the nature of human ability than summary intellectual indices (Zimmerman & Woo-Sam, 1985).

Based on these principles, intricate subtest profile interpretation systems have achieved wide popularity in psychological training and practice (Aiken, 1996; Groth-Marnat, 1997; Kaufman, 1994a; Sattler, 2001). 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, and almost all use texts that advocate subtest analysis (Alfonso, Oakland, LaRocca, & Spanakos, 2000). As would be expected from such training, school psychologists frequently analyze cognitive subtest profiles in their practice (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Among their sample of clinicians, for example, Pfeiffer et al. found that almost 70% reported factor

Marley W. Watkins, Department of Educational and School Psychology, Pennsylvania State University.

The author is indebted to Gary Canivez, Joe Glutting, Joe Kush, Paul McDermott, and Eric Youngstrom for their generous professional support.

Correspondence concerning this article should be addressed to Marley W. Watkins, Department of Educational and School Psychology, Pennsylvania State University, University Park, PA 16802. E-mail: mww10@psu.edu.

scores/profile analysis to be a useful feature of intelligence tests and 29% reported that they derived specific value from individual subtests. This review will consider the scientific support for the widespread psychological practice of IQ subtest analysis.

A HISTORICAL REVIEW OF IQ SUBTEST ANALYSIS

Subtest Scatter

Attempts to analyze IQ subtest variations date back more than 70 years (Zachary, 1990). 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 greater potential than indicated by averaged IQ composites.

Diagnostic Accuracy

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 and opined that "we regard scatter analysis as inefficient and inappropriate" (p. 45).

The quantitative combination of results from 94 studies (N=9,372) also demonstrated that subtest scatter and scatter between Verbal IQ (VIQ) and Performance IQ (PIQ) failed to uniquely distinguish children with learning disabilities (Kavale & Forness, 1984). For example, the average VIQ-PIQ difference for children with learning disabilities 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" (p. 139).

An invaluable source of data on IQ subtest analysis is the Dunedin Multidisciplinary Health and Development Study (DMHDS; Silva, 1990). The DMHDS included an epidemiological sample of more than 1,000 New Zealand children assessed with a battery of psychological, sociological, and medical measures every two years from birth to adulthood. Its representative sample, comprehensive assessment battery, and longitudinal design make DMHDS IQ test results unique in the professional literature.

Using data from the DMHDS, Moffitt and Silva (1987) reported on the clinical significance and stability of Wechsler Intelligence Scale for Children–Revised (WISC–R; Wechsler, 1974) VIQ and PIQ scatter. They concluded that perinatal, neurological, and health problems did not cause extreme VIQ-PIQ discrepancies (i.e., > 90th percentile) and found that neither behavior problems nor motor problems were significantly related to VIQ-PIQ scores. Further, 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 two years later. Thus, "VIQ-PIQ discrepancies are of doubtful diagnostic value" (Moffitt & Silva, 1987, p. 773).

Although not supported by DMHDS data, many psychologists associate VIQ-PIQ differences on Wechsler scales with brain damage. For example, Kaufman (1994b) reported extensive evidence on a VIQ>PIQ difference pattern found among patients with right cerebral hemisphere damage and indicated that these score differences "suggest strong discriminant validity and instructional value of the V-P discrepancy for neuropsychological assessment purposes" (p. 201). However, Macmann and Barnett (1994) pointed out that more than half of the reported VIQ>PIQ differences were not statistically significant and would result in many errors if used to make diagnostic decisions.

Neither was analysis of the subtest scatter on the standardization sample of the Wechsler Adult Intelligence Scale–Revised (WAIS–R; Wechsler, 1981) supportive of scatter's unique diagnostic value. Matarazzo and Prifitera (1989) noted that while some supportive WAIS–R subtest scatter research had been published, lack of cross-validation and replication made interpretation of subtest scatter "art and not science" (p. 186).

Finally, a narrative review of 70 years of research on subtest scatter also arrived at pessimistic conclusions concerning its diagnostic utility (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 observed that "extensive scatter proved to be both typical and 'normal,' and thus of limited use as a diagnostic feature" (p. 878).

Academic Achievement

Beyond diagnostic accuracy, subtest scatter was also found to be unrelated to academic achievement. After entering WISC–R subtest level (i.e., general ability) in a regression model, Hale and Saxe (1983) found that subtest scatter did not contribute to the prediction of academic achievement. These results were replicated by Kline, Snyder, Guilmette, and Castellanos (1992) who reported that scatter had no incremental validity beyond general ability in predicting achievement. Further, scatter was consistently found to be ineffectual in developing educational intervention strategies (Kramer et al., 1987).

Subtest Profiles

Diagnostic Accuracy

Given the popularity of the Wechsler scales (Sparrow & Davis, 2000), Wechsler subtest profiles have been the source of much research. For example, the validity of using Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and WISC–R subtests for diagnosing learning disabilities 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" (p. 150).

A different approach to meta-analysis was applied by Mueller, Dennis, and Short (1986). They statistically clustered the WISC–R subtest data of 119 samples of normal 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 et al. concluded that Wechsler subtest profiles were not helpful in differentiating among children with emotional and learning impairments and 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 et al. (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 et al. found that subtest patterns did not significantly differ across the various groups and concluded that "WISC profiles . . . cannot contribute to differential diagnosis" (p. 1593). Further, after controlling for general intelligence, subtests exhibited little incremental validity when predicting parent ratings of child psychopathology.

Academic Achievement

It is widely recognized that IQ scores covary positively with academic achievement (Neisser et al., 1996). Hale and Saxe (1983) hypothesized that if subtest profiles are useful in distinguishing between children with and without learning problems, then profiles should account for variance in academic achievement beyond that contributed by general ability. They tested this hypothesis and found that after controlling for general ability, WISC–R subtest profile information accounted for 8% of the variance in concurrent reading performance. A similar study by Hale and Raymond (1981) also found that general ability was responsible for the preponderance of variability in achievement. Kline et al. (1992) extended these results by applying similar methods to several IQ tests. General ability explained 29% to 43% of achievement test variance and subtest profiles explained another 7% to 11% variance. Kline et al. 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% to 90% of the predictable variance in scholastic performance is accounted for by general ability, with only 10% to 20% accounted for by all other scores in IQ tests.

Empirical Subtest Profiles

Most research on the diagnostic accuracy and predictive validity of IQ subtests has relied on clinically derived subtest profiles. Alternatively, multivariate statistical methods can be used to form empirical subtest groupings. If such empirically generated profiles exhibit adequate psychometric properties, then they might be useful for diagnosis and prediction of socially important criteria.

The temporal stability of empirical subtest profiles was tested by Moffitt, Caspi, Harkness, and Silva (1993), who used WISC–R data from the exemplary DMHDS project to examine the degree of precision that resulted when children were categorized into IQ subtest profiles at one age and then again two years later. Subtest profiles were cocategorized with low agreement (i.e., only around 15% better than chance) when general level of intelligence was not considered. From these results, Moffitt et al. concluded that "the important and replicable fact that can be gleaned from the repeated assessment of IQ is the elevation, or mean height, of the scores in the profile" and suggested that "the pattern of IQ profiles over time is merely error" (p. 475).

Subtest profiles formed after controlling for general ability have also been found to be unstable among smaller and less representative samples. For example, subtest strengths and weaknesses found on initial testing disappeared more than 60% of the time within one month with the WISC–R standardization retest sample and more than 80% of the time within three years with a large sample of students enrolled in special education classes (McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Consequently, McDermott et al. attributed the temporal stability of empirically formed subtest profiles primarily to general ability.

The importance of general ability to the initial formation of subtest groupings has also been observed. For example, the WISC–R standardization sample was shown to contain seven core profiles distinguished primarily by level of general ability (McDermott, Glutting, Jones, Watkins, & Kush, 1989). Although four core profiles of the Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967) were differentiated by general ability level, two other core profiles were also marked by VIQ-PIQ differences. Similar results were reported when WAIS–R (Wechsler, 1981) standardization sample subtest scores were statistically grouped (McDermott, Glutting, Jones, & Noonan, 1989).

Empirically generated subtest profiles can serve as a normative standard against which subtest profiles obtained from clinical groups are tested. 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, et al., 1989) to 1,222 students with learning disabilities, 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 the standardization sample.

Another normative comparison applied the 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 learning disabilities (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 learning disabilities.

Summary

Subtest scatter had been determined to be clinically ineffectual as early as 1937 (Harris & Shakow, 1937). 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 demonstrated that neither subtest scatter nor subtest profiles demonstrated acceptable accuracy in discriminating among diagnostic groups. Likewise, subtest scatter and profiles demonstrated little incremental validity over general ability in predicting academic achievement. Additionally, subtest profiles of clinical groups, when normatively compared with profiles statistically derived from the standardization samples of major IQ batteries, were usually typical in relation to the general population, rendering it implausible that most subtest profiles were clinically unique. However, even these empirically generated subtest profiles failed to exhibit adequate temporal stability unless they included general ability; that is, profiles based on subtest shape alone were unstable. General ability, rather than subtest profiles, was also the most powerful predictor of academic achievement. Thus, a selective historical review of subtest analysis reiterated the pervasive role of general ability (Jensen, 1998; Neisser et al., 1996) but provided no scientific support for the use of IQ subtest analysis in differential diagnosis or the prediction of academic achievement.

CURRENT PERSPECTIVES ON IQ SUBTEST ANALYSIS

Although subtest analysis with older IQ tests has not proven fruitful, modern tests and methods might alter these conclusions. Complex, unpredictable changes in subtest relationships may obtain when tests are revised (Strauss, Spreen, & Hunter, 2000). Additionally, current IQ measures contain more subtests, involve more representative standardization samples, and tend to be more theory based than their predecessors (Sattler, 2001).

Subtest Scatter

Tables of subtest scatter are included in both Wechsler Intelligence Scale for Children–Third Edition (WISC–III; Wechsler, 1991) 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 scatter tables. Thus, clinical interest in subtest scatter remains high.

Diagnostic Accuracy

The value of WISC–III subtest scatter as a diagnostic indicator was analyzed by Daley and Nagle (1996) among 308 children with learning disabilities. 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). Likewise, WISC–III subtest scatter did not significantly discriminate between 45 children with learning disabilities and 34 children without learning disabilities (Mayes, Calhoun, & Crowell, 1998). Similarly, Dumont and Willis (1995) found no evidence to support the diagnostic value of scatter within subtests. WISC–III subtest scatter was also ineffectual as an indicator of learning disabilities among 170 Australian adolescents (Greenway & Milne, 1999). Although a relationship was found between scatter and emotional disturbance, it was inconsistent and had little diagnostic utility.

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 learning disabilities from the 2,200 children of the WISC–III standardization sample. Likewise, the number of subtests deviating from examinees' VIQ, PIQ, and Full Scale IQ (FSIQ) by ±3 points exhibited no diagnostic utility in distinguishing children from the WISC–III standardization sample from 684 children with learning disabilities (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" (p. 308).

Academic Achievement

Although diagnostically inaccurate, subtest scatter might be related to such important social criteria as academic achievement. This hypothesis was tested by Watkins and Glutting (2000), who analyzed the incremental validity of WISC–III subtest scatter in predicting academic achievement. They sequentially regressed subtest level and scatter 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 ($R = .72–.75$) and exceptional ($R = .36–.61$) children. Profile scatter did not aid in the prediction of achievement beyond general ability level for either group.

The standardization sample ($N = 2{,}974$) of the Woodcock-Johnson Psycho-Educational Battery–Revised (WJ–R; Woodcock & Johnson, 1989) provided another powerful test of the value of subtest scatter in predicting academic achievement (McGrew & Knopik, 1996). The number of significant cognitive strengths and weaknesses was calculated for each child, and children with high and low numbers of intercognitive strengths and weaknesses were then compared on academic achievement criteria. The presence of a large number of significant cognitive strengths or weaknesses was not related to academic problems in reading, writing, or mathematics.

Subtest Profiles

Diagnostic Accuracy

SCAD profile. Kaufman (1994a) noted that children with disabilities often score relatively low on WISC–III Symbol Search (*SS*), Coding (*CD*), Arithmetic (*AR*), and Digit Span (*DS*) subtests. Kaufman coined the acronym *SCAD* for this pattern of subtest scores and asserted that it is "a potent land mine for identifying children with neurological impairment or with exceptionalities" (p. 224).

In accord with this hypothesis, Prifitera and Dersh (1993) found that the SCAD profile was demonstrated by a significantly higher proportion of children with learning disabilities than children from the WISC–III standardization sample. Watkins, Kush, and Glutting (1997a) also found the SCAD profile to exceed the normative rate among 363 students with learning and emotional disabilities. Further overrepresentation of SCAD profiles was reported by Ward, Ward, Hatt, Young, and Mollner (1995) among 444 students with disabilities.

More recently, Mayes et al. (1998) found the SCAD profile among 11% of their sample of 45 students with learning disabilities. Although there was a greater proportion of SCAD profiles found among children with disabilities, when SCAD profiles were used to classify students into disabled and nondisabled groups, SCAD scores operated at near-chance levels. Accordingly, the SCAD profile demonstrated little utility for the differential diagnosis of disabilities.

ACID profile. The *ACID* profile (*A*R, *C*D, Information [*I*N], and *D*S subtests) has also been advanced as characteristic of learning-disabled students (Vargo, Grossner, & Spafford, 1995). Prifitera and Dersh (1993) compared percentages of children with the ACID profile from learning-disabled (LD) and attention-deficit/hyperactivity disorder (ADHD) samples, with the percentages found in the WISC–III standardization sample. They found a greater prevalence of ACID profiles in the clinical samples, with approximately 5% of the children with LD ($n=99$) and 12% of the children with ADHD ($n=65$) evidencing such a profile. In contrast, the ACID profile occurred in only 1% of the cases from the standardization sample.

The prevalence of ACID profiles among other samples of students with disabilities has been inconsistent (Frederickson, 1999). Mayes et al. (1998) found the ACID profile among 8.9% of their sample of 45 students with learning disabilities. Ward et al. (1995) found the ACID profile in 4.7% of a sample of 382 children with learning disabilities. Watkins, Kush, and Glutting (1997b) reported a prevalence rate of 4.1% among 612 students with learning disabilities. Swartz, Gfeller, Hughes, and Searight (1998) found that 6% of children with ADHD and 3.2% of children with learning disabilities in their sample ($N=81$) exhibited the ACID profile. However, the prevalence of the ACID profile among Daley and Nagle's (1996) sample of 165 students with learning disabilities was only 1.0%.

Although generally more prevalent among children with learning disabilities than among children without disabilities, the ACID profile was unable to accurately classify students into disabled and nondisabled groups. Watkins et al. (1997b), for example, noted that a randomly selected child with a learning disability would exhibit a more severe ACID profile than a randomly selected child from the WISC–III standardization sample 60% of the time. This accuracy dropped to 54% when children with other disabilities were included in the comparison group. In recognition of this low diagnostic accuracy, Ward et al. (1995) judged the ACID profile to "have little utility in differential diagnosis" (p. 275).

Freedom from distractibility profile. Factor analytic studies of the WISC–III have consistently identified a two-subtest factor (AR and DS) that has historically been called Freedom from Distractibility (FD; Wechsler, 1991). Many authors have hypothesized that performance on these two subtests "is greatly facilitated by attention and concentration, whereas it is impaired by distractibility and anxiety" (Kaufman, 1994a, p. 209). Consequently, researchers and practitioners have assumed that low FD scores are clinical indicators of ADHD. For example, the WISC–III manual reported that children with ADHD attained relatively low FD scores in comparison with their verbal comprehension (VC) and perceptual organization (PO) abilities.

A quantitative test of the ability of the FD factor to differentially identify children with ADHD was reported by Anastopoulous, Spisto, and Maher (1994). They administered the WISC–III and several behavior rating scales to 40 children with ADHD and found that, on average, FD scores were significantly lower than VC and PO scores. FD scores also correlated significantly with teacher measures of inattention ($r=-.49$), but not with teacher ratings of impulsivity, hyperactivity, or other internalizing or externalizing problems. Only 23% of the students with ADHD exhibited characteristic FD profiles, however, suggesting low diagnostic accuracy if used to identify individuals with ADHD.

Similar ambiguous relationships between FD scores and behavior ratings were reported by Lowman, Schwanz, and Kamphaus (1996). They found a moderate relationship between FD scores and teacher ratings of achievement problems. However, teacher ratings of attention and hyperactivity were not significantly related to FD scores. Likewise, FD scores of 126 children with attention and learning problems were not significantly related to parent and teacher ratings of hyperactivity and attention or laboratory measures of sustained attention/concentration (Riccio, Cohen, Hall, & Ross, 1997). Given this pattern of relations, Riccio et al. recommended that low performance on the FD should not be interpreted "as indicating the presence or absence of ADHD" (p. 36) and Lowman et al. "cautioned against using the FFD [*sic*] score as a measurement of ADHD symptoms" (p. 20).

Inadequate diagnostic accuracy and incongruent construct validity evidence for the FD has been reported by other researchers (Gussin & Javorsky, 1995; Reinecke, Beebe, & Stein, 1999). Factor analytic reports also suggested that the FD factor may be related more to quantitative skills or memory functioning than to attention (Keith & Witta, 1997). This constellation of evidence led Kaufman and Lichtenberger (2000) to conclude that the WISC–III FD score cannot be used as a diagnostic test for ADHD.

Wechsler developmental index. Wechsler's Deterioration Index (WDI) was originally developed as an indicator of cognitive impairment that was hypothesized to be sensitive to brain injury in adults (Livesay, 1986). Conceptually, the WDI was composed of two groups of Wechsler subtest scores: (a) *hold* subtests, which were considered to be insensitive to deterioration in brain injury; and (b) *don't hold* subtests, which were judged vulnerable to intellectual decline.

Application of the WDI with children was suggested by Bowers et al. (1992), given that neuropsychological deficits have often been hypothesized to account for learning and attentional difficulties in children. Bowers et al. recommended that the WDI be renamed the Wechsler Developmental Index because children's cognitive skills are not deteriorating but, rather, assumed to be developing unevenly. Research with the WISC–R found that children in learning disability programs scored significantly higher on the WDI than did children not placed in special education programs (Bowers et al., 1992; Klein & Fisher, 1994).

Neither WISC–R study, however, analyzed the accuracy of the WDI in making learning disability diagnoses for individual students; that is, the ability of mean group differences to accurately diagnose individuals. This deficiency was remedied by Watkins (1996), who found that the WDI performed at chance levels when distinguishing 611 students diagnosed as learning disabled from those diagnosed as emotionally disabled ($n=80$) or mentally retarded ($n=33$) as well as from 2,200 simulated random normal cases. Thus, the WDI was an inaccurate diagnostic indicator of learning disabilities.

Learning disability index. The Learning Disability Index (LDI; Lawson & Inglis, 1984, 1985) was also hypothesized to relate to specific neuropsychological deficits among students with learning disabilities. Lawson and Inglis conjectured that Wechsler subtests are sensitive to the presence of learning disabilities in direct proportion to their verbal saturation. Consequently, LDI scores are calculated from the second principal component of the Wechsler scales because it reflects a verbal-performance dimension. The first component constitutes a general intellectual dimension, and ignoring it in the calculation of the LDI substantially removes general ability from LDI scores.

Comparisons of groups of students with and without learning disabilities found significantly higher mean WISC–R LDI scores among students with learning disabilities than among students in regular education (Clampit & Silver, 1990; Lawson & Inglis, 1985). Statistically significant group LDI differences were subsequently interpreted as evidence that the LDI is diagnostically effective. For example, Kaufman (1990) concluded that the LDI is "quite valuable for distinguishing learning-disabled children from normal children" (p. 354).

As with the WDI, however, previous investigations of the LDI did not report its diagnostic accuracy when making decisions about individual children. Watkins, Kush, and Schaefer (2002) filled that evidential lacuna by comparing the WISC–III LDI scores of students previously diagnosed with learning disabilities ($N=2{,}053$) with those of students without learning disabilities ($N=2{,}200$). Subsamples of youth with specific reading ($n=445$) and math ($n=168$) disabilities permitted further assessment of the diagnostic efficiency of the LDI. Results revealed that the LDI rendered a correct diagnostic decision only 55% to 64% of the time. According to Swets (1996), diagnostic accuracy rates between .50 and .70 show low accuracy, .70 to .90 represent medium accuracy, and .90 to 1.00 denote high accuracy. Thus, diagnostic accuracy of the LDI is inadequate for scientific practice.

Diagnosis of emotional disabilities. Although children and adolescents with emotional disabilities (ED) have often demonstrated lower IQ scores than their peers without ED (Teeter & Korducki, 1998), there is considerable variability across studies. PIQ scores have generally been higher than VIQ scores for children and adolescents with ED and delinquency, but the significance and diagnostic accuracy of these differences are equivocal (Lynam, Moffitt, & Stouthamer-Loeber, 1993; Wong & Cornell, 1999). Typically, the magnitude of PIQ > VIQ difference scores does not reach levels that would be considered especially unusual in the standardization population. Likewise, there has been no consensus regarding which specific subtest profiles *should* characterize children or adults with emotional disabilities (Schretlen, Bobholz, & Benedict, 1992). Even children with statistically unusual subtest profiles on the Kaufman Assessment Battery for Children (Kaufman & Kaufman, 1983) were no more likely to receive special education than was a comparison group with more typical profiles (Glutting, McGrath, Kamphaus, & McDermott, 1992). Based on this literature, Teeter and Korducki (1998) concluded that "there are no distinctive Wechsler patterns that can provide reliable, discriminative information about a child's behavior or emotional condition" (p. 124). Likewise, subtest patterns were found to be inefficient in identifying adults with psychiatric disorders (Piedmont, Sokolove, & Fleming, 1989).

Academic Achievement

In contrast to the robust relationship between general ability and achievement (Jensen, 1998), specific subtest profiles have been unable to explain much variation in achievement measures. For example, neither the SCAD profile nor the ACID profile was a robust predictor of academic achievement among children with disabilities (Watkins et al., 1997a, 1997b). Nor were unusual subtest profiles on the Differential Ability Scales (DAS; Elliott, 1990) significantly associated with academic performance (Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998).

Further, the incremental validity of subtests for forecasting academic performance has been weak. For example, Watkins and Glutting (2000) sequentially regressed subtest level and shape information onto reading and mathematics achievement scores for children with and without disabilities. Profile elevation accounted for 52% to 56% of the variance in achievement among the 1,118 children from the nationally representative WISC–III and Wechsler Individual Achievement Test (WIAT; Wechsler, 1992) linking sample and 13% to 37% of the variance in achievement for 538 students with disabilities. Addition of subtest profile shape information accounted for an additional 5% to 8% of the variance in academic achievement. Only two shape patterns contributed to the prediction of achievement: relatively high verbal scores positively predicted reading and math achievement and relatively low scores on the AR subtest negatively predicted math achievement.

Even weaker predictive results have been found with the standardization sample of the DAS. McDermott and Glutting (1997) reported that DAS subtests incremented information by only 5.5% beyond the 34% available from general ability when compared with concurrent academic achievement. Nor did DAS factor scores provide substantial improvements over general ability in the prediction of academic criteria (Youngstrom, Kogos, & Glutting, 1999). Further, DAS subtests did not explain any unique variance in achievement beyond general and specific factor scores, even for children with significant DAS subtest and factor variability (Kahana, Youngstrom, & Glutting, 2002).

Empirical Subtest Profiles

By applying multivariate statistical methods to large representative samples, it is often hoped that subtest profiles that reliably discriminate between clinical and normal groups can be identified. Typically, standardization samples are used for initial subtest profile formation. These profiles can then be applied to clinical groups to determine their stability, diagnostic utility, and predictive validity.

The WISC–III has received the most attention in this search for normative subtest profiles. For example, Konold, Glutting, McDermott, Kush, and Watkins (1999) cluster analyzed the 10 mandatory WISC–III subtests and obtained eight core profiles. Although defined predominately by overall ability level, VIQ-PIQ differences were also reflected in several profiles. Similar conclusions were drawn from a cluster analysis of all 13 WISC–III subtests (Glutting, McDermott, & Konold, 1997). Nine core profiles resulted, with five determined primarily by general ability level. Nevertheless, several profiles also demonstrated variation in the four dimensions uncovered by factor analytic investigations of the WISC–III (Keith & Witta, 1997). These studies provide little support for the reliable disaggregation of individual subtests into profiles independent of their factor membership.

To determine whether empirical subtest profiles reliably replicated in clinical groups, the eight core profiles identified by Konold et al. (1999) in the WISC–III normative sample were applied to 1,025 students with learning and emotional disabilities (Glutting, McDermott, Watkins, Kush, & Konold, 1997). It was found that profiles overlapped around 94% between the normal and exceptional groups. Congruent with results from the WISC–R (Watkins & Kush, 1994) and WAIS–R (Maller & McDermott, 1997), children with learning and emotional disabilities were no more likely than children in general to exhibit rare or unusual WISC–III subtest profiles.

Recognizing that subtests are marked by limited reliability and construct integrity, some researchers have created empirical profiles from factor scores. For example, Donders (1996) clustered scores from the four factors of the WISC–III and obtained six core profiles: three distinguished primarily by level of performance, but three also manifesting differences in VC, PO, or Perceptual Speed (PS) scores. In contrast to children, fewer profiles and factorial influences have been found with adults. Donders, Zhu, and Tulsky (2001) submitted the four factor scores from the WAIS–III to cluster analysis and found five core profiles. Three were predominantly characterized by general ability level but two also displayed variability on the PS factor.

More complex empirical groupings were reported when IQ factor scores and academic achievement scores were simultaneously clustered (Glutting, McDermott, Prifitera, & McGrath, 1994). Modest replication was achieved (Ward, Ward, Glutting, & Hatt, 1999), but no studies have assessed the relative prevalence of these normative IQ factor and achievement profiles in clinical groups. Nor have studies tested the temporal stability of empirically formed profiles. Thus, it is not possible at this time to draw any conclusions regarding the stability or predictive validity of normative subtest profiles. The failure of WISC–R, WAIS–R, and WISC–III empirical subtest profiles to reliably discriminate between clinical and normal groups suggests, however, that they lack diagnostic utility.

Summary

Subtest scatter has continued to receive clinical attention despite previous literature reviews demonstrating that it was ineffective for distinguishing clinical from normal groups. Recent research results are remarkably consistent with past reviews: subtest scatter is an invalid diagnostic indicator and incapable of incrementally predicting academic achievement. Several studies are particularly important because they demonstrated no relationship between subtest scatter and academic achievement in large, nationally representative samples of students without disabilities (Kahana et al., 2002; McDermott & Glutting, 1997; McGrew & Knopik, 1996; Watkins & Glutting, 2000; Youngstrom et al., 1999). Based on their review of IQ subtest scatter research, Kline et al. (1996) suggested that psychologists "have pursued scatter analysis . . . with little success. It is time to move on" (p. 11). This 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.

Not all unique subtest profiles could be reviewed here, given that more than 75 subtest patterns have been identified for the Wechsler scales alone (McDermott, Fantuzzo, & Glutting, 1990). Nevertheless, results have been consistent in indicating that subtest profiles offer little diagnostic or predictive validity advantage over global IQ indices. 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). Clearly, abundant scientific evidence and expert consensus recommend against the use of subtest profile analysis for the differential diagnosis of psychopathology.

IDENTIFICATION OF COGNITIVE STRENGTHS AND WEAKNESSES WITH SUBTEST ANALYSIS

Although there is general agreement that IQ subtest–based diagnosis should be eschewed, the use of subtest profiles for *hypothesis generation* 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).

These assessment-for-intervention concepts have been operationalized via elaborate subtest profile interpretation systems (Groth-Marnat, 1997; Kaufman, 1994a; Sattler, 2001) that are taught to psychologists in training (Alfonso et al., 2000) and applied widely in clinical practice (Pfeiffer et al., 2000). To operate as envisioned, however, these interpretative systems must meet three sequential conjunctive conditions. First, subtest profiles must be robustly associated with performance in such socially important endeavors as academic achievement and psychosocial behavior. If subtest profiles are not substantially related to important social criteria, then hypotheses generated from subtest variation cannot be useful. Second, subtest-based hypotheses must be consistently confirmed by other information; that is, "unless an interpretation is supported by multiple pieces of data, it may not be strongly validated" (Kaufman & Lichtenberger, 2000, p. 178). Third and finally, interventions that result from validated hypotheses must exhibit treatment validity; that is, they must selectively remediate cognitive, achievement, or behavioral weaknesses.

Relationship between Subtests and Important Social Criteria

Subtest interpretation systems provide hundreds of hypotheses to consider when IQ subtest patterns are obtained (Kaufman, 1994a; Sattler, 2001). For example, low performance on the Picture Arrangement (PA) and Comprehension (CM) subtests suggests poor social adjustment; equal VIQ and PIQ scores indicate an absence of emotional distress; a large difference in performance between digits forward and digits backward indicates anxiety; and low scores on DS, AR, and CD subtests identify anxiety, attentional deficits, or both

(Banas, 1993; Drummond, 2000; Groth-Marnat, 1997; Kellerman & Burry, 1997). The enormous number of hypotheses makes it impossible to review the entire scientific literature. Nevertheless, it is instructive to examine in greater detail the evidence relating to several subtest-based hypotheses.

Specific Hypotheses

Social adjustment. The relationship between social adjustment and performance on PA and CM subtests has received considerable attention. Lipsitz, Dworkin, and Erlenmeyer-Kimling (1993) administered Wechsler scales and two measures of social adjustment to groups of high-risk and normal comparison children. PA scores showed no relation with either measure of social adjustment. Further, Campbell and McCord (1996) found that PA scores were not significantly better than FSIQ scores in predicting participants' ability to interpret the nonverbal behavior of others. These negative results for the PA subtest have been replicated with different samples, criterion tests, and IQ tests (Beebe, Pfiffner, & McBurnett, 2000; Campbell & McCord, 1999). Consequently, Kamphaus (2001) concluded that "making an inference regarding social adjustment/judgment based on PA scores is contraindicated by research" (p. 467).

However, conflicting outcomes have been reported for the CM subtest. Campbell and McCord (1999) found that relatively poor CM performance did not predict clinically significant social or peer relationship problems. In contrast, Lipsitz et al. (1993) found that CM scores were significantly related with one adjustment measure among at-risk children. However, other Wechsler subtests not hypothesized to reflect social adjustment were also significantly correlated with that adjustment measure. Lipsitz et al. suggested that this univariate pattern of correlations demonstrated the powerful role of general intelligence rather than a specific relation between CM and social adjustment. Nevertheless, Beebe et al. (2000) reported that CM scores demonstrated incremental validity beyond FSIQ in predicting mother-reported conduct problems and teacher-reported adaptability ($r_{\text{partial}} = -.22$ and .23, respectively). CM was, however, not significantly related to five other measures of child adjustment and behavior once general intelligence was controlled. Beebe et al. noted that only 5% of the children with low teacher-reported adaptability displayed a significant relative weakness on the CM subtest and cautioned clinicians "against the use of subtest scores to support cognitive explanations for poor social functioning on a case-by-case basis" (p. 100).

In summary, the PA subtest consistently failed to demonstrate a relationship with social adjustment once general intelligence was controlled. Although the CM subtest was sometimes related to ratings of students' conduct, there was inconsistent evidence regarding its relation to social adjustment. Even when found, relations between CM and social functioning were small and inconsistent. Thus, "normative interpretation of Picture Arrangement or Comprehension as measuring social competence is not warranted" (Campbell & McCord, 1999, p. 222).

Coding speed. Kaufman and Lichtenberger (2000) opined that "changes in the rate of responding during [Coding] may be related to motivation, distraction, fatigue, boredom, and so forth. Thus it is a good idea to note the number of symbols copied during each of the four 30-second periods within the 120-second limit" (p. 115). Nicholson and Alcorn (1994) were even more adamant that "there should be a steady increase in the number of symbols completed at the end of each 30-second period" (p. 8). Nevertheless, they provided no evidence to support the putative relationship between examinees' rates of responding on the CD subtest and emotional conditions.

To test the hypothesis that it is typical for children to show an increase in the number of symbols copied over successive time intervals, Dumont, Farr, Willis, and Whelley (1998) monitored the rate of response on the CD subtest of 351 children for each of four successive 30-second intervals. They found that *no* child increased in rate of production across the final three intervals. In fact, most children showed decreases in rate across some intervals. Nor was the pattern of responding on the CD subtest related to any disability. In fact, children in the WISC–III standardization sample also demonstrated a small average decrement in production after the first 30-second CD interval (Sattler, 2001). Thus, a slight decrease in response rate on CD is the norm rather than the exception. This resounding refutation of the CD speed hypothesis prompted Dumont et al. to wonder "how often well-meaning clinicians may have been led to believe that a student might be unmotivated, depressed, or suffering from deficits in learning ability because the student demonstrated the same decline in Coding speed shown by 97.7% of the children in the present sample" (p. 116).

Processing speed. After noting that the WISC–III contains a PS factor composed of CD and SS subtests, Blaha and Wallbrown (1996) speculated that processing speed might be related to reading disabilities. Kaufman (1994a), in contrast, asserted that the WISC–III PS factor is reflective of uncooperative test-taking behaviors and could more aptly be renamed "Freedom from Bad Attitude" (p. 209). Based on his clinical analysis of the WISC–III, Kaufman hypothesized that children with uncooperative test-taking behaviors are more likely to show a PO > PS score pattern.

The purported relationship of processing speed to reading and test-taking behaviors appears to be based on clinical impressions and personal testimonials. When tested with 283 children without disabilities and 636 children referred for special education evaluation, the WISC–III PS factor made no contribution to the prediction of WIAT reading scores beyond that provided by FSIQ (Glutting, Youngstrom, Ward, Ward, & Hale, 1997). Nor did the processing speed component of the DAS contribute to reading among a nationally representative sample of 1,185 children once general ability was considered (Youngstrom et al., 1999). Finally, processing speed exhibited no significant relationship to reading achievement among the standardization sample of the WJ–R (McGrew, Keith, Flanagan, & Vanderwood, 1997). These results are consistent with fundamental reading research, which has found that speed of processing letters, but not general speed of processing, predicts reading (Neuhaus, Foorman, Francis, & Carlson, 2001).

Although processing speed across its full range has not shown a relationship to achievement, it is possible that subtest profiles with unusually weak or strong processing speed scores would display differential reading test performance. This was tested with the DAS standardization sample (Oh, Glutting, & McDermott, 1999). Children with rare (i.e., prevalence ≤5%) processing speed strengths and weaknesses were matched on age, race, gender, parents' education level, and general ability with an equal number of comparison participants without unusual processing speed profiles. Relative to these matched participants, children with processing weaknesses and strengths showed no significant differences in reading achievement or on six teacher-rated indices of behavioral adjustment. Additionally, referred children with normal visual-motor skills and low academic achievement did not significantly differ from normative standards on the WISC–III PS factor (Tiholov, Zawallich, & Janzen, 1996).

Nor did processing speed succeed in a test of its relationship with uncooperative test-taking behavior among a nationally representative sample of 640 children (Oakland, Broom, & Glutting, 2000). WISC–III PS scores shared only 2% variance with uncooperative test-taking behaviors, as quantified by the *Guide to the Assessment of Test-Session Behavior* (GATSB; Glutting & Oakland, 1993). Further, GATSB uncooperative scores were not significantly related to an index created by subtracting PS scores from PO scores. Oakland et al. concluded that Kaufman's PS hypothesis exhibited little clinical utility and recommended reliance on normative behavioral measures of test-taking behavior in preference to distal cognitive indices.

General Hypotheses

Subtest profiles are often assumed to be related to a wide variety of learning and behavioral variables (Kaufman, 1994a; Sattler, 2001). Thus, hypotheses about learning and behavior are commonly generated from subtest profiles.

Academic achievement. "Variables worthy of scientific attention provide information not already available; nonincremental sources of variance do not" (Lubinski, 2000, p. 415). Accordingly, a number of studies have investigated the incremental validity of IQ subtests and specific factors in forecasting academic achievement in schools and performance in vocational training courses. Results consistently find that subtest profiles and specific factors share a small amount of variance (i.e., up to 8%) with achievement and training performance beyond that accounted for by general ability (Hale & Saxe, 1983; Hale & Raymond, 1981; Kahana et al., 2002; Kline et al., 1992; McDermott & Glutting, 1997; McGrew & Knopik, 1996; Ree & Carretta, 1997; Watkins & Glutting, 2000; Youngstrom et al., 1999). For example, the WISC–III was found to provide only two specific shape patterns that contributed to prediction of achievement: relatively high verbal scores positively predicted both reading and math achievement, and relatively low scores on the AR subtest negatively predicted math achievement. However, neither the SCAD profile nor the ACID profile was a robust predictor of academic achievement (Watkins, et al., 1997a, 1997b).

Even among researchers who posit influences beyond general ability, academic achievement is not generally assumed to be determined by cognitive subtests acting independently. Rather, achievement is presumed to be primarily determined by the higher-order general ability (g) factor, followed by first-order ability factors. For example, auditory processing and crystallized intelligence factors were found to be related to specific reading subskills beyond *g* in the WJ–R standardization sample (McGrew et al., 1997). McGrew et al. cautioned, however, that "practitioners should not misinterpret this research to support any form of individual subtest interpretation" (p. 205).

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 co-normed 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 only 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 child who is quiet during testing is assumed to be retiring in other social 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). Consistent with previous studies, test-session behaviors were correlated with teacher ratings at .12 for 72 students from the GATSB normative sample and at .16 for 140 referred students (Glutting et al., 1996). Thus, more than 97% of the variation in scores on test-session and classroom behavior rating scales is unique. Given this weak relationship, Oakland and Glutting (1998) encouraged clinicians to "refrain from drawing conclusions as to the generalizability of test observations to conditions outside the immediate test situation" (p. 301).

Because test-session behaviors tend to be situationally specific, however, they might be related to concurrent test scores. A review of the evidence on this relationship found that, on average, the correlation between test behaviors and IQs obtained during the same test session was –.34 (Glutting et al., 1996). Thus, information regarding the test session may be useful for validating the integrity of obtained IQ scores. For example, among the GATSB normative sample, those children with inappropriate test-taking behaviors averaged WISC–III IQ scores from 7 to 10 points lower than did children with more suitable test behaviors (Oakland & Glutting, 1998). However, there was little differential variability across IQ subtests (Oakland et al., 2000). In fact, when the relation was analyzed in reverse, global ability accounted for 9.2% of test-session behaviors and the addition of WISC–III subtests explained another 3.2% (McDermott & Glutting, 1997).

Classroom behaviors. Many 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 (1994a), children with the 5% 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 the groups with and without unusual subtest profiles 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, et al., 1998).

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 hypothesized relationships between subtest profiles and socially important academic and psychosocial outcomes is, at best, weak: subtest profile information contributes 2% to 8% variance beyond general ability to the prediction of achievement and 2% to 3% to the prediction of learning behaviors and test-session behaviors. Hypothesized relationships between subtest profiles and measures of psychosocial behavior 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. It lacks utility for making special education placement decisions, and produces scores for people that lack both reliability and validity" (p. 98). Thus, conclusions drawn from subtest analysis "are based on the clinician's acumen and not on any sound research base" (Kamphaus, 2001, p. 598).

Confirmation/Disconfirmation of Hypotheses: Flaws in Reasoning

As previously documented, the hypothetical-deductive process is likely to begin with incorrect or weak hypotheses regarding the relation between subtest profiles and socially important achievement and psychosocial behaviors. One danger of introducing false hypotheses is that erroneous conclusions about examinees can result (Dumont et al., 1998). This would not be of great concern if the confirmation or refutation of an hypothesis is an objective and accurate process (Kaufman, 1994a, p. 31). Unfortunately, ambiguity and error contaminate the hypothesis confirmation-disconfirmation process because cognitive errors frequently accompany and impair human decision making (Faust, 1984). Cognitive errors are well documented and have been consistently demonstrated by both novice and seasoned clinicians (Faust, 1986). Among the most ubiquitous of these flaws in human reasoning are underutilization of base rates, misjudgment of covariation, association of availability in memory with probability of occurrence, estimation of probabilities on the basis of similarity or representativeness, overreliance on confirmatory hypothesis-testing strategies, and a tendency to over-emphasize initial impressions and underestimate confidence intervals (Baron, 1994; Faust, 1984, 1986; Tracey & Rounds, 1999). Of course, fundamental limitations in human information processing capabilities are also universally applicable.

Most subtest profile analysis systems do not recognize that clinical decision making is exquisitely sensitive to these well-known human-reasoning limitations. Thus, the typical clinician is not only vulnerable, but also insensitive to their operation. Even worse, subtest analysis advocates suggest that diagnostically unreliable subtest profiles somehow become more useful when integrated informally and impressionistically with a complex array of objective and subjective assessment data. However, the research literature has unequivocally documented that clinicians are most susceptible to suboptimal decision making in exactly this type of situation (Faust, 1986) and should rely on actuarial rather than clinical predictions (Dawes, Faust, & Meehl, 1989). Faust (1990) suggested that this "common belief in the capacity to perform complex configural analysis and data integration might thus be appropriately described as a shared professional myth" (p. 478).

The complex interaction of cognitive errors and limitations of reasoning that can cause the clinical hypothesis testing process to go awry have been trenchantly described by Faust (1986):

> Despite what supervisors tell their students about integrating data and examining configural relations, the typical cognitive processes underlying psychodiagnosis are likely much closer to that of this example: The clinician proceeds to collect sufficient information to formulate and support (not test) hypotheses. As data are collected or analyzed, the clinician formulates hypotheses about the patient, often quite early in the process. Hypotheses are based on a few salient cues. Subsequent data collection or analysis is overly influenced by these hypotheses; although they may be further elaborated or refined, they are rarely changed substantially. . . . Much of the subsequent search may be little more than an attempt to find sufficient evidence to confirm conclusions. The final conclusions are based not on complex configural analysis but on "counting noses.". . . Data that might conflict with conclusions are either explained away (e.g., as test artifact), ignored, or molded to fit the hypothesis through mental gymnastics. The dynamic formulation used to explain the assumed pathological state is shaped by additional bad judgment habits. . . . The process becomes an exercise in redundancy, extending the initial diagnostic conclusions to questions of cause while ensuring that a satisfactory answer is obtained regardless of its accuracy. In fact, no matter what their accuracy is, the search for such explanations is likely to increase confidence. (p. 424)

Beyond the judgmental difficulties inherent in a clinical hypothesis approach, basic psychometric principles predict a high rate of erroneous decisions. By beginning the decision-making process with an essentially random component (i.e., the subtest profile) and then searching for confirmation, the clinician cannot increase, and may decrease, judgment accuracy when trying to detect a low-prevalence strength or weakness (Meehl & Rosen, 1955). Thus, flawed decision-making processes inflate the probability of unsound clinical hypotheses being accepted and subsequently used to generate interventions (Meehl, 1997).

Treatment Validity of Subtest Profile-Based Interventions

Major subtest interpretative systems allow the identification of specific subtest profiles (reflective of aptitudes) that can be matched to interventions (treatments) so that individual clients uniquely benefit. Regardless of the verity of subtest-based hypotheses, identified cognitive aptitudes must lead to specific treatments that are differentially predictive of treatment success for particular examinees. As an example, a case study presented by Kaufman (1994a, pp. 348–360) illustrates that relatively low performance on a constellation of IQ subtests was interpreted as a deficit in auditory short-term memory (aptitude), requiring a focus on experiential learning activities (treatment) that was presumed to produce optimal instructional benefit for that child. Generically, this is called an *aptitude by treatment interaction* (ATI; Cronbach, 1975).

Several ATI models that differ on presumed aptitudes have been widely disseminated. Such models include matching learning modality (i.e., visual, verbal, kinesthetic, and tactile), cognitive-processing mode (i.e., simultaneous versus sequential), or neuropsychological function to instructional methods. Research has failed to support any of these models (Arter & Jenkins, 1979; Good, Vollmer, Creek, Katz, & Chowdhri, 1993). In fact, Gresham and Witt (1997) reported that they "could not locate a single study demonstrating a significant ATI based on neuropsychological assessment, interpretation, and treatment prescription with children having mild learning problems" (p. 253). Belief in the existence of replicable ATIs, although seemingly logical, has been incapable of attaining scientific support over the past four decades (Cronbach, 1975; McNemar, 1964; Reschly, 1997).

Furthermore, IQ tests have been unable to demonstrate treatment validity, regardless of ATI matching. That is, they have not been shown to lead to effective treatments, instructional programs, or strategies to improve academic skills (Gresham & Witt, 1997; Reschly & Grimes, 1990). As noted by Bray, Kehle, and Hintze (1998) "in general, intelligence tests are excellent predictors of many tasks, but do not substantially aid in the planning of academic interventions" (p. 216).

METHODOLOGICAL CONSIDERATIONS

Although IQ subtest analysis has not demonstrated adequate diagnostic utility or treatment validity, it continues to be endorsed by assessment specialists and applied widely in training and practice. This enthusiasm has been expressed most vividly by Kaufman (1994a), who asserted that "I believe in profile interpretation" (p. 26). Many authors have found the popularity of IQ subtest analysis to greatly outstrip its meager scientific support (Bray et al., 1998; Gresham & Witt, 1997; Watkins, 2000). In fact, the widespread acceptance of IQ subtest analysis has variously been described as a reliance on clinical delusions, illusions, myths, or folklore.

The following 14 interrelated issues explain some of the fundamental methodological problems in research and practice that 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, 1994a).

1. Subtests are implicitly assumed to possess measurement precision similar to global IQ measures. They do not. For example, the internal consistency reliability coefficients of WISC–III VIQ, PIQ, and FSIQ scores are .95, .91, and .96, respectively. In contrast, the corresponding coefficients of WISC–III subtests range from .69 to .87, with a median value of .78 (Sattler, 2001). Only 3 of the 13 WISC–III subtests meet the reliability coefficient criterion of ≥.85 recommended by Hansen (1999) for making decisions about individuals and none meets the more stringent criterion of ≥.90 (Hopkins, 1998).

Additionally, standardization samples produce "best case" reliability estimates because standardization examiners are carefully trained and monitored for adherence to standardization procedures and test protocols are checked for accuracy by the test company. None of these "best case" procedures are typical in clinical practice. Consequently, errors in administration and scoring of IQ tests are common in clinical practice (Slate, Jones, Coulter, & Covert, 1992). In fact, examiner scoring errors were found to double the standard error of measurement of the WISC–R (Klassen & Kishor, 1996). Thus, actual reliability estimates will be lower by an unknown degree than those reported in standardization manuals (Feldt & Brennan, 1993; Thorndike, 1997).

Because internal consistency coefficients do not reflect all sources of measurement error, it is informative to also consider temporal stability estimates. Short-term test-retest coefficients of .94, .87, and .94 were reported for the WISC–III VIQ, PIQ, and FSIQ, respectively, for 353 children (Wechsler, 1991). However, the median short-term subtest stability coefficient was .74. VIQ, PIQ, and FSIQ stability coefficients were .87, .87, and .91, respectively, when 667 students enrolled in special education programs were twice evaluated with the WISC–III across a 2.87-year span (Canivez & Watkins, 1998). In contrast, the median long-term subtest stability coefficient was .68.

Furthermore, the increased error generated by the use of difference scores makes even the best subtest-to-subtest comparison unreliable (i.e., the reliability of the difference between WISC–III Block Design and Vocabulary is .76). For example, the median correlation between WISC–III VIQ-PIQ differences over 2.87 years was only .40 (Canivez & Watkins, 1998). Although 191 students exhibited a significant VIQ-PIQ difference (≥15 points) at initial testing, only 44% maintained such a large discrepancy upon retesting. Similar unstable results have also been reported for older adults who were retested with the WAIS–R (Ivnik, Smith, Malec, Petersen, & Tangalos, 1995). Cahan and Cohen (1988) detailed the low statistical power and high error rate involved in testing for the statistical significance of subtest score differences. These statistical difficulties were confirmed by Krauskopf (1991) and determined to be intractable by Silverstein (1993).

2. The dimensionality of cognitive abilities is often ignored and interpretation is inappropriately focused only on the lowest-level dimension (i.e., subtests). Factor analyses consistently reveal that approximately 50% of the common variance of a diverse set of cognitive tests comprise a general ability factor that is robustly predictive of academic and vocational outcomes (Jensen, 1998). At a more molecular level, several narrow factorial dimensions account for smaller amounts of common variance (e.g., spatial, verbal reasoning, quantitative, etc.). Individual subtests are at the lowest level of the IQ hierarchy. The score variation of an individual subtest, therefore, is due to the combined influence of the general ability factor, narrow ability factors, factors specific to each subtest, and, finally, measurement error. Interpretation of subtests as if they measured only one attribute rejects this nuanced account of subtest variability, ignores the primary source of common variance, and falls prey to the nominalistic fallacy—believing that a name reflects reality.

Further, most of the variance of general intelligence batteries have typically been found to be due to general intelligence rather than specific factors. For example, one partitioning of the common subtest variance of the WISC–III normative sample estimated that general ability accounted for 35% of the variance, four narrow factors accounted for an additional 18% of the variance, and all 13 subtests together accounted for another 28% of the variance (Blaha & Wallbrown, 1996). Alternative factor analytic methods have found contributions to total WISC–III score variance to be 71% for general ability, 19% for first-order factors, and 10% for error (Gustafsson & Undheim, 1996). Among multiple-aptitude batteries for adults, the proportion of general ability variance is around 64% and reliable nongeneral ability variance is about 16% (Ree & Carretta, 1997). Thus, reliance on lower-order subtests ignores the majority of explanatory and predictive variance carried by IQ tests.

Even worse, some measurement experts have noted that not all common variance may be psychologically important. For example, some portion of the common variance of IQ subtests is construct irrelevant method variance (Lubinski, 2000). This point was recognized by Thorndike (1994), who speculated that it is possible "that most of the differential profile patterns are really little more than unrecognized error variance" (p. 178). To be interpretable, then, subtests must demonstrate replicable incremental validity with external criteria *beyond* general intelligence and narrow ability factors (Lubinski, 2000).

3. Much IQ subtest research relies on statistical tests (e.g., ANOVA, correlation) that cannot assess both the linear and nonlinear components of subtest profiles (see Cattell, 1949). As noted by Mosel and Roberts (1954), profile comparisons require statistical treatments that are sensitive to trends in both level and shape. Additionally, univariate analyses are often conducted when multivariate methods would be preferable.

4. Most studies assume that groups of similarly diagnosed people represent meaningful, homogeneous categories. Yet it is clear from the research literature that diagnostic reliability is sometimes suspect and there is often considerable heterogeneity among people with the same diagnosis (Garfield, 1978).

5. There is an overreliance on small samples of convenience that span broad age ranges. Reasonably equivalent subtest measurement error across age levels is assumed. This assumption is untenable given that internal consistency reliability coefficients vary across ages and subtests. Across ages, for example, the reliability of the WISC–III Block Design subtest grows from .77 at age 7 years to .92 at age 15 years. Across subtests, at age 15 years the WISC–III Object Assembly subtest reliability is .60 compared with the Vocabulary subtest's .91. Furthermore, special characteristics of the sample may make generalization of results questionable and small, unstable samples may make it difficult to arrive at accurate estimates of population effect sizes.

6. Subtest analysis relies on subtests maintaining the same relative relations across time. That is, it is assumed that the relationships among subtests remain invariant from initial standardization of the test until it is revised after 10 to 20 years of use. Flynn's (1987) research resoundingly invalidates this assumption. The phenomenon known as the Flynn effect suggests that IQ scores increase by about 3 points each decade, but the subtests that contribute to that global IQ increase do so disproportionately. For example, the Picture Arrangement subtest increased by 1.9 points between the WISC–R and

WISC–III normative samples, whereas the Information subtest declined by .3 points (Flynn, 1999a). Likewise, VIQ and PIQ scores were differentially affected. Inconsistent subtest, VIQ, and PIQ growth across time on the WISC–R and WISC–III was also demonstrated for a sample of children with learning disabilities (Truscott & Frank, 2001). From these studies it is clear that IQ subtest relationships change over time in a complex, unpredictable manner (Flynn, 1999b). Truscott and Frank noted that this "serves as another reason that practitioners should be wary of subtest analysis" (p. 330).

7. Many studies employ IQ subtest profiles for both initial formation of diagnostic groups and subsequent searches for profiles that define those groups. For example, children may be wholly or partially categorized as learning disabled based upon their WISC–III scores and then their WISC–III scores are examined for profiles that identify them as learning disabled. Such circular reasoning endangers external validity.

8. Claims for discovery of unique IQ subtest profiles are rarely made against the null hypothesis that such profiles are actually commonplace in the normal population and therefore unremarkable. It is not possible to know whether any profile is distinctive for a specific diagnostic group without knowledge of normal variation.

9. 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 and thereby transformed into person-relative metrics from their original population-relative metric (McDermott et al., 1990). For example, two hypothetical students' normative (population-relative) and ipsative (person-relative) WISC–III verbal scores are displayed in the Table. These two students share identical ipsative scores, but their normative scores are very different.

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 et al. (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 DAS and WISC–III. Both practical and theoretical analyses suggest that the mathematical properties of ipsative methods are profoundly different than those of familiar normative methods (Hicks, 1970; McDermott et al., 1990, 1992). Thus, ipsative subtest scores cannot be interpreted as if they possessed the reliability and validity of normative scores.

10. Identification of pathognomonic IQ subtest profiles has generally been based upon statistically significant group differences. That is, the mean subtest score of a group of children with a particular disorder is compared with 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, group mean-score differences do not support this interpretation (Frederickson, 1999). 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).

This situation illustrates reliance on classical validity

Table

Normative (Population-Relative) and Ipsative (Person-Relative) Wechsler Verbal Scores for Two Hypothetical Examinees

	Student A		***Student B***	
Subtest	**Norm Score**	**Ipsative Score**	**Norm Score**	**Ipsative Score**
Information	3	–5	9	–5
Similarities	7	–11	3	–1
Vocabulary	10	+2	16	+2
Comprehension	12	+4	18	+4
Mean	8	0	14	0

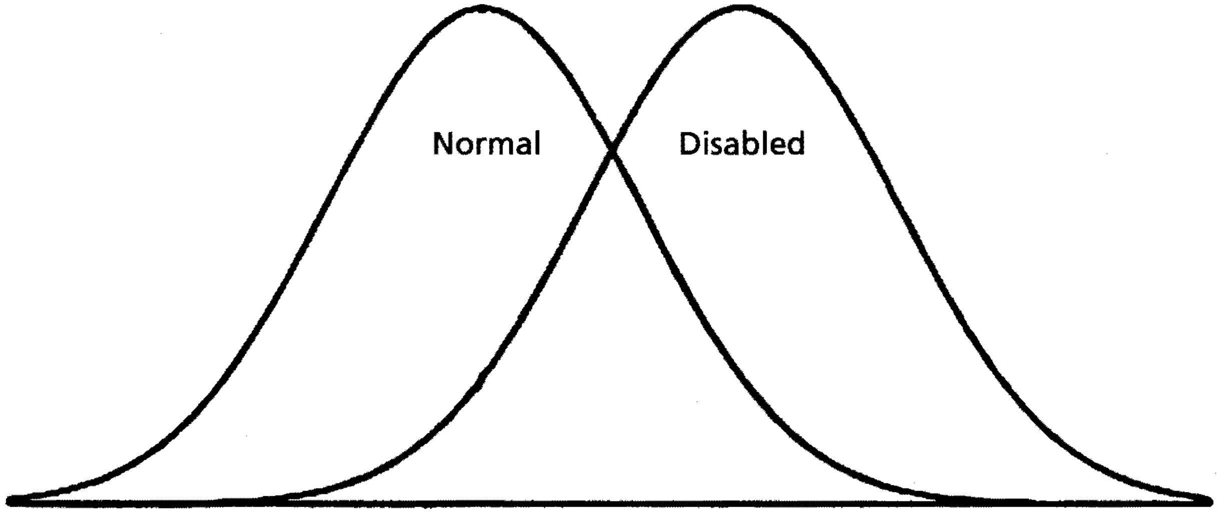

Figure 1. Two hypothetical overlapping score distributions.

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*. Figure 1 illustrates this dilemma. It displays hypothetical score distributions of children from normal and disabled populations. Group mean differences are clearly discernable, but the overlap between distributions makes it difficult to accurately determine group membership for those individuals within the overlapping distributions.

Errors in assigning individuals to normal or disabled groups are inevitable in psychology (Zarin & Earls, 1993). There are four possible outcomes when 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.

The relative proportion of correct and incorrect diagnostic decisions depends on the cutting score used. For example, cutting score X in Figure 2 produces a high true positive and a low true negative rate. That is, it correctly identifies those who are disabled but makes many mistakes for those who are not disabled. In contrast, cutting score Z makes few false positive errors but many false negative errors. Figure 2 graphically displays the trade-offs between sensitivity and specificity that are always encountered when test scores are used to differentiate groups (Zarin & Earls, 1993).

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, the full range of that test's diagnostic utility can be determined. 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).

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

11. Beyond cutting 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

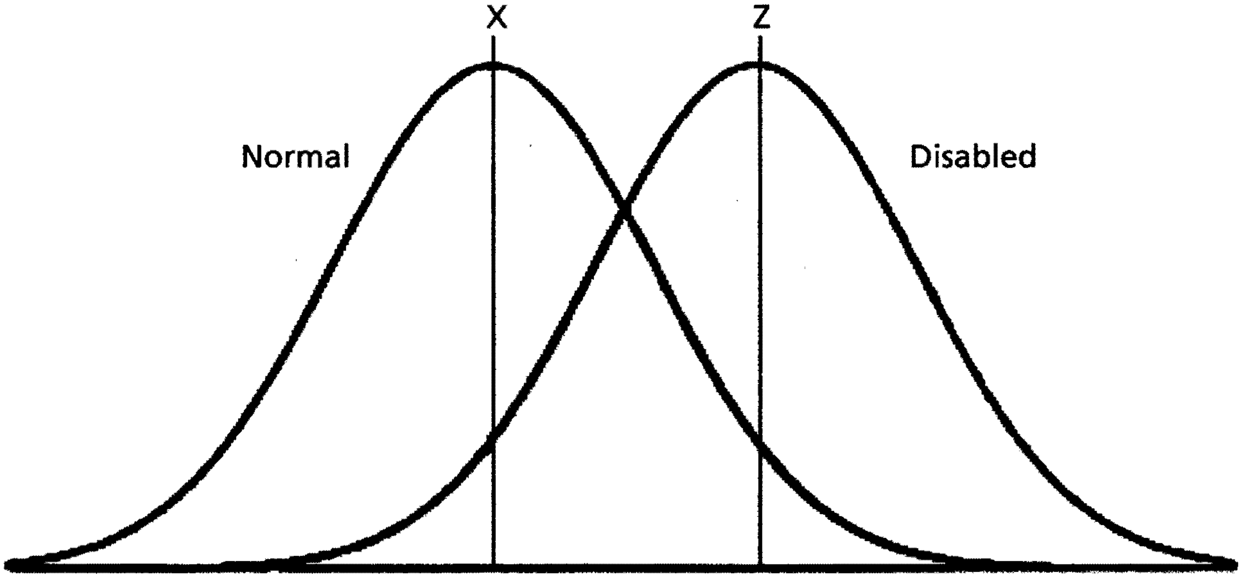

Figure 2. Two hypothetical overlapping score distributions with cutting scores.

identify accurately (Meehl & Rosen, 1955). This issue is relevant for much psychological practice and research because many disabilities are, almost by definition, unusual or rare.

12. Directly related to the previous methodological issues is the use of inverse probabilities encouraged by much subtest profile research. It is generally not understood that the probability of a particular score on a diagnostic test given membership in a diagnostic group is different than 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). This quandary can be illustrated with an hypothetical subtest analysis example. A small group of children with learning disabilities is located and WISC–III subtest scores are analyzed. It is found that many exhibit the ACID profile. Thus, the probability of the ACID profile is high given that the child is learning disabled. However, clinical use of subtest profiles is predicated on a different probability—namely, determining the probability that a referred child is learning disabled given an ACID profile. Retrospective analyses will systematically overestimate prospective accuracy (Dawes, 1993).

13. There is an implicit reliance on nonscientific methods of knowledge reflected in much research and practice with IQ subtests. First, there is an alarming eagerness to accept personal experience and insight as equal or superior to objective investigation as a source of knowledge (Cromer, 1993; Gambrill, 1990). For example, Kaufman (1994a) asserted that "our judgment, knowledge of psychology, and clinical training are more important than the obtained IQs" (p. 26) and Blumberg (1995) suggested that interpretation is "more an art than a science" (p. 97). However, experience is not necessarily synonymous with expertise (Gambrill, 1990) and "may lead to nothing more than learning to make the same mistakes with increasing confidence" (Skrabanek & McCormick, 1990, p. 28). Preference for impressionistic, subjective judgment over obtained cognitive indices ignores the substantial evidence regarding the value of actuarial assessment (Dawes et al., 1989) and high error rates associated with clinical judgment (Dawes, 1994).

Second, this review has explicated in detail that little supportive evidence for subtest analysis has been found. However, many advocates of subtest analysis appear to believe that subtest analysis is scientifically appropriate as long as there is no evidence *against* it. That is, they seem to believe that the burden of proof lies with the skeptic. Formally called the *ad ignorantium* fallacy, this appeal to lack of knowledge (Baron, 1994) turns the scientific method on its head and runs counter to professional testing standards and ethical codes (AERA, APA, & NCME, 1999; APA, 1992). As noted by Cromer (1993), "the burden of proof must be on the believer" (p. 156).

14. The complexity of nomothetic and idiographic dimensions are often ignored or misinterpreted in subtest analysis research and practice. For example,

Kaufman (1994a) asserted that critics of subtest analysis "rely almost exclusively on *group* data" (p. 36), whereas individual intellectual assessment is "unique to each person evaluated" (p. 32). However, "the uniqueness of a particular event can never be used as a ground for rejecting nomothetic formulations" (Meehl, 1996, p. 311). IQ scores are based on the average performance of members of the standardization sample. Critics who disdain group data as a foundation of criticism simultaneously embrace interpretations based on such group-based metrics as factor scores, subtest specificity indices, reliability coefficients, and so on. All test performance is the performance of individuals and permits probabilistic decisions about individuals, but there is no science of interpretation that is unique to one person (McDermott & Glutting, 1997).

References

Aiken, L. R. (1996). *Assessment of intellectual functioning* (2nd ed.). New York: Plenum.

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.

American Psychological Association. (1992). Ethical principles of psychologists and code of conduct. *American Psychologist, 47*, 1597–1611.

Anastopoulous, A. D., Spisto, M. A., Maher, M. C. (1994). The WISC–II freedom from distractibility factor: Its utility in identifying children with attention deficit hyperactivity disorder. *Psychological Assessment, 6*, 368–371.

Arter, J. A., & Jenkins, J. R. (1979). Differential-diagnosis-prescriptive teaching: A critical appraisal. *Review of Education Research, 49*, 517–555.

Banas, N. (1993). *WISC–III prescriptions: How to work creatively with individual learning styles*. Novato, CA: Academic Therapy Publications.

Baron, J. (1994). Thinking and deciding (2nd ed.). New York: Cambridge University Press.

Beebe, D. W., Pfiffner, L. J., & McBurnett, K. (2000). Evaluation of the validity of the Wechsler Intelligence Scale for Children–Third Edition comprehension and picture arrangement subtests as measures of social intelligence. *Psychological Assessment, 12*, 97–101.

Blaha, J., & Wallbrown, F. H. (1996). Hierarchical factor structure of the Wechsler Intelligence Scale for Children–III. *Psychological Assessment, 8*, 214–218.

Blumberg, T. A. (1995). A practitioner's view of the WISC-III. *Journal of School Psychology, 33*, 95–97.

Bowers, T. G., Risser, M. G., Suchanec, J. F., Tinker, D. E., Ramer, J. C., & Domoto, M. (1992). A developmental index using the Wechsler Intelligence Scale for Children: Implications for the diagnosis and nature of ADHD. *Journal of Learning Disabilities, 25*, 179–185.

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.

Cahan, S., & Cohen, N. (1988). Significance testing of subtest score differences: The case of nonsignificant results. *Journal of Psychoeducational Assessment, 6*, 107–117.

Campbell, J. M., & McCord, D. M. (1996). The WAIS–R comprehension and picture arrangement subtests as measures of social intelligence: Testing traditional interpretations. *Journal of Psychoeducational Assessment, 14*, 240–249.

Campbell, J. M., & McCord, D. M. (1999). Measuring social competence with the Wechsler picture arrangement and comprehension subtests. *Assessment, 6*, 215–223.

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.

Cattell, R. B. (1949). r_p and other coefficients of pattern similarity. *Psychometrika, 14*, 279–298.

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.

Cromer, A. (1993). *Uncommon sense: The heretical nature of science*. New York: Oxford University Press.

Cronbach, L. J. (1975). Beyond the two disciplines of scientific psychology. *American Psychologist, 12*, 671–684.

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.

Dawes, R. M. (1994). *House of cards: Psychology and psychotherapy built on myth*. New York: Fress Press.

Dawes, R. M., Faust, D., & Meehl, P. E. (1989). Clinical versus actuarial judgment. *Science, 243*, 1668–1674.

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

Donders, J. (1996). Cluster subtypes in the WISC–III standardization sample: Analysis of factor index scores. *Psychological Assessment, 8*, 312–318.

Donders, J., Zhu, J., & Tulsky, D. (2001). Factor index score patterns in the WAIS–III standardization sample. *Assessment, 8*, 193–203.

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

Drummond, R. J. (2000). *Appraisal procedures for counselors and helping professionals* (4th ed.). Upper Saddle River, NJ: Prentice-Hall.

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.

Dumont, R., Farr, L. P., Willis, J. O., & Whelley, P. (1998). 30–second interval performance on the coding subtest of the WISC–III: Further evidence of WISC folklore? *Psychology in the Schools, 35,* 111–117.

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. (1984). *The limits of scientific reasoning.* Minneapolis: University of Minnesota Press.

Faust, D. (1986). Research on human judgment and its application to clinical practice. *Professional Psychology: Research and Practice, 17,* 420–430.

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.

Flynn, J. R. (1987). Massive IQ gains in 14 nations: What IQ tests really measure. *Psychological Bulletin, 101,* 171–191.

Flynn, J. R. (1999a). Evidence against Rushton: The genetic loading of WISC–R subtests and the causes of between-group IQ differences. *Personality and Individual Differences, 26,* 373–379.

Flynn, J. R. (1999b). Reply to Rushton: A gang of *g*s overpowers factor analysis. *Personality and Individual Differences, 26,* 391–393.

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

Frederickson, N. (1999). The ACID test—or is it? *Educational Psychology in Practice, 15,* 2–8.

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

Garfield, S. L. (1978). Research problems in childhood diagnosis. *Journal of Consulting and Clinical Psychology, 46,* 596–601.

Glutting, J. J., McDermott, P. A., & Konold, T. R. (1997). Ontology, structure, and diagnostic benefits of a normative subtest taxonomy from the WISC–III standardization sample. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 349–372). New York: Guilford.

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., Prifitera, A., & McGrath, E. A. (1994). Core profile types for the WISC–III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. *School Psychology Review, 23,* 619–639.

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., Youngstrom, E. A., Oakland, T., & Watkins, M. W. (1996). Situational specificity of generality of test behaviors for samples of normal and referred children. *School Psychology Review, 25,* 64–107.

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.

Good, R., Vollmer, M., Creek, R. J., Katz, L., & Chowdhri, S. (1993). Treatment utility of the Kaufman Assessment Battery for Children: Effects of matching instruction and student processing strength. *School Psychology Review, 22,* 8–26.

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.

Gustafsson, J., & Undheim, J. O. (1996). *Individual differences in cognitive functions.* In D. C. Berliner & R. C. Calfee (Eds.), *Handbook of educational psychology* (pp. 186–242). New York: Macmillan.

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.

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.

Hansen, J. C. (1999). *Test psychometrics*. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 15–30). Boston: Allyn and Bacon.

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

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

Hopkins, K. D. (1998). *Educational and psychological measurement and evaluation* (8th ed.). Needham Heights, MA: Allyn and Bacon.

Ivnik, R. J., Smith, G. E., Malec, J. F., Petersen, R. C., & Tangalos, E. G. (1995). Long-term stability and intercorrelations of cognitive abilities in older persons. *Psychological Assessment*, *7*, 155–161.

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

Kahana, S. Y., Youngstrom, E. A., & Glutting, J. J. (2002). Factor and subtest discrepancies on the Differential Abilities Scale: 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). New York: Academic Press.

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

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

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

Kaufman, A. S. (1994b). A reply to Macmann and Barnett: Lessons from the blind men and the elephant. *School Psychology Quarterly*, *9*, 199–207.

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

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.

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.

Kehle, T. J., Clark, E., & Jenson, W. R. (1993). The development of testing as applied to school psychology. *Journal of School Psychology*, *31*, 143–161.

Keith, T. Z., & Witta, E. L. (1997). Hierarchical and cross-age confirmatory factor analysis of the WISC–III: What does it measure? *School Psychology Quarterly*, *12*, 89–107.

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

Klassen, R. M., & Kishor, N. (1996). A comparative analysis of practitioners' errors on WISC–R and WISC–III. *Canadian Journal of School Psychology*, *12*, 35–43.

Klein, E. S., & Fisher, G. S. (1994, March). *The usefulness of the Wechsler deterioration index as a predictor of learning disabilities in children*. Paper presented at the meeting of the National Association of School Psychologists, Seattle, WA.

Kline, R. B., Snyder, J., & Castellanos, M. (1996). Lessons from the Kaufman Assessment Battery for Children (K–ABC): Towards a new cognitive assessment model. *Psychological Assessment*, *8*, 7–17.

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.

Konold, T. R., Glutting, J. J., McDermott, P. A., Kush, J. C., & Watkins, M. W. (1999). Structure and diagnostic benefits of a normative subtest taxonomy developed from the WISC–III standardization sample. *Journal of School Psychology*, *37*, 29–48.

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.

Krauskopf, C. J. (1991). Pattern analysis and statistical power. *Psychological Assessment*, *3*, 261–264.

Kubiszyn, T. W., Meyer, G. J., Finn, S. E., Eyde, L. D., Kay, G. G., Moreland, K. L., Dies, R. R., & Eisman, E. J. (2000). Empirical support for psychological assessment in clinical health care settings. *Professional Psychology: Research and Practice*, *31*, 119–130.

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.

Lawson, J. S., & Inglis, J. (1985). Learning disabilities and intelligence test results: A model based on a principal components analysis of the WISC–R. *British Journal of Psychology*, *76*, 35–48.

Lipsitz, J. D., Dworkin, R. H., & Erlenmeyer-Kimling, L. (1993). Wechsler comprehension and picture arrangement subtests and social adjustment. *Psychological Assessment, 5,* 430–437.

Livesay, J. (1986). Clinical utility of Wechsler's deterioration index in screening for behavioral impairment. *Perceptual and Motor Skills, 25,* 1191–1194.

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.

Lubinski, D. (2000). Scientific and social significance of assessing individual differences: "Sinking shafts at a few critical points." *Annual Review of Psychology, 51,* 405–444.

Lynam, D., Moffitt, T., & Stouthamer-Loeber, M. (1993). Explaining the relation between IQ and delinquency: Class, race, test motivation, school failure, or self-control? *Journal of Abnormal Psychology, 102,* 187–196.

Macmann, G. M., & Barnett, D. W. (1994). Some additional lessons from the Wechsler series: A rejoinder to Kaufman and Keith. *School Psychology Quarterly, 9,* 223–236.

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., & Prifitera, A. (1989). Subtest scatter and premorbid intelligence: Lessons from the WAIS–R standardization sample. *Psychological Assessment, 1,* 186–191.

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

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, A. R. (1992). Illusions 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., & Noonan, J. V. (1989). Typology and prevailing composition of core profiles in the WAIS–R standardization sample. *Psychological Assessment, 1,* 118–125.

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 and Clinical Science.

McDermott, P. A., Marston, N. C., & Stott, D. H. (1993). *Adjustment Scales for Children and Adolescents.* Philadelphia: Edumetric and 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., Keith, T. Z., Flanagan, D. P., & Vanderwood, M. (1997). Beyond *g*: The impact of Gf-Gc specific cognitive abilities research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review, 26,* 189–210.

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.

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.

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

Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical-statistical controversy. *Psychology, Public Policy, and Law, 2,* 293–323.

Meehl, P. E. (1997). Credentialed persons, credentialed knowledge. *Clinical Psychology: Science and Practice, 4,* 91–98.

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., Caspi, A., Harkness, A. R., & Silva, P. A. (1993). The natural history of change in intellectual performance: Who changes? How much? Is it meaningful? *Journal of Child Psychology and Psychiatry, 34,* 455–506.

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.

Mosel, J. N., & Roberts, J. B. (1954). The comparability of measures of profile similarity: An empirical study. *Journal of Consulting Psychology, 18,* 61–66.

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.

Neisser, U., Boodoo, G., Bouchard, T. J., Boykin, A. W., Brody, N., Ceci, S. J., Halpern, D. F., Loehlin, J. C., Perloff, R., Sternberg, R. J., & Urbina, S. (1996). Intelligence: Knowns and unknowns. *American Psychologist, 51,* 77–101.

Neuhaus, G., Foorman, B. R., Francis, D. J., & Carlson, C. D. (2001). Measures of information processing in rapid automatized naming (RAN) and their relation to reading. *Journal of Experimental Child Psychology, 78,* 359–373.

Nicholson, C. L., & Alcorn, C. L. (1994). *Educational applications of the WISC–III: A handbook of interpretive strategies and remedial recommendations.* Los Angeles: Western Psychological Services.

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., & Glutting, J. J. (1998). Assessment of test behaviors with the WISC–III. In A. Prifitera & D. H. Saklofske (Eds.), *WISC–III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 289–310). New York: Academic Press.

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.

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], 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.

Reinecke, M. A., Beebe, D. W., & Stein, M. A. (1999). The third factor of the WISC–III: It's (probably) not freedom from distractibility. *Journal of the American Academy of Child and Adolescent Psychiatry, 38,* 322–328.

Reschly, D. J. (1997). Diagnostic and treatment utility of intelligence tests. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 437–456). New York: Guilford.

Reschly, D. J., & Grimes, J. P. (1990). Best practices in intellectual assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology II* (pp. 425–440). Washington, DC: National Association of School Psychologists.

Riccio, C. A., Cohen, M. J., Hall, J., & Ross, C. M. (1997). The third and fourth factors of the WISC–III: What they don't measure. *Journal of Psychoeducational Assessment, 15,* 27–39.

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.

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

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

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.

Schretlen, D., Bobholz, J. H., & Benedict, R. H. B. (1992, August). *How useful is WAIS–R analysis for predicting psychiatric diagnoses?* Poster session presented at the annual convention of the American Psychological Association, Washington, DC.

Silva, P. A. (1990). The Dunedin Multidisciplinary Health and Development Study: A 15-year longitudinal study. *Pediatric and Perinatal Epidemiology, 4,* 96–127.

Silverstein, A. B. (1993). Type I, type II, and other types of errors in pattern analysis. *Psychological Assessment, 5,* 72–74.

Skrabanek, P., & McCormick, J. (1990). *Follies and fallacies in medicine.* Buffalo, NY: Prometheus Books.

Slate, J. R., Jones, C. H., Coulter, C., & Covert, T. L. (1992). Practitioners' administration and scoring of the WISC–R: Evidence that we do err. *Journal of School Psychology, 30,* 77–82.

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.

Strauss, E., Spreen, O., & Hunter, M. (2000). Implications of test revisions for research. *Psychological Assessment, 12,* 237–244.

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

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

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). New York: Academic Press.

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

Thorndike, R. L. (1994). [Review of the book *Clinical assessment of children's intelligence*]. *Journal of Psychoeducational Assessment, 12,* 172–179.

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

Tiholov, T. T., Zawallich, A., & Janzen, H. L. (1996). Diagnosis based on the WISC–III processing speed factor. *Canadian Journal of School Psychology, 12*, 23–34.

Tracey, T. J., & Rounds, J. (1999). *Inference and attribution errors in test interpretation*. In J. W. Lichtenberg & R. K. Goodyear (Eds.), *Scientist-practitioner perspectives on test interpretation* (pp. 113–131). Boston: Allyn and Bacon.

Truscott, S. D., & Frank, A. J. (2001). Does the Flynn effect affect IQ scores of students classified as LD? *Journal of School Psychology, 39*, 319–334.

Vargo, F. E., Grossner, G. S., & Spafford, C. S. (1995). Digit span and other WISC–R scores in the diagnosis of dyslexic children. *Perceptual and Motor Skills, 80*, 1219–1229.

Ward, S. B., Ward, T. B., 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, 12*, 267–276.

Ward, T. J., Ward, S. B., Glutting, J. J., & Hatt, C. V. (1999). Exceptional LD profile types for the WISC–III and WIAT. *School Psychology Review, 28*, 629–643.

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., & 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). Prevalence and diagnostic utility of the WISC–III SCAD profile among children with disabilities. *School Psychology Quarterly, 12*, 235–248.

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

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. (1967). *Wechsler Preschool and Primary Scale of Intelligence*. New York: Psychological Corporation.

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

Wechsler, D. (1981). *Wechsler Adult Intelligence Scale-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.

Weiner, I. B. (1989). On competence and ethicality in psychodiagnostic assessment. *Journal of Personality Assessment, 53*, 827–831.

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

Wong, W.-K., & Cornell, D. G. (1999). PIQ > VIQ discrepancy as a correlate of social problem solving and aggression in delinquent adolescent males. *Journal of Psychoeducational Assessment, 17*, 104–112.

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

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*. New York: 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.