15

Multifactored and Cross-Battery Ability Assessments: Are They Worth the Effort?

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

(2003).
in C.R. Reynolds & R. Kamphaus (Eds.),
Handbook of Psychological and Educational Assessment in Children.
New York: Guilford Press.

The practical merit of intelligence tests has been debated extensively. In the mid-1990s, as a consequence of the controversy surrounding Herrnstein and Murray's (1994) book *The Bell Curve: Intelligence and Class Structure in American Life*, the American Psychological Association formed a task force charged with developing a scientific report on the known meaning and efficacy of scores from tests of intelligence. The final report was published in the *American Psychologist* (Neisser et al., 1996). This account is especially intriguing in the context of a chapter examining the utility of multifactor ability assessments, because it offered no evidence that would support either the diagnostic or prescriptive relevance of subtest scores, factor scores, or other derived indices. Instead, IQ tests were defended solely on the basis of the more parsimonious construct coverage provided by global, or *g*-based, measures of intelligence (Neisser et al., 1996).

Psychologists today expend considerable effort administering and scoring the many subtests within most instruments that are intended to assess cognitive ability. Such an investment is presumably made in order to garner clinically useful information not available from interpretation of the single, global score. Accordingly, the trend among publishers of individually administered intelligence tests has been toward creating longer instruments that provide an ever-increasing diversity of discrete subtest scores and factor indices. A partial listing of some instruments introduced in recent years illustrates this trend.

Compared to the 10 mandatory (and 2 supplementary) subtests in the Wechsler Intelligence Scale for Children—Revised (WISC-R; Wechsler, 1974), the updated Wechsler Intelligence Scale for Children—Third Edition (WISC-III; Wechsler, 1991) is slightly longer, having added a new subtest. The Wechsler Adult Intelligence Scale—Third Edition (WAIS-III; Wechsler, 1997) contains 14 subtests versus the 11 subtests of its predecessor (Wechsler, 1981), a 27% increase in overall length. The Differential Ability Scales (DAS; Elliott, 1990) consists of 14 cognitive subtests. The revised Woodcock–Johnson Psycho-educational Battery (WJ-R; Woodcock & Johnson, 1989), including both ability and achievement subtests, allowed for the administration of 29 separate measures, while the new Woodcock–Johnson Psycho-educational Battery—

Third Edition (WJ-III; Woodcock & Johnson, 2000) includes 43 subtests!

Furthermore, in an attempt to capture all major components from Carroll's (1993) three-stratum model of intelligence, clinicians are now encouraged to move beyond the boundaries of specific, individually administered tests of intelligence. In their place, they are directed to employ multifactored, cross-battery assessments (Flanagan & McGrew, 1997; McGrew & Flanagan, 1998).

This chapter examines the relative efficacy of multifaceted abilities. The chapter is divided into five main sections. The first of these sections serves as a foundation; it establishes the amount and quality of validity evidence supporting the interpretation of global measures of intelligence. The second section reports on a series of empirical studies that assess continuing utility claims for the myriad specific abilities evaluated by subtest profiles. The third section moves away from subtest analysis and discusses research that has evaluated the validity of factor scores from individually administered tests of intelligence. The fourth section presents several troubling conceptual and practical issues associated with the interpretation of factor scores. The fifth section then scrutinizes evidence concerning the interpretation of scores (factor and subtest) obtained from cross-battery assessments.

WHAT *g* PREDICTS AND DEFINES

Validity Issues

As will be demonstrated shortly, utility is well established for global ability. This simple fact permits *g*-based estimates of intelligence to serve as a contrast for comparing the relative truth and value of multiple ability components. An even more important point—and one generally overlooked in the ability-testing literature—is that preference for interpretation should be given to *g*-based scores over other, more elaborate interpretative schemes. The reason is simply that global intelligence satisfies a foundational law of science: the law of parsimony. The common interpretation of parsimony is "Keep it simple." More formally, the law of parsimony (also known as "Occam's Razor") states that "what can be explained by fewer principles is explained needlessly by more" (Jones, 1952, p. 620). Because the number of subtest and factor scores interpreted during an ability assessment is usually large, it therefore becomes imperative that this added information offer practical, diagnostic, or treatment benefits for the individual being assessed, and that these benefits extend *above and beyond* the level of help afforded by interpretation of a single, *g*-based score (Brody, 1985; Reschly, 1997; Reschly & Grimes, 1990). Should the analysis of subtest or factor scores fail to fulfill these promises, their relevance becomes moot.

Treatment versus Predictive Validity

Multiple sources of evidence can be used to validate interpretations of test scores (Messick, 1989). However, in diagnostic assessment, two types of validity evidence are primary. Diagnostic, score-based interpretations become valid to the extent that they (1) are associated with a viable *treatment* for individuals experiencing a particular psychological problem/disorder, or (2) accurately *predict* (either concurrently or in the future) with a high probability that a given person will develop a problem/disorder (Cromwell, Blashfield, & Strauss, 1975; Glutting, McDermott, Konold, Snelbaker, & Watkins, 1998; Gough, 1971).

Psychologists have come to believe that treatment validity is the most important validity evidence for psychological tests, IQ and otherwise. This belief is unfortunate, because it occurs at the expense of prediction. Prediction is valuable in its own right, because we may never be able to remediate all of the negative circumstances that can influence a person's growth and well-being.

Presented below are several common outcomes *predicted* by *g*-based IQs. The presentation is representative of variables associated with general intelligence, but is not meant to be exhaustive. Such treatment is beyond the scope of this chapter, and readers are referred to the accompanying citations for more thorough discussions.

Scholastic Achievement

The substantial relationship between general intelligence and school achievement is

perhaps the best-documented finding in psychology and education (Brody, 1997; Neisser et al., 1996). Broadly speaking, *g*-based IQs correlate approximately .70 with standardized measures of scholastic achievement and .50 with grades in elementary school (Brody, 1985; Jensen, 1998). These correlations are somewhat higher than those obtained in the later school years; because of range restrictions, the correlations decrease progressively as individuals advance through the educational system. The typical correlation between *g* and standardized high school achievement is between .50 and .60; for college, coefficients vary between .40 and .50; for graduate school, correlations range between .30 and .40.

Jensen (1998) has indicated that *g*-based IQs predict academic achievement better than any other measurable variable. The reason he cites for the strong association is that school learning itself is *g*-demanding. Thorndike (1984) similarly concluded that 80%–90% of the *predicted* variance in scholastic performance is accounted for by *g*-based IQs, with only 10%–20% accounted for by *all* other scores in IQ tests. Thus the available evidence strongly suggests that global ability is the most important variable for estimating a person's academic achievement.

Years of Education

General intelligence is correlated with the number of years of a person's formal education and training. For instance, Jensen (1998) showed that, on average, years of education correlate .60 to .70 with *g*-based IQs. Jencks (1972) found longitudinal correlations above .50 between the IQs of preadolescents and the final grade level they completed. Likewise, in a review of 16 studies, Ceci (1991) reported correlations of .50 to .90 between measures of overall intelligence and an individual's years of education. Thus research results reveal a strong positive association between overall ability levels and years of education.

Job Training and Work Performance

Because of their strong correlation, there is much debate in the literature regarding whether intelligence or educational level is the variable more directly related to one's level of job performance (Ceci & Williams, 1997; Wagner, 1997; Williams & Ceci, 1997). Regardless of the interrelationship, some basic findings emerge. The average validity coefficient ranges between .20 and .30 for ability tests high in *g* and job performance (Hartigan & Wigdor, 1989). The coefficients rise to .50 when corrected for range restrictions and sources of measurement error (Hunter & Hunter, 1984; Ree & Earles, 1993).

Consequently, general ability provides surprisingly good prediction of job performance, and does so across a variety of occupations. Although the size of the correlations may not appear to be very high, the most impressive point to remember is that tests of general ability have a higher rate of predicting job performance than variables commonly employed to make such decisions, including class rank, grade point average, previous job experience, results from interviews, and performance on occupational interest inventories (Jensen, 1998).

Social Correlates

Global intelligence shows significant, but more moderate, criterion validity for personality and social dispositions. Typically, the independent contribution of IQ to any given social variable is small (a correlation of approximately .20; Glutting, Youngstrom, Oakland, & Watkins, 1996). At the same time, even such small correlations can have a striking impact in certain segments of the ability continuum. For example, adolescents with IQs of 90 and lower are more likely to have conduct disorder and to be arrested for juvenile delinquency than those with average or better IQs (Kazdin, 1995; Moffitt, Gabrielli, Mednick, & Schulsinger, 1981). Similarly, individuals with IQs of 80 or below experience an increased incidence of various social misfortunes, such as becoming disabled on the job or divorcing within the first 5 years of marriage (Jensen, 1998).

Summary of *g*-Based Interpretations

There is a tendency among some professionals to dismiss global intelligence as having

mere historical value, and thereafter to tout the merits of viewing intelligence as a multidifferentiated construct. However, an extensive body of empirical evidence demonstrates the practical, prognostic utility of *g*-based IQs. This literature supports the notion that the *g*-based IQ is among the most dominant and enduring of influences associated with many consequential outcomes within our culture. To those who would dismiss the import of global ability because it does not also serve to remedy what it predicts, we would urge that the inherent value of predictors be appreciated. There are countless predictors of life's vicissitudes, including predictors of the weather, of accident risk, of AIDS infection, and of future achievements. We would hate to see them all ignored because they fail to fix what they forecast.

INTERPRETATION OF COGNITIVE SUBTESTS

Reliance upon subtests to hypothesize about children's cognitive strengths and weaknesses is endemic in psychological training and practice (Aiken, 1996; Alfonso, Oakland, LaRocca, & Spanakos, 2000; Blumberg, 1995; Bracken, McCallum, & Crain, 1993; Gregory, 1999; Groth-Marnat, 1997; Kamphaus, 1993; Kaufman, 1994; Kaufman & Lichtenberger, 2000; Kellerman & Burry, 1997; Prifitera, Weiss, & Saklofske, 1998; Sattler, 1992; Truch, 1993). Interpretation of individual subtests is a vestigial practice, but recommendations that rely essentially upon one or two subtests can still be found (Banas, 1993). This is especially true for neuropsychological assessment (Lezak, 1995). More commonly, however, interpretation of individual subtests is eschewed (Kamphaus, 1993). For example, Kaufman and Lichtenberger (2000) concluded that "the key to accurately characterizing a child's strong and weak areas of functioning is to examine his or her performance across several subtests, not individual subtest scores in isolation" (p. 81). In support of Kaufman and Lichtenberger's conclusion, Table 15.1 illustrates that only 3 of the 12 WISC-III subtests contributing to the WISC-III's four factors meet the reliability coefficient criterion of $\geq .85$ recommended by Hansen (1999) for making decisions about individuals, and that none meet the more stringent criterion of $\geq .90$ (Hopkins, 1998; Salvia & Ysseldyke, 1998). Furthermore, the increased error generated by the use of

TABLE 15.1. Reliability of the WISC-III

Subtest or index	Internal consistency[a]	Short-term test–retest[a]	Long-term test–retest[b]
Information	.84	.85	.73
Similarities	.81	.81	.68
Arithmetic	.78	.74	.67
Vocabulary	.87	.89	.75
Comprehension	.77	.73	.68
Digit Span	.85	.73	.65
Picture Completion	.77	.81	.66
Coding	.79	.77	.63
Picture Arrangement	.76	.64	.68
Block Design	.87	.77	.78
Object Assembly	.69	.66	.68
Symbol Search	.76	.74	.55
Verbal IQ	.95	.94	.87
Performance IQ	.91	.87	.87
Verbal Comprehension	.94	.93	.85
Perceptual Organization	.90	.87	.87
Freedom from Distractibility	.87	.82	.75
Processing Speed	.85	.84	.62
Full Scale IQ	.96	.94	.91

[a]Data from Wechsler (1991).
[b]Data from Canivez and Watkins (1998).

difference scores makes even the best subtest-to-subtest comparison unreliable (e.g., the reliability of the difference between Block Design and Vocabulary is .76).

Elaborate interpretative systems (Kaufman, 1994; Kamphaus, 1993; Sattler, 1992) have been developed to identify specific cognitive subtest patterns that are assumed to reflect neurological dysfunction (Arizona Department of Education, 1992; Drebing, Satz, Van Gorp, Chervinsky, & Uchiyama, 1994; Ivnik, Smith, Malec, Kokmen, & Tangalos, 1994), to be related to learning disabilities (LDs) (Banas, 1993; Kellerman & Burry, 1997; Mayes, Calhoun, & Crowell, 1998; McLean, Reynolds, & Kaufman, 1990), and/or to be prognostic of emotional and behavioral impairments (Blumberg, 1995; Campbell & McCord, 1999). In fact, more than 75 patterns of subtest variation have been identified for the Wechsler scales alone (McDermott, Fantuzzo, & Glutting, 1990).

Diagnostic Efficiency Statistics

Identification of pathognomonic cognitive 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 (e.g., LDs) is compared to the mean subtest score of a group of children without the problem. Statistically significant subtest score differences between the two groups are subsequently interpreted as evidence that the profile is diagnostically effective.

However, mean-score difference methods are inadequate to reach this conclusion. Almost 50 years ago, Meehl and Rosen (1955) made it clear that efficient diagnosis depends on the psychometric instruments employed *and* on a consideration of base rates (i.e., prevalence) of the criterion condition in both nondisabled and clinical populations. More recently, Elwood (1993) asserted that "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; emphasis in original). As outlined in Table 15.2, Kessel and Zimmerman (1993) listed several diagnostic efficiency indices that allow a test's accuracy to be analyzed in relation to two pervasive alternative interpretations: base rate and chance (Cohen, 1990; Rosnow & Rosenthal, 1989).

An extension of the diagnostic efficiency statistics in Table 15.2 was originally developed in engineering as a way to tell how well a radar operator is able to distinguish signal from noise (Hanley & McNeil,

TABLE 15.2. Diagnostic Efficiency Statistics

Statistic	Description
Sensitivity	True-positive rate. Proportion of participants with a disorder who are identified by a positive test result.
Specificity	True-negative rate. Proportion of participants free of a disorder who are correctly identified by a negative test result.
Positive predictive power	Proportion of participants identified by a positive test result who truly have the target disorder.
Negative predictive power	Proportion of participants identified by a negative test result who truly do *not* have the target disorder.
False-positive rate	Proportion of participants identified by a positive test result who truly do *not* have the target disorder.
False-negative rate	Proportion of participants identified by a negative test result who truly have the target disorder.
Hit rate	Proportion of participants with *and* without the target disorder who were correctly classified by the test.
Kappa	Proportion of agreement between the test and actual condition of the participants (disordered vs. nondisordered) beyond that accounted for by chance alone.

1982). The methodology was then adapted and reformulated for biostatistical applications (Kraemer, 1988; Murphy et al., 1987; Swets, 1988), and it was recently recommended for use with psychological assessment data (McFall & Treat, 1999). Designated the "receiver operating characteristic" (ROC), this procedure entails plotting the balance between the sensitivity and specificity of a diagnostic test while systematically moving the cut score across its full range of values. As illustrated in Figure 16.1, the diagonal dashed line is the "random ROC," which reflects a test with zero discriminating power. The more clearly a test is able to discriminate between individuals with and without the target disorder, the farther its ROC curve will deviate toward the upper left corner of the graph.

The accuracy of an ROC can be quantified by calculating the area under its curve (AUC). Chance diagnostic performance corresponds to an AUC of .50, whereas perfect diagnostic performance equates to 1.00. The AUC is independent of cut score and does not assume that the underlying score distributions are normal. It is interpreted in terms of two children: one drawn randomly from the distribution of children with the target disorder, and one selected randomly from the population of children without the problem. The AUC is the probability of the test's correctly rank-ordering the children into their appropriate diagnostic groups. According to Swets (1988), AUCs between .50 and .70 are characterized as showing low accuracy; those between .70 and .90 represent medium accuracy; and those between .90 and 1.00 denote high accuracy. Diagnostic utility statistics, including the ROC and its AUC, should be applied when subtest profiles are hypothesized as being able to distinguish between children with and without disorders.

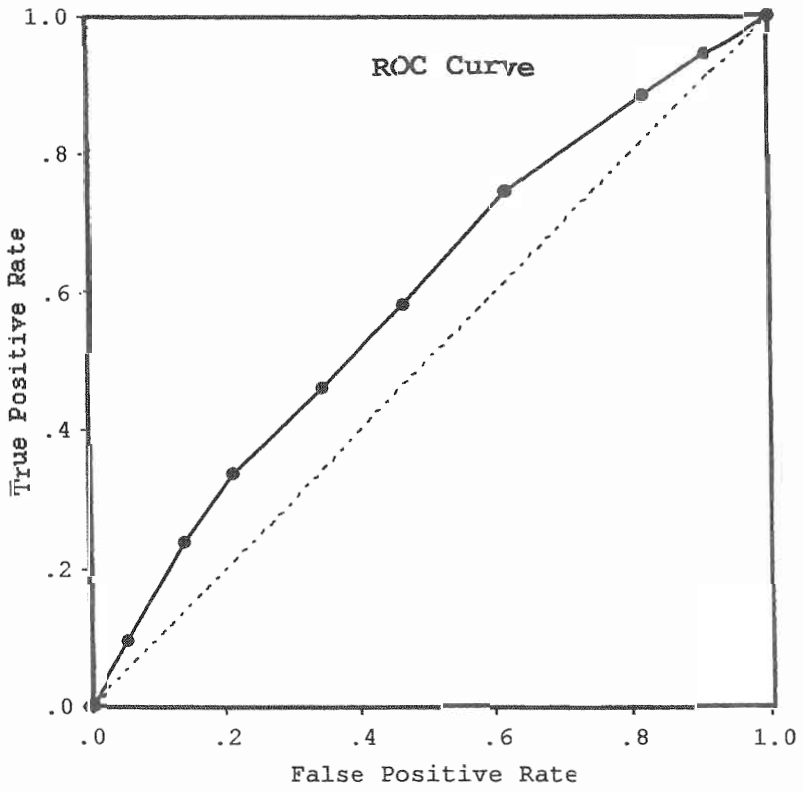

FIGURE 15.1. Receiver operating characteristic (ROC) of Wechsler Development Index (WDI), used to distinguish between participants with and without learning disabilities.

Diagnosis of Neurological Dysfunction

Wechsler's (1958) Deterioration Index (WDI) was originally developed as an indicator of cognitive impairment that was hypothesized to be sensitive to brain injury in adults. Conceptually, the WDI was composed of two groups of Wechsler subtest scores: (1) "hold" subtests, which were considered insensitive to brain injury (Vocabulary, Information, Object Assembly, and Picture Completion); and (2) "don't hold" subtests, which were judged vulnerable to intellectual decline (Digit Span, Similarities, Coding, and Block Design).

Application of the WDI with children was suggested by Bowers and colleagues (1992), given that neuropsychological deficits have often been hypothesized to account for LDs and attentional difficulties (Accardo & Whitman, 1991; Goodyear & Hynd, 1992). Bowers and colleagues recommended that the WDI be renamed the Wechsler *Developmental* Index, because children's cognitive skills are not deteriorating; rather, they are assumed to be developing unevenly. Klein and Fisher (1994) applied the WDI to children in LD programs and found that they scored significantly higher on the WDI (i.e., showed more problems) than children in regular education programs. Based on these statistically significant group differences, Klein and Fisher concluded that the WDI is useful for predicting which students would be found eligible for LD services.

However, mean-difference statistics cannot be used to justify this conclusion. Watkins (1996) replicated the Klein and Fisher (1994) study, but also applied more appropriate diagnostic efficiency procedures. Results revealed that the WDI performed at near-chance levels when distinguishing students diagnosed with LDs (n = 611) from those diagnosed with emotional disabilities (n = 80) or mental retardation (n = 33), as well as from randomly simulated, normal cases (n = 2,200). Based upon formulas provided by Hsiao, Bartko, and Potter (1989), the AUC for this study (see Figure 15.1) summed to .57 (compared with a chance rate of .50 for AUCs and a low accuracy rate of between .50 and .70). It was concluded that mean group differences were insufficient to determine diagnostic efficacy, and that the WDI must be definitively validated before it can be applied in actual practice.

Diagnosis of LDs

The ACID Profile

Several subtest profiles have long, storied histories in the field of psychodiagnosis. The most venerable is the "ACID" profile, characterized by low scores on Wechsler Arithmetic, Coding, Information, and Digit Span subtests. With development of the most recent revision of the Wechsler, the WISC-III, diagnostic merit of the ACID profile has once again been advanced (Groth-Marnat, 1997). Prifitera and Dersh (1993) compared percentages of children showing WISC-III ACID profiles in samples with LDs and attention-deficit/hyperactivity disorder (ADHD) to percentages showing the ACID profile in the WISC-III standardization sample. Their findings uncovered a greater incidence of ACID profiles in the clinical samples, with approximately 5% of the children with LDs and 12% of the children with ADHD showing the ACID profile, while such a configuration occurred in only 1% of the cases from the WISC-III standardization sample. Based upon this data, Prifitera and Dersh concluded that ACID profiles "are useful for diagnostic purposes" because "the presence of a pattern or patterns would suggest strongly that the disorder is present" (pp. 50–51). Ward, Ward, Hatt, Young, and Mollner (1995) investigated the prevalence of the WISC-III ACID profile among children with LDs (n = 382) and found a prevalence rate of 4.7% (vs. the expected rate of 1%). Likewise, upon achieving similar ACID results for a sample of children with LDs (n = 165), Daley and Nagle (1996) suggested practitioners that "investigate the possibility of a learning disability" (p. 330) when confronted by an ACID profile.

Watkins, Kush, and Glutting (1997a) evaluated the discriminative and predictive validity of the WISC-III ACID profile among children with LDs. As in previous research (Kaufman, 1994), ACID profiles were more prevalent among children with

LDs (n = 612) than among children without LDs (n = 2,158). However, when ACID profiles were used to classify students into groups with and without LDs, they operated with considerable error. At best, only 51% of the children identified by a positive ACID profile were previously diagnosed as having LDs. These data indicated that a randomly selected child with an LD had a more severe ACID profile than a randomly selected child without an LD about 60% of the time (AUC = .60). Although marginally better than chance, the degree of accuracy was quite low (cf. classificatory criteria presented by Swets, 1988).

The SCAD Profile

Preliminary empirical support was provided by Prifitera and Dersh (1993) for another subtest configuration hypothesized to be indicative of LDs. They combined subtests from the WISC-III Freedom from Distractibility and Processing Speed factors to create a new profile. This profile was more common in a sample of children with LDs (n = 99) and in another sample of children with ADHD (n = 65) than within the WISC-III standardization sample. Using the outcomes as guidance, Prifitera and Dersh suggested that the subtest configuration would be "useful in the diagnosis of LD and ADHD" (p. 53).

Kaufman (1994) coined an acronym for this new profile: "SCAD" (for the Symbol Search, Coding, Arithmetic, and Digit Span subtests). He recommended that the SCAD index be subtracted from the sum of Picture Completion, Picture Arrangement, Block Design, and Object Assembly to create a comparison between SCAD and the Perceptual Organization factor. Kaufman opined that Arithmetic, Coding, and Digit Span have "been quite effective at identifying exceptional groups from normal ones, and . . . are like a land mine that explodes on a diversity of abnormal populations but leaves most normal samples unscathed" (p. 213). Kaufman concluded that the SCAD profile is "an important piece of evidence for diagnosing a possible abnormality" (p. 221), which "won't identify the type of exceptionality, but [the profile is] likely to be valuable for making a presence–absence decision and helping to pinpoint specific areas of deficiency" (p. 214).

The foregoing claims were tested by Watkins, Kush, and Glutting (1997b) with children who were enrolled in LD and emotional disability programs (n = 365). When these children were compared to the WISC-III standardization sample via diagnostic utility statistics, an AUC of .59 was generated. This finding suggests that the SCAD profile is not substantially more useful in making this diagnostic decision than any randomly chosen, irrelevant variable (McFall & Treat, 1999). Thus, contrary to Kaufman's (1994) assertion, SCAD subtest scores were not found to be important evidence for diagnosing exceptionalities.

Subtest Variability

Heterogeneous variability among subtest scores is a traditional diagnostic indicator of LDs. Subtest variability can be quantified in three ways (Schinka, Vanderploeg, & Curtiss, 1997). The first method examines the range (i.e., difference between an examinee's highest and lowest subtest scaled scores). The second method involves evaluating variances, using the variance formula applicable to the subtest scores of an individual examinee. Finally, researchers look at the number of subtests differing from the individual examinee's mean score by ±3 points.

The diagnostic utility of all three variability metrics was tested by Watkins (1999) and Watkins and Worrell (2000). Children from the WISC-III standardization effort were compared to children enrolled in LD programs (n = 684). Results included AUCs ranging from .50 to .54. Thus WISC-III subtest variability exhibited low diagnostic utility in distinguishing children with LDs from those without identified problems from the WISC-III standardization sample.

Diagnosis of Emotional and Behavioral Disorders

Despite long-standing assumptions, subtest profiles have consistently failed to demonstrate utility in predicting students' social and behavioral functioning (Beebe, Pfiffner, & McBurnett, 2000; Dumont, Farr, Willis, & Whelley, 1998; Glutting et al., 1998;

Glutting, McGrath, Kamphaus, & McDermott, 1992; Kramer, Henning-Stout, Ullman, & Schellenberg, 1987; Lipsitz, Dworkin, & Erlenmeyer-Kimling, 1993; McDermott & Glutting, 1997; Piedmont, Sokolove, & Fleming, 1989; Reinecke, Beebe, & Stein, 1999; Riccio, Cohen, Hall, & Ross, 1997; Rispens et al., 1997) and have been discounted as valid indicators of children's mental health. Thus Teeter and Korducki (1998) concluded that "in general there appears to be a consensus in the literature that there are no distinctive Wechsler [subtest] patterns that can provide reliable, discriminative information about a child's behavior or emotional condition" (p. 124). In contrast, instruments designed specifically to assess child behavior, such as teacher- and parent-completed rating scales, have produced highly accurate differential diagnoses (i.e., AUCs > .90; Chen, Faraone, Biederman, & Tsuang, 1994).

Hypothesis Generation

Although cognitive subtest profiles are not accurate in diagnosing childhood psychopathology, profile interpretation is frequently relied upon to identify distinctive abilities useful for hypothesis generation (Gregory, 1999). This practice implicitly assumes that cognitive subtest profiles are predictive of performance in important endeavors, such as children's academic achievement and/or their classroom conduct. For example, Kaufman (1994) asserted that "insightful subtest interpretation" (p. 32) allows an examiner to understand why a student experiences learning difficulties and how to remediate them.

As illustrated earlier in this chapter, global intelligence has a well-documented, robust relationship with academic achievement. However, the excellent predictive validity of the *g*-based IQ cannot be assumed to generalize to subtest profiles. One way to test the utility and validity of subtest scores is to decompose profiles into their elemental components. The unique, incremental predictive validity of each component can then be analyzed separately to determine what aspect(s), if any, of the subtest profile can be used to estimate academic performance.

To this end, Cronbach and Gleser (1953) reported that subtest profiles contain only three types of information: "elevation," "scatter," and "shape." Elevation information is represented by a person's aggregate performance (i.e., mean, normative score) across subtests. Profile scatter is defined by how widely scores in a profile diverge from its mean; scatter is typically operationalized by the standard deviation of the subtest scores in a profile. Finally, shape information reflects where "ups and downs" occur in a profile. Even if two profiles have the same elevation and scatter, their high and low points may be different. Shape is thus defined by the rank order of scores for each person (Nunnally & Bernstein, 1994).

Watkins and Glutting (2000) tested the incremental validity of WISC-III subtest profile level, scatter, and shape in forecasting academic performance. WISC-III subtest profiles were decomposed into the three elements just described and sequentially regressed onto reading and mathematics achievement scores for nonexceptional (n = 1,118) and exceptional (n = 538) children. Profile elevation was statistically and practically significant for both nonexceptional (R = .72 to .75) and exceptional (R = .36 to .61) children. Profile scatter did not aid in the prediction of achievement. Profile shape accounted for an additional 5%–8% of the variance in achievement measures: One pattern of relatively high verbal scores positively predicted both reading and mathematics achievement, and a pattern of relatively low scores on the WISC-III Arithmetic subtest was negatively related to mathematics. Beyond these two somewhat intuitive patterns, profile shape information had inconsequential incremental validity for both nonexceptional and exceptional children. In other words, it was the averaged, norm-referenced information (i.e., elevation) contained in subtest profiles that best predicted achievement. This information is essentially redundant to the prognostic efficacy available from omnibus intelligence scores (i.e., Verbal IQ [VIQ], Performance IQ [PIQ]) and global ability (i.e., the Full Scale IQ [FSIQ]) and is consistent with outcomes obtained in previous studies (Glutting et al., 1992, 1998; Hale & Saxe, 1983; Kline, Snyder, Guilmette, & Castellanos, 1993). From these findings, it was concluded that subtest

scatter and shape offer minimal assistance for generating hypotheses about children's academic performance.

Methodological Issues

Subtest analysis has also undergone serious methodological challenges. Specifically, within the last 15 years several methodological problems have been identified that operate to negate, or equivocate, essentially all research into children's ability profiles (Glutting et al., 1998; McDermott et al., 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992; Watkins & Kush, 1994).

Circular Reasoning and Selection Bias

Prominent among the methodological problems identified is the circular use of ability profiles for *both* the initial formation of diagnostic groups *and* the subsequent search for profiles that might inherently define or distinguish those groups. This problem is one of self-selection. The consequence is that self-selection unduly increases the probability of discovering group differences. Another factor affecting outcomes is the nearly exclusive use of children previously classified or those referred for psychoeducational assessments. Both classified and referral samples (the majority of whom are subsequently classified) are unrepresentative of the population as a whole and subject to selection bias (Rutter, 1989).

Solutions to Methodological Problems

It is possible to overcome the problems of circular reasoning and selection bias (Glutting, McDermott, Watkins, Kush, & Konold, 1997; Glutting et al., 1998; Sines, 1966; Wiggins, 1973). Three steps are necessary. First, rather than concentrating exclusively on exceptional or referral samples, researchers should use epidemiological samples from the general population (i.e., large, unselected cohorts), because such samples are representative of the child population as a whole. Second, the epidemiological samples should be further divided on the basis of their score configurations, rather than according to whether children fit predetermined diagnostic categories (e.g., "children with LDs," "normal children," and the like). In other words, the epidemiological sample should be used to identify groups with unusual versus common ability score profiles. The identification of "unusual" profiles can be accomplished with a variety of methods. Examples include the traditional approaches of whether or not statistically significant normative or ipsative score differences are present. Alternatively, more current univariate–normative and univariate–ipsative base rate approaches could be used (e.g., a prevalence/base rate occurring in less than 5% of the child population), as well as multivariate prevalence approaches (cf. Glutting et al., 1998). Third, once classified on observed score configurations (e.g., groups with unusual vs. common ability profiles), the groups should subsequently be compared across a variety of important criteria, external to the ability test itself.

When we took *all* of the methodological factors described above into account, we were able to locate only a single investigation within the last 15 years that supported the interpretation of subtest scores (Prifitera & Dersch, 1993). By contrast, a substantial number of studies satisfying the dual criteria failed to find relationships between unusual subtest configurations on such tests as the WISC-III, the DAS, and the Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983) and performance on meaningful external criteria (Glutting et al., 1992, 1998; Glutting, McDermott, et al., 1997; McDermott et al., 1990, 1992; McDermott & Glutting, 1997; Watkins & Glutting, 2000; Watkins et al., 1997b).

Ipsative Assessment

Evidence reported in previous sections of this chapter suggests that subtest profiles are invalid indicators of childhood psychopathology, and that most of the predictive power carried by subtests resides in their level (i.e., their role as a vehicle of *g*) rather than in their shape or scatter. One psychometric source of this invalidity was described by McDermott and colleagues (1990, 1992) and later by McDermott and Glutting (1997). In essence, the operationalization of a large number of current interpretative systems moves away from norma-

tive measurement, and instead rests upon ipsative interpretation of test scores (Cattell, 1944). As described by Kaufman (1994), ipsative measurement is concerned with how a child's subtest scores relate to his or her personalized, average performance and discounts the influence of global intelligence. Thus Kaufman suggested that "it is of greater interest and potential benefit to know what children can do well, relative to their own level of ability, than to know how well they did [normatively]" (p. 8).

Ipsative measurement is operationalized by taking an individual's subtest scores and averaging them. Then each subtest score is subtracted from the child's personal grand mean. Subtest scores that deviate negatively from the personalized mean are considered to reflect cognitive weaknesses, and those that deviate positively are assumed to represent cognitive strengths. As Silverstein (1993) has cautioned, however, these repeated subtest versus grand-mean comparisons entail repeated statistical comparisons that produce excessive Type I and Type II error rates, for which there is no satisfactory solution.

More importantly, after employing a large number of statistical techniques across multiple samples (both large epidemiological samples and cohorts of exceptional children), McDermott and colleagues (1990, 1992) and McDermott and Glutting (1997) concluded:

> Ipsative measures have insufficient reliability for educational decisions, are significantly less reliable than normative measures, and are relatively insensitive to sources of individual variation that characterize omnibus ability measures. Further, any argument in favor of ipsatized assessment certainly is vitiated by the fact that such approaches fail to predict outcomes as well as normative approaches. And, were all of this not the case, we would still be left with uncertainty about the meaning of ipsative constructs and their limited utility for either group or individual studies. (McDermott et al., 1992, p. 521)

The preceding findings prompted Silverstein to observe that "the assumption of clinical meaningfulness [of subtest deviations around a personalized mean] may ultimately prove to be the fundamental error in pattern analysis" (1993, p. 73).

INTERPRETATION OF FACTOR SCORES FROM INDIVIDUALLY ADMINISTERED TESTS OF INTELLIGENCE

Factor scores are stronger candidates for interpretation than subtest profiles. Factor scores have better reliabilities than subtest scores (as per the Spearman–Brown prophecy; Traub, 1991), as illustrated in Table 15.1. And, because they theoretically represent phenomena beyond the sum of subtest specificity and measurement error, factor scores potentially escape the psychometric weaknesses that undermine analyses of conjoint subtest patterns. Factor score interpretation is also consistent with standards for good assessment practice, such as the "top-down" hierarchical approach recommended by authorities on intelligence testing (cf. Kamphaus, 1993; Kaufman, 1994; Sattler, 1992). Therefore, it is possible that ability constructs measured by factor deviation quotients (i.e., factor score IQs) from such tests as the WISC-III and DAS might show strong associations with important achievement, emotional, and/or behavioral criteria.

Criterion-Related Validity

Despite their psychometric advantages, the utility of factor scores has not been well researched. A major issue that remains is to demonstrate the validity of factor scores—and, specifically, to determine whether factor scores provide substantial improvements in predicting important criteria above and beyond levels afforded by general ability. Glutting, Youngstrom, Ward, Ward, and Hale (1997) assessed the ability of the four factors underlying the WISC-III (Verbal Comprehension, Perceptual Organization, Freedom from Distractibility, and Processing Speed), relative to the FSIQ, to predict performance in four areas of achievement (reading, mathematics, language, and writing). Two groups were examined: a nationally representative epidemiological sample (n = 283) and a sample of children referred for psychoeducational assessments (n = 636). In general, the four factor scores did not show any substantial increase in the prediction of achievement criteria after the FSIQ was partialed out. The Freedom from Distractibility factor showed the largest cor-

relations after the FSIQ was controlled for, but it only uniquely accounted for between 1.4% and 5.2% of the variance in the various achievement measures. Results showed that the FSIQ was the most parsimonious and powerful predictor of academic achievement obtainable from the WISC-III. Using factor scores to estimate achievement levels, even in specific content areas, led to more complex models (and more laborious calculations for the practitioner) that provided either no or meager dividends. This relationship held true for both nonreferred and referred samples.

The research described above addressed only the inability of factor scores from the WISC-III to inform academic achievement. It is possible that factor scores from other IQ tests might yet tell something relevant about children's academic performance. Youngstrom, Kogos, and Glutting (1999) examined this issue. The incremental validity of the DAS's three factors (Verbal, Nonverbal, and Spatial Ability), relative to the test's General Conceptual Ability (GCA) score, was investigated in terms of predicting standardized achievement in three areas (word reading, basic number skills, and spelling). Results with an epidemiological sample (n = 1,185) showed that even when factor scores provided a statistically significant increment above the GCA score, the improvement was too small to be of clinical significance. Consequently, the outcomes extended prior findings with the WISC-III: that the more differentiated ability estimates provided by factor scores has not yet been found to better predict achievement than *g*.

Reversing the Hierarchical Order of Predictors

It could be argued that it is inappropriate to partial global ability (the FSIQ or GCA) prior to letting the ability factors predict achievement. In other words, in the two aforementioned studies, the hierarchical strategy should have been reversed (i.e., partialing the effect of the factor scores and then letting the FSIQ or GCA predict achievement). This strategy has some intuitive appeal. However, as noted at the beginning of this chapter, to sustain such logic, psychologists would have to repeal the law of parsimony. We would have to accept the novel notion that when many things essentially account for no more, or only marginally more, predictive variance in academic achievement than that accounted for by merely one thing (global ability), we should adopt the less parsimonious system.

Obviously, in the preceding analyses, there was a high degree of multicollinearity (i.e., redundancy) among the predictors as a consequence of global ability's being drawn in large part (but not entirely) from the underlying factor scores. However, in situations where variables are all highly interrelated, more things (such as factor scores) will nearly always predict as well as, or even *marginally* better than, one thing (global ability)—but that is exactly why such multicollinearity is a violation of parsimony and not a virtue. Therefore, it is incumbent among advocates of factor score interpretation to present convincing empirical support in their favor—support that clearly extends above and beyond the contribution provided by the parsimonious *g* variable.

Validity of Processing Speed Factors

The role and function of specific factor scores have also been investigated. One such factor is processing speed. The construct of "processing speed" has received considerable scholarly attention through the information-processing theories of cognitive psychology (see Kranzler, 1997, for a review). Likewise, the discovery of a processing speed factor on the DAS (Keith, 1990) and the inclusion of a processing speed factor on the WISC-III make it likely that clinicians will interpret this dimension during routine clinical assessments. Oh and Glutting (1999) investigated the utility of processing speed factors from the DAS and WISC-III, respectively. An epidemiological sample was employed. From the cohort, groups with unusual strengths and weaknesses in processing speed were identified according to a rarity criterion (i.e., the strengths or weaknesses occurred in $\leq$5% of the child population). The group with these strengths and weaknesses were then matched to a control group on the demographic variables of race, gender, and parents' educational levels, as well as on overall ability level. The group and its control were compared across multi-

ple, norm-referenced measures of achievement (and, in the DAS study, also across six teacher-rated indices of behavioral adjustment). In both studies, children with unusual strengths and weaknesses in processing speed were found to exhibit no significant differences in achievement or classroom adjustment from their respective controls. Consequently, these results suggested that measures of processing speed provide psychologists with no diagnostic help.

Factor Scores versus *g*

All of the foregoing results should come as no surprise. Kranzler (1997) summarized the evidence on general versus specific factors by noting:

> On IQ tests with at least several subtests measuring different abilities, *g* constitutes by far the single largest independent component of variance (e.g., Jensen, 1980). In fact, psychometric *g* usually explains more variance than all group factors *combined*. . . . Furthermore, the predictive validity of tests in education and vocational settings is overwhelmingly a function of *g*. (p. 152; emphasis in original)

Lubinski and Benbow (2000) concurred that general intelligence is the most potent predictor of academic performance for students in grades K–12, and attributed this ubiquitous finding to the fact that the K–12 educational curriculum is relatively uniform for most students. They hypothesized that specific mathematical, spatial, and verbal reasoning factors should become more important predictors of educational–vocational criteria as people begin to pursue more specialized educational and vocational training in young adulthood. According to their theory, this cognitive differentiation should be most apparent for students with high ability. Jensen (1998) also suggested that abilities are more differentiated at the upper end of the intelligence range, and supplied the analogy that rich people spend their money on a greater variety of things than do poor people. It would be worthwhile to test this hypothesis with jointly standardized cognitive and achievement tests that span the broad age and cognitive ability ranges specified by Lubinski and Benbow (2000).

Results from Structural Equation Modeling

Several authors recently contested the overwhelming research evidence in favor of general ability and suggested that specific factors have important effects beyond *g* (McGrew, Keith, Flanagan, & Underwood, 1997). Evidence presented to support these claims was based upon complex structural equation modeling (SEM) applied to the WJ-R (Keith, 1999). Researchers' conclusions seem to have moved from scientific caution to clinical certainty in only 2 years and two studies. For example, in the first study, McGrew and colleagues (1997) indicated that "the current results only suggest that *some* specific *Gf-Gc* abilities may be important for understanding *some* academic skills at *some* developmental levels" (p. 205, emphasis in original); by 1999, however, Keith concluded that "psychologists and educators who wish to understand students' reading and mathematics learning will gain more complete understanding of those skills for groups and individuals via the assessment of these specific abilities" (p. 257).

The conclusions must be tempered by several considerations. First, all of the studies cited just above used the WJ-R to formulate both general and specific intellectual factors. Flanagan, McGrew, and Ortiz (2000) report an unpublished study that apparently included the WISC-R and another that applied the WISC-III, but provided insufficient information to permit the evaluation of methods and results. Beyond the unknown generalizability to other intelligence tests, there is some danger that WJ-R cognitive and academic scales are confounded. In terms of generalizability, the WJ-R processing speed factor was related to math achievement (Keith, 1999; McGrew et al., 1997); however, the WISC-III and DAS processing speed factors, as noted earlier, demonstrated little incremental validity in predicting achievement and behavior (Glutting, Youngstrom, et al., 1997; Oh & Glutting, 1999; Youngstrom et al., 1999; Youngstrom & Glutting, 2000). When considering shared method variance, Keith (1999) and McGrew and colleagues (1997) both reported that the WJ-R Auditory Processing factor (*Ga*) was related to the WJ-R Letter–Word Identification and Word Attack subtests. The WJ-R *Ga* factor is com-

posed of two subtests: Incomplete Words and Sound Blending. McGrew and colleagues equated this auditory processing factor to "phonological awareness (*Ga*) in reading" (p. 196). However, phonological awareness is usually considered to be an important component of reading itself (Adams, 1990; Stahl & Murray, 1994), is often included as a skill in the reading curriculum (Carnine, Silbert, & Kameenui, 1997), and can be developed through instruction with subsequent enhancement of children's' reading skills (Bus & van Uzendoorn, 1999; Ehri et al., 2001). Thus the "cognitive" subtests appear to be inexorably confounded with their contrasting "academic" subtests.

Second, the aforementioned studies based their conclusions solely on SEM, which is a multivariate correlational technique designed to identify relationships among *latent* variables (i.e., constructs). Thus the methodology provides results that are best interpreted as relationships between pure constructs measured without error. SEM is, of course, an excellent method for testing theory, but it can be less than satisfactory for direct diagnostic applications. The observed test scores employed by psychologists are not latent variables, and they clearly contain measurement error (i.e., reliability coefficients less than 1.00). Basing diagnostic decisions on theoretically pure constructs is impossible in practice. Even approximating true scores would require clinicians to perform complex, tedious calculations for which no published algorithms yet exist. For example, attempting to employ SEM to describe the association between a cognitive ability and achievement would demand both (1) a known, quantifiable relationship between the measured variables and the latent variable, and (2) a way of correcting the individual's scores on predictor and criterion to approximate the true scores. In practical terms, this would involve using the factor loadings from the measurement model as regression coefficients to predict the individual's factor score based on the observed subtest scores. Not only is this more complicated than current practice, but the estimated factor loadings will change depending on the reference sample (unless it is a large, representative, epidemiological sample) and on the combination of subtests used to measure the factor.

A careful parsing of published claims reveals a subtle distinction between what can be inferred from SEM results and what can be accomplished during day-to-day assessments. For example, McGrew (1997) suggested that research finding negligible effects for specific ability factors after considering general ability (e.g., Glutting, Youngstrom, et al., 1997; Youngstrom et al., 1999) was predictive in nature, but that

> to translate specific ability research into practice, to use it to develop meaningful interventions for students with learning problems, an *explanatory* approach is needed. That is, it is not enough to know simply that ability 'x' *predicts* reading comprehension; to translate research into practice it is necessary to know whether or not ability 'x' *affects* reading comprehension. (p. 197; emphasis in original)

Likewise, Keith (1999) proposed that a more complete "understanding" (p. 257) of academic skills could be obtained via assessment of specific cognitive factors. From a theoretical perspective, science seeks the simplest explanations of complex facts and uses those explanations to craft hypotheses that are capable of being disproved (Platt, 1964). Testing of hypotheses typically involves prediction of one kind or another (Ziskin, 1995). Thus, accurate prediction should flow from explanation and understanding of natural phenomena, but understanding without prediction is an inherently weak scientific proof. In clinical practice, an approach which "involves not confusing the ability to *explain* with the ability to *predict*" (Tracey & Rounds, 1999, p. 125; emphasis in original) is recommended to reduce bias and errors in clinical judgment (Garb, 1998). Thus both theory and practice suggest an approach that emphasizes prediction.

Multiple Regression versus SEM

An advantage of the multiple-regression analyses used by certain researchers (e.g., Glutting et al., 1997; Youngstrom et al., 1999) is that they rely on the same measured factor indices clinicians employ in practice. Factor index scores are imperfect,

and this measurement error is present both in the regression analyses and in clinical practice. The advantage of SEM is that it provides estimates of the "true" relationship between such constructs as ability and achievement, with the measurement model removing the effects of measurement error.

The critical issue for both the regression and SEM approaches is to demonstrate effects sufficiently large to have meaningful consequences. In other words, when factor scores are considered to be clinically interpretable (i.e., to show statistically significant or rare strengths or weaknesses), it is still necessary to demonstrate their consequences for individual decision making. For example, Youngstrom and Glutting (2000) found that unusual discrepancies between Verbal Ability and Spatial Ability on the DAS provided a statistically significant improvement ($p < .00005$) to the prediction of reading achievement, above and beyond levels produced by general ability. However, the significant regression coefficient (.21) was then translated to show its consequence for clinical decisions. The comparison revealed that for every 5-point increase in the difference between a child's Verbal Ability and Spatial Ability scores, there was a 1-point change in reading. Even when children showed unusually large Verbal Ability versus Spatial Ability discrepancies (i.e., ≥ 29 points), which occured in less than 5% of the DAS standardization sample, the difference translated into a 6-point change in predicted word knowledge. This amount of predicted change possesses only limited clinical relevance, because it barely exceeds the *standard error of measurement* of the reading measure (i.e., 4 points)!

CONCEPTUAL AND PRACTICAL PROBLEMS WITH INTERPRETING FACTOR SCORES

Besides the lack of incremental, criterion-related validity, there are several troubling conceptual and practical issues associated with the interpretation of factor scores. We now discuss four of these issues, and provide an alternative recommendation addressing the proper diagnostic application of IQ tests.

Contemporary Pressures for Increased Productivity Are Strong

The inclusion of more subtests (and factors) in an ability battery extends administration time. Unfortunately, most psychologists are increasingly confronted with growing caseloads as a consequence of pressures generated from commercial mental health insurance carriers and recent federal regulations that affect school caseloads. At the same time, the Centers for Medicare and Medicaid Services federal guidelines stipulate that Medicaid will not pay for time spent scoring, interpreting, or writing assessment reports. Instead, only the "face-to-face" time spent on test administration will be reimbursed. This policy is significant, because these federal standards are often imitated by other third-party payers—particularly when adoption of the standards offers the possibility of decreased reimbursement.

Similarly, managed care organizations have begun to constrain psychological assessment reimbursement rates (Groth-Marnat, 1999). For example, one national managed care organization only allows 1 hour for administering, scoring, and interpreting a WAIS-III or WISC-III (Eisman et al., 1998), even though published data indicate that these tests require more than twice as long on average (e.g., median values are 75 minutes to administer, 20 minutes to score, and 20 minutes to interpret; Ball, Archer, & Imhof, 1994; Camara, Nathan, & Puente, 1998). The net effect of longer ability tests in these times is that psychologists are caught between societal demands for higher efficiency and a new generation of longer, more time-consuming ability tests.

It is possible to quantify the impact that changes in test administration and interpretation could have in terms of cost. For example, published estimates are available that document the number of practicing school psychologists and their median salary, the median number of assessments completed in a year, and the length of time typically spent in giving and scoring tests. Using these estimates, we find that a 1-hour change in the length of the average evaluation yields a more than $55 million change in costs to educational systems each year! Specifically, the following equation shows:

1-hour change in assessment × $33.33 per hour × 72 assessments per year × 23,000 practitioners = $55,194,480 per year

The hourly rate is based on the median salary and work hours reported in Thomas (2000). The median number of assessments per year is based on remarkably similar figures from two independent surveys: Curtis, Hunley, and Baker (1996) found a median of 72, and Thomas obtained a median of 73. The number of practitioners is based on the report by Lund, Reschly, and Martin (1998).

The $55 million figure is only an estimate, but it is a conservative one for several reasons. One is that the numbers constitute median, not mean, values; therefore, they are less influenced by extreme cases with unusually large salaries or caseloads. In a broader sense, this result is a substantial underestimate of the cumulative effect of a change in assessment practice, because the example only considers school psychologists. There are other large practicing constituencies that spend substantial time in assessment activities (see Camara et al., 1998, for details about the assessment practices of clinical psychologists and neuropsychologists). Clearly, the addition of clinical psychologists, counseling psychologists, and neuropsychologists to the formula can only increase the estimated fiscal impact of changes in assessment practices.

Surveys suggest that there is room for streamlining current assessment-related activity, and that test administration time and scoring contribute substantially to the length of the assessment process (Brown, Swigart, Bolen, Hall, & Webster, 1998; Camara et al., 1998). Based on a review of 271 records from 59 school psychologists, the average time spent on an assessment case was 12.3 hours (median = 11.7, *SD* = 4.1), with test administration consuming the most time (*M* = 2.9 hours, *SD* = 1.2) and the combination of administration and scoring lasting an average of 6.3 hours (*SD* = 2.4; Lichtenstein & Fischetti, 1998).

The reality in most settings is that the demand for evaluation and services far outstrips capacity, with school psychologists spending the majority of their time in assessment-driven activities, and relatively little time in consultation, counseling, or other service delivery roles (Reschly & Wilson, 1997). The time savings offered by adoption of shorter assessment batteries could be viewed as a potential transfer of resources. Each hour not spent in assessment or interpretation is an hour available to provide support services and consultation, or at least to assess a child on a waiting list sooner. As the gross estimates above show, small changes in the assessment procedure (e.g., 1 hour is 8.1% of the average assessment cycle) can yield resource reallocations involving tens to hundreds of millions of dollars each year within the psychoeducational system alone.

Longer Tests May be No Better Diagnostically

The second issue has both theoretical and practical implications. As established at the outset of this chapter, the trend in intelligence testing has been toward developing longer IQ tests that provide psychologists with a wide variety of specific abilities, as reflected by the presence of more subtest scores and factor indices. Practically speaking, no test can hope to evaluate all specific abilities (Horn, 1988). One could imagine an assessment using the "breadth of specificity" created by administering nonredundant subtest measures found among the 14 cognitive subtests from the DAS, the 13 from the WISC-III, and the 21 cognitive subtests from the WJ-R. Such a combined ability measure is not typical of current evaluations, and it would probably be unappealing both to the psychologist and certainly to the individual being tested. Moreover, while such a battery would measure many specific abilities, such an extensive (and time-consuming) assessment would still fail to capture all of the specific abilities identified or proposed for the realm of intelligence (cf. Gardner, 1983; Guilford, 1967; Sternberg, 1988).

Some Variables Beyond *g* are Important

Third, as demonstrated earlier, it has not yet been proven whether the proliferation of factor and subtest scores found in longer IQ

tests actually make a meaningful contribution to differential diagnosis and treatment planning. Most psychologists would agree that at least some abilities beyond *g* are clinically relevant. Examples are the verbal–nonverbal dichotomy in Wechsler's tests and the crystallized–fluid distinctions in the WJ-III, the Stanford–Binet Intelligence Scale: Fourth Edition (SB4; Thorndike, Hagen, & Sattler, 1986), and the Kaufman Adolescent and Adult Intelligence Test (KAIT; Kaufman & Kaufman, 1993).

There is substantial factor-analytic support for both verbal–visual and crystallized–fluid abilities (Carroll, 1993; Keith & Witta, 1997; McGrew, 1997; McGrew et al., 1997; Roid, Prifitera, & Weiss, 1993). More importantly, the external, diagnostic relevance of visual/fluid constructs is evident when psychologists evaluate (1) children and adults with hearing impairments, (2) those with language disorders, (3) individuals whose dominant language is not English, and (4) those with inadequate exposure to formal academic training. In each instance, a verbal/crystallized score is *not* likely to reflect a person's true ability. Similarly, a nonverbal/fluid score is *not* likely to represent the ability of those with (1) visual impairments, (2) physical limitations, or (3) certain forms of acute brain injury. Therefore, both components seem necessary to capture a more accurate estimate than one global score.

Several writers have reviewed the literature regarding the incremental efficacy of verbal versus visual dimensions relative to *g*-based IQs. Jensen (1998) indicates that visual–spatial abilities (along with visual–motor abilities) provide the greatest incremental validity of any second-order ability over and above the criterion variance predicted by g. Hunt (1995) found that visual–spatial reasoning is an important part of understanding mathematics. Children with depressed verbal/crystallized IQs relative to their visual/fluid IQs show more reading problems than normally would be expected (Moffitt & Silva, 1987; Watkins & Glutting, 2000; Youngstrom & Glutting, 2000). Finally, depressed verbal IQs are also more common among children and adolescents with conduct disorder (Kazdin, 1995; Moffitt et al., 1981).

The review above demonstrates that there is an empirical basis for hypotheses generated from discrepancies between verbal/crystallized and visual/fluid abilities. That is, certain outcomes can be predicted with greater precision than that which would result from *g*-based IQs alone. Psychologists therefore must pay careful attention to variation between an individual's verbal/crystallized and visual/fluid IQs.

The Number of Meaningful Variables beyond *g* Appears to Be Small

Fourth, the simple fact that a specific ability can be measured does not necessarily mean that the ability has diagnostic merit (Briggs & Cheek, 1986). As demonstrated in previous sections of this chapter (and elsewhere), a case in point is the well-known Wechsler Freedom from Distractibility factor. The diagnostic and treatment validity of the Freedom from Distractibility factor remains as conjectural today as it was over 40 years ago when Cohen (1959) first discovered the dimension. Indeed, more recent treatment reviews and analyses of diagnostic data raise serious concerns about the importance and utility of deviation IQs based on the Freedom from Distractibility factor (Barkley, 1998; Cohen, Becker, & Campbell, 1990; Kavale & Forness, 1984; Riccio et al., 1997; Wielkiewicz, 1990).

The assessments of factors for other specific abilities or groups of abilities, such as processing speed, sequential and simultaneous processing, Bannatyne categories, deterioration indexes, and the like, are also of theoretical interest. The problem is that their diagnostic and treatment validity is even less well investigated than that of the Freedom from Distractibility factor. Moreover, what little empirical information is available for these abilities is discouraging. Therefore, while a quest for more specific ability constructs is tempting, the amount of empirical support for using most newer constructs advanced over the past 25 years is disappointingly meager. At least for now, those involved in applied clinical assessment would have difficulty empirically justifying the utilization of more assessment tasks than those available in much shorter measures.

Alternative Recommendation

Groth-Marnat (1999) noted that "selection of instruments is a crucial cost consideration especially in cost containment efforts," and hypothesized that it "may be that simpler, briefer tests can make comparable predictions (p. 819). Rather than emphasizing the identification of new or different discrete abilities, taking a different tack might be more useful. An alternative to longer IQ tests would be to develop instruments whose subtests are chosen to possess high loadings on *g*. These measures could possibly be designed to be shorter than, and yet to assess theoretical *g* nearly as well as, longer ability tests. At the same time, these tests could concentrate on the identification of ability dimensions beyond *g* whose diagnostic validities are well established. Examples of this alternative trend are the verbal/crystallized and visual/fluid abilities measured by compact instruments such as the four-subtest Wide Range Intelligence Test (WRIT; Glutting, Adams, & Sheslow, 2000) and the four-subtest Wechsler Abbreviated Scale of Intelligence (WASI; Psychological Corporation, 1999).

CROSS-BATTERY ASSESSMENTS

Interpretation of cognitive test performance has traditionally been based upon subtests contained in a single instrument. Thus the interpretation of subtest profiles has generally been restricted to the individual cognitive test from which the subtests were derived. Recently, however, expansion of profile interpretation to subtests extracted from a variety of cognitive tests has been suggested (Flanagan & McGrew, 1997). Energized by the factor-analytic work of Carroll (1993), and relying upon newer theories about the structure of intelligence (Horn & Noll, 1997), Flanagan and McGrew (1997) have advocated a cross-battery approach to assessing and interpreting intelligence.

Flanagan and McGrew (1997) have asserted that a "synthesized Carroll and Horn–Cattell Gf-Gc model of human cognitive abilities . . . is the most comprehensive and empirically supported model of the structure of cognitive abilities" (p. 316). Operating from this foundation, they have hypothesized that human cognitive abilities can be classified at two levels of hierarchical generality: (1) approximately 70 narrow abilities, which are in turn subsumed by (2) 10 broad abilities. This formulation omits general intelligence. The reason cited for doing so is that *g*

> has little practical relevance to cross-battery assessment and interpretation. That is, the cross-battery approach was designed to improve psychoeducational assessment practice by describing the unique *Gf-Gc pattern of abilities* of individuals that in turn can be related to important occupational and achievement outcomes and other human traits. (McGrew & Flanagan, 1998, p. 14; emphasis in original)

As explained by McGrew and Flanagan, "a global composite intelligence test score is at odds with the underlying Gf-Gc cross-battery philosophy" (p. 382), which "uncovers the individual skills and abilities that are more diagnostic of learning and problem-solving processes than a global IQ score" (p. 383).

Given the large number of abilities identified in the Gf-Gc model, all existing intelligence tests are considered to be "incomplete because they measure between three and five cognitive abilities [i.e., factors], reflecting only a subset of known broad cognitive abilities" (McGrew & Flanagan, 1998, p. 5). To conduct a compete cognitive assessment, therefore, an intelligence test should be augmented with "the most psychometrically sound and theoretically pure tests (according to ITDR [*Intelligence Test Desk Reference*] criteria . . . so that a broader, more complete range of Gf-Gc abilities can be assessed" (McGrew & Flanagan, 1998, p. 357).

McGrew (1997) and McGrew and Flanagan (1998) have published the procedures necessary to operationalize their cross-battery approach. First, subtests from all major intelligence tests have been characterized according to the 10 broad cognitive domains specified in Gf-Gc theory. Calling upon these subtest classifications, the second step in cross-battery assessment entails the examiner's selecting at least two subtests from existing intelligence tests to adequately represent each of the 10 broad cognitive abilities. Then mean scores from each pairing

are calculated to form a factor, and are compared to other factors, to determine whether the abilities are significantly different.

The cross-battery approach is well articulated and noteworthy in many respects. Nonetheless, many theoretical and psychometric issues have not been adequately addressed with respect to cross-battery assessments. We now elucidate and discuss nine prominent concerns: (1) comparability of subtest scores obtained from different instruments, (2) effects associated with modifying the presentation order of subtests, (3) sampling and norming issues, (4) procedures used to group subtests into factors, (5) use of ipsative score interpretation, (6) extent of established external validity, (7) relative efficiency and economy of the assessment process, (8) vulnerability to misuse, and (9) determining the correct number of factors to examine and retain.

Comparability of Scores from Different Tests

All cross-battery comparisons implicitly assume that subtest scores are free from extraneous influences. Regrettably, a host of variables beyond those contributed by differentiated, cross-battery ability constructs could be responsible for score differences. Bracken (1988) identified 10 psychometric reasons why tests measuring similar constructs produce dissimilar results. Among the problems identified are errors introduced by differences in floor effects, ceiling effects, item gradients, and so forth. Similarly, Flynn (1999) has demonstrated that individuals invariably score lower on newer than on older ability tests (i.e., the well-documented "Flynn effect"), and has reported that IQ *subtests* show differential changes across time that are not normally distributed. McGrew and Flanagan (1998) confess that

> scores yielded by cross-battery assessments, taken together, represent an unsystematic aggregate of standardized tests. That is, cross-battery assessments employ tests that were developed at different times, in different places, on different samples, with different scoring procedures, and for different purposes. (p. 402)

Flanagan and colleagues (2000) have since asserted that "the potential error introduced due to cross norm groups is likely negligible" (p. 223), but have provided no evidence to support this claim. However, in light of Bracken's and Flynn's work, it seems reasonable to conclude that subtest scores from cross-battery assessments are likely to be profoundly influenced by extraneous, contaminating influences—variables that subsequently can result in erroneous decisions about children's cognitive strengths and weaknesses.

Order Effects

Another uncontrolled influence inherent in cross-battery assessments is that subtests are administered out of their normative sequence. An example will help to clarify the problem. Let us assume, for instance, that the WISC-III Block Design subtest was administered out of order following administration of the WJ-R battery. A logical question in such circumstances is this: Would the child's Block Design score be lower than, higher than, or unchanged from what it would have been had Block Design been administered within the standard WISC-III test order?

Flanagan (2000) has asserted that "within the context of the cross-battery approach, order of subtests is a trivial matter" (p. 10). However, this statement is not based on data regarding subtest order effects, but rather upon an assumption regarding cross-battery assessment procedures. Namely, if a subtest score is unduly affected by administration order, it is assumed that it will deviate from the other subtests within its broad ability cluster and thus require supplemental subtests to be administered. Under this assumption, the "true" ability measures would cluster together and reveal the discrepant subtest score as spurious. However, this presupposes that erroneous subtest scores always deviate from the remainder of the subtests in their ability cluster, and it ignores the possibility that subtests could be spuriously affected in the direction of other subtests in their ability cluster. For example, let us assume that the WISC-III Block Design subtest is paired with the K-ABC Triangles subtest as constituents of the Visual Processing ability cluster. Let us further assume that the "true" Block Design score is 10, but

that an out-of-order effect has caused it to drop to 8. If the hypothetical examinee's "true" and obtained Triangles scores is 6, then it would appear that the Visual Processing subtests are not significantly discrepant (i.e., 6 vs. 8). However, the Block Design "true" score of 10 is significantly different from the Triangles score; according to cross-battery procedures, another subtest should be administered to measure the Visual Processing cluster more adequately. Thus Flanagan's assumption that cross-battery procedures will correct out-of-order testing effects is faulty. There simply are no data on this issue. Anecdotal reports are available, however, which suggest that subtest scores change according to their administration position (Daniel, 1999). Furthermore, documented order effects are established for the administration of structured interview modules (Jensen, Watanabe, & Richters, 1999).

Practice effects are known to be substantial, especially for nonverbal subtests (Glutting & McDermott, 1990a, 1990b), and pose a related threat to cross-battery validity because the cross-battery approach requires the administration of multiple, similar subtests that were not co-normed. Let us consider yet another scenario: A child is administered Block Design in two different orderings. In one, Block Design is the first subtest administered, followed by Triangles (from the K-ABC) and then Diamonds (a chip construction task from the WRIT). In the second sequence, a child receives Block Design as the third subtest behind Diamonds and Triangles.

There is likely to be a substantial practice effect between the two hypothetical Block Design scores. In the first sequence, the child would not have had the benefit of exposure and practice with similar tasks (triangles, chips constructed of varying numbers of diamonds, and then blocks); in the second sequence, the child would have received such benefits. Test norms used to convert children's raw scores into standard scores are all based on the assumption that the tasks are novel, or at least that no children receive varying exposure to similar measures. Consequently, concerns about order effects and practice effects lead us to conclude that cross-battery procedures are likely to distort performance in an incalculable manner. Thereby, it is incumbent on cross-battery advocates to demonstrate that these effects pose no threat to valid interpretation of test scores.

Sampling and Norming Issues: Size and Representativeness

Ideally, cross-battery factor identification would be accomplished by factor analyses of large, nationally representative samples of children who were administered multiple intelligence tests. This, however, was not done with the cross-battery model. Instead, several small, unrepresentative samples of children completing a small number of intelligence tests were simultaneously analyzed (Flanagan & McGrew, 1998; McGhee, 1993; Woodcock, 1990). In terms of factor analyses, most used by McGrew (1997) as the basis for his categorization system came from the WJ-R concurrent validity samples summarized by Woodcock (1990). One data set included WJ-R and WISC-R scores from 89 third graders. A second included scores from 70 children age 9 on the WJ-R, WISC-R, K-ABC, and SB4. A third involved scores from 53 adolescents age 17 on the WJ-R, WAIS-R, and SB4. Children and adolescents participating in all three studies were from schools in the Dallas–Fort Worth area. Finally, a fourth study included WJ-R and WISC-R scores from 167 children in grades 3 and 5, from schools in Anoka County, Minnesota. Woodcock indicated that each study was analyzed via confirmatory factor analyses. Unfortunately, the sample sizes were simply too small for proper analysis, given the number of variables and parameters involved (Marsh, Hau, Balla, & Grayson, 1998). Furthermore, the samples were all grossly unrepresentative of the national population.

Another study cited by McGrew (1997) included 114 minority children (85 African American and 29 Hispanic) in sixth through eighth grade who were administered 16 WJ-R subtests, 10 KAIT subtests, and one WISC-III subtest. In addition to inadequate sample size and representativeness, analyses were marred by excessive respecification of models based on statistical criteria (Kline, 1998). As noted by Gorsuch (1988), "this procedure has the worst [characteristics] of both exploratory and confirmatory factor analysis and cannot be recommended" (p.

235). Even after capitalizing on sample characteristics with respecifications, model fit statistics did not meet commonly accepted levels necessary to claim plausibility (i.e., goodness-of-fit index (GFI) fit for the final model did not exceed .80, although fit statistics of ≥.90 are recommended; Kline, 1998). Thus the empirical foundation for subtest classifications reported in McGrew (1997) and McGrew and Flanagan (1998) seems weak.

Procedures Employed to Categorize Subtests

Theory should play a prominent role in the selection and organization of subtests. Therefore, it is laudable that proponents of cross-battery assessment have explicitly described their underlying rationale and worked to integrate theory into the structure of assessments. To identify candidate subtests and arrange them into a multifactored battery, the empirical data described above were supplemented by subjective ratings from 10 scholars. As explained by McGrew (1997), "these individuals were asked to logically classify the tests contained in one or more of the intelligence batteries according to the narrow ability factor definitions" (p. 160). However, McGrew reported that "no interrater reliability figures were calculated," and that "when noticeable differences were observed, I made a decision based on a detailed review of Carroll's narrow ability definitions and my task analysis of the test" (p. 160). Thus there is no evidence that experts demonstrated good agreement in assigning subtests into higher-order categories. Although there certainly is a place for the rational derivation of scales in measurement, the approach documented by McGrew did not achieve acceptable standards for constructing a typology of subtests (Bailey, 1994). In addition, the test categorizations originally provided by McGrew were modified by McGrew and Flanagan (1998) and again by Flanagan and colleagues (2000) based upon unspecified logical analyses. When considered together with the weak factor-analytic results, it seems fair to surmise that placement of subtests within cross-battery factors was more a matter of speculative deduction than of demonstrable fact.

Ipsative Interpretation

McGrew and Flanagan (1998) explicitly agreed with extant criticisms of ipsative score interpretation (McDermott et al., 1990, 1992). That is, they accepted that ipsative assessment of subtests from a single cognitive test (i.e., intratest interpretation) "is inherently flawed" (p. 415) due to the unreliability of subtests, the narrow conceptualization of intelligence expressed by subtest scores, and their lack of external validity. However, McGrew and Flanagan concluded that "*some* of the limitations of the ipsative approach to interpretation can be circumvented" (p. 415; emphasis in original) by cross-battery assessment, because it is based upon clusters of subtests that are more reliable and based upon current theories of the structure of intelligence. "Thus, most ipsative test interpretation practice and research have not benefited from being grounded in a well-validated structure of human cognitive abilities" (p. 415). At the same time, McGrew and Flanagan expressed caution regarding complete acceptance of ipsative methods, and suggested that "when significant intra-individual differences are found using Gf-Gc cross-battery data, they should be corroborated by other sources of data" (p. 417).

In essence, then, McGrew and Flanagan (1998) have maintained that the use of cross-battery ability clusters and the Gf-Gc theoretical foundation of their work make ipsative assessment less problematic for cross-battery assessments. However, no empirical data have been advanced in support of this claim. The ipsative interpretive system advocated by Flanagan and colleagues (2000) differs from other popular ipsative systems (e.g., Kaufman, 1994) *only* in its use of factor scores (calculated as the average of at least two subtests) instead of subtest scores. Thus cross-battery ipsative measurement is operationalized by taking factor scores, calculating their grand mean for a given child, and then comparing each factor to the child's personalized mean. Factor scores will nearly always be more reliable than individual subtests. Even so, all other problems elucidated by McDermott and colleagues (1990, 1992) and McDermott and Glutting (1997) remain unsolved. Consequently, it is not apparent how ipsatiza-

tion within the cross-battery framework serves to reduce the mathematical and psychometric weaknesses inherent in the interpretation of ipsatized profiles!

External Validity

Floyd and Widaman (1995) noted that "the ultimate criterion for the usefulness of a factor solution is whether the obtained factor scores provide information beyond that obtained from the global score for the entire scale" (p. 296). Briggs and Cheek (1986) suggested that "factor analysis is not an end in itself but a prelude to programmatic research on a particular psychological construct" (p. 137). As explained by McGrew and Flanagan (1998), a foundational assumption of cross-battery assessment is that "individual skills and abilities . . . are more diagnostic of learning and problem-solving processes than a global IQ score" (p. 383); they asserted that "the cross-battery approach was developed as a means of potentially improving aptitude–treatment interaction (ATI) research" (p. 374). Flanagan and colleagues (2000) proclaimed that "The cross-battery approach defined here provides a systematic means for practitioners to make valid, *up-to-date* interpretations of the Wechsler Intelligence Scales, in particular, and to augment them in a way consistent with the empirically supported *Gf-Gc* theoretical model" (p. 209; emphais in original); asserted that the measurement of Gf-Gc factors "via Wechsler-based cross-battery assessment, supercedes global IQ in the evaluation of learning and problem-solving capabilities"; and stated that the "intracognitive data gleaned from Wechsler-based cross-battery assessments can be translated into educational recommendations" (p. 209).

These claims are sweeping, given the previously reviewed research on specific versus general cognitive factors and the historic failure of aptitude profiles to inform treatments. Reliance on ATI effects is, of course, optimistic, considering the historically unfavorable research literature (Cronbach & Snow, 1977; Gresham & Witt, 1997). Especially strong is the claim that the cross-battery approach leads to the "valid" interpretation of Wechsler scales. However, no new data have been offered to support this broad assertion (Flanagan et al., 2000). Perhaps most telling is a conclusion Flanagan and colleagues (2000) reach themselves: "the diagnostic and treatment validity of the *Gf-Gc* cross-battery approach, like traditional assessment approaches, is not yet available" (p. 288).

Efficiency and Economy

Cross-battery methodology increases test length and complexity at several decision points. For instance, the examiner must determine how many of the 10 postulated broad abilities to measure. Each broad ability area must then be measured by at least two subtests, so if all 10 are selected, then at least 20 subtests would be required. Finally, if the two subtests that measure a specific broad ability are statistically discrepant, then an additional subtest should be administered to clarify the composition of that broad ability cluster.

For the sake of argument, let us compare a 20-subtest cross-battery assessment to the typical WISC-III administration. Most practitioners use the 10 mandatory subtests, rarely administering the optional subtests (Symbol Search and Digit Span) and almost never giving Mazes (cf. Glutting, Youngstrom, et al., 1997). Thus the cross-battery protocol would be roughly twice as long as the modal WISC-III administration. In addition, given the lack of published scoring software or conversion tables, cross-battery approaches are likely to take longer to score. Even assuming that this method adds only 70 minutes to administration and 45 minutes to scoring and interpretation (both estimates are based on the median length of time reported for the WISC-III by practicing clinical psychologists and neuropsychologists; Camara et al., 1998), in practical terms the 20-subtest cross-battery assessment would yield an increased expense of well over $100 million per year within the psychoeducational realm alone (based on our estimates provided earlier in this chapter).

In contrast, McGrew and Flanagan (1998) asserted that the increase in test administration time associated with the cross-battery approach is "negligible" (p. 387), because only portions of complete IQ batteries are used. However, McGrew and

Flanagan and Flanagan and colleagues (2000) have presented model case studies that seem to contradict this conclusion. In the McGrew and Flanagan case study, 14 WJ-R and WISC-III subtests that represented seven broad cognitive areas were first administered. Two broad cognitive areas contained statistically discrepant subtest scores, so at least two more subtests should have been administered to better define these two broad factors (see the Gf-Gc flowchart, p. 405). However, this decision rule was ignored, and only one additional subtest was administered. Thus this cross-battery model case study required 15–16 subtests, depending on adherence to cross-battery decision rules. In the Flanagan and colleagues case study, 14 subtests representing seven cognitive areas were first administered. It was then noted that subtests within three of the seven areas were significantly different. According to their flowchart (p. 267), this should have resulted in administration of at least three more subtests for a minimum of 17 subtests. However, through a complex series of rationalizations, it was determined that two of these cognitive areas should not be further explored while two other cognitive areas should receive detailed attention. This resulted in the administration of a total of 18 subtests. Making a very conservative estimate that each additional subtest beyond the standard 10-subtest battery would require only 6 minutes to administer, score, interpret, and report, cross-battery assessments would increase the length of each cognitive assessment by about 30–48 minutes. Based upon previous presented financial estimates, cross-battery assessment as modeled in McGrew and Flanagan and Flanagan and colleagues would increase yearly psychoeducational assessment expenses by roughly $27.5 to $44.1 million.

The admittedly rough estimates offered here do not include a variety of hidden costs associated with the cross-battery approach. For example, this procedure is likely to require increases in the time spent writing reports. The cross-battery approach also necessitates the purchase of multiple tests and expensive protocols. Furthermore, in many instances the cross-battery approach will incur increased costs for training in the various instruments, or else practitioners run the risk of increased error in administration and scoring. Such expenses could certainly be justified if the new interpretive practices resulted in more worthwhile predictions or educational programming; as shown throughout our presentation on cross-battery assessments, however, the issue of added validity is suspect and certainly open to debate.

Vulnerability to Misuse

Gf-Gc theory, although supported by considerable research as a theory of the structure of intelligence, is still only a theory and not fact (Sternberg, 1996). Even its advocates acknowledge that "there is still much work to do in the factorial study of cognitive abilities. The time is not yet ripe for closing the curtains on this field, as some have suggested" (Carroll, 1995, p. 430). For example, there is no general factor in the Gf-Gc model, whereas factor analyses of intelligence tests persist in finding a robust general factor (Jensen, 1998). With the WISC-III, for example, Keith and Witta (1997) concluded that "the test is first and foremost a measure of general intelligence, or g" (p. 105). In addition, there is no widely accepted explanation as to why the correlation between the Gf factor and the *g* factor is often so close to unity as to suggest only one construct (Gustafsson & Undheim, 1996).

Cross-battery methods have not been validated simply because they are based on Gf-Gc theory. In particular, as previously noted, no evidence to date has conclusively demonstrated that cross-battery assessments are reliable and valid. Repeated statements that cross-battery approaches are based on contemporary, current, modern, or comprehensive theory do not constitute evidence. However, school psychologists and school districts may be prematurely operationalizing cross-battery methods. For example, the Learning Disabilities Assessment Model of the Washington Elementary School District in Phoenix, Arizona (n.d.) utilizes a cross-battery ipsative procedure as one step in the diagnosis of an LD. Seven Gf-Gc factors are delineated and ipsatively compared, but there is no consideration of the theoretical relationships between these factors and various academic achievement dimensions; nor is

there any discrimination among factors regarding importance or scope. In addition, cognitive strengths and weaknesses identified in this manner are not considered in a normative framework. This appears to be directly contrary to a statement by Flanagan and colleagues (2000): "In the absence of empirical evidence that supports the practice of intraindividual or ipsative analysis, it is recommended that ipsative intracognitive analysis be de-emphasized or that it be used in conjunction with interindividual analysis" (p. 284).

Determining the Number of Factors Underlying a Group of Subtests

In factor analysis, one of the most important decisions is determining the appropriate number of dimensions necessary to describe the structure of the data adequately. Various different statistical algorithms have been offered as potential determinants of the number of factors or components to retain in an analysis (see Gorsuch, 1988, for a review). Five of these techniques deserve mention here, although there are other heuristics available.

The "Kaiser criterion" is one of the most widely adopted decision rules, and it is the default criterion employed by exploratory factor analysis procedures in popular statistical software, such as SPSS (SPSS, 1999) and SAS (SAS Institute, Inc., 1990). According to this criterion, components or factors are retained if they possess eigenvalues greater than or equal to 1.0. There is some intuitive appeal to this rule, because it maintains that a component or factor must explain at least as much variance as any single variable contributing to the analysis. Put more simply, a component should be "larger" than a variable in terms of the variance explained.

A second popular procedure is Cattell's scree test, in which the eigenvalues of successive components are plotted and printed. The analyst then takes a straight edge and draws a best-fit line through the "scree" of small eigenvalues. The first point that clearly falls above this line is interpreted as being the smallest component that should be retained for subsequent analyses. The scree test is not as popular as some other alternatives, because it involves subjective judgment in plotting the line (Zwick & Velicer, 1986).

The third approach has begun to supplant the previous two rules in many literatures, including cognitive ability testing. The trend now is to use a chi-square "goodness-of-fit" test, fitting an unrestricted solution to the data. This procedure uses maximum-likelihood (ML) procedures to iteratively estimate the population parameter values that would be most likely to produce the observed data if the specified model were true. The goodness-of-fit test compares the predicted covariances between variables to the actually observed covariances, weighting the discrepancies by sample size. The resulting statistic has a chi-square distribution (if the assumptions of multivariate normality and large sample size are met), with significant values indicating that there is a reliable discrepancy between the model and the observed data. The chi-square technique has gained popularity rapidly, probably both because the statistical software needed to perform these analyses is increasingly available, and also because the approach forms a bridge between exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). Whereas other types of EFA determine the number of factors via post hoc criteria, analysts using ML EFA can specify the number of factors a priori, and then determine the goodness of fit of a model containing that number of factors. ML EFA is more liberal and less theory-driven than CFA, in that it does not require a priori specification of which variables load on which factors (Kline, 1998). In practice, investigators have typically used ML EFA to test the adequacy of several models specifying different numbers of factors. The most parsimonious model producing the lowest chi-square statistic becomes the accepted model in this approach (see Table 6.7 in the WISC-III manual for an example of this sort of application; Wechsler, 1991, p. 195).

The other two procedures—Horn's parallel analysis (HPA; Horn, 1965) and the method of minimum average partials (MAP)—have been available for decades, but have not been incorporated into popular statistical software (Zwick & Velicer, 1986). The omission has hindered widespread adoption of these procedures. Both

approaches have intuitively meaningful interpretations. HPA addresses the fact that principal-components analysis (PCA) summarizes observed variance, even though it may be the product of measurement or sampling error. In theory, if k uncorrelated variables were submitted to PCA, the analysis should produce k components, each with eigenvalues of 1.0. In practice, analyzing a set of variables uncorrelated in the population will yield a first principal component with an eigenvalue somewhat larger than 1.0. How much larger depends on the number of variables (more variables will result in larger first components, all else being equal) and the number of cases (fewer cases lead to less precise estimates, and therefore larger estimated first components when the true population eigenvalue would be 1.0). HPA involves generating artificial data sets with numbers of cases and variables identical to those found in the actual, observed data. The artificial variables are created randomly, implying that the random variables should be uncorrelated in the population. Both the actual and the artificial data are submitted to separate PCAs, and then eigenvalues are compared. Factors (components) are retained only when eigenvalues in an actual data set exceed those in the artificial data. Put another way, components are only considered interpretable if they are larger than what one might observe by chance in analyzing data where the variables are known not to correlate in the population.

The MAP method (Velicer, 1976) relies on the conceptual definition of a factor as a dimension summarizing the correlation between variables. To perform MAP, the investigator submits the data to PCA and saves all component scores. Then the investigator examines partial correlations between the indicator variables, after controlling for the first principal component (in SPSS, this could be achieved using the PARTIAL CORR procedure, specifying the saved principal-component score as a covariate). Next, the investigator calculates the partial-correlation matrix—controlling for the first and second components; then the first, second, and third components; and so on. The average magnitude of the partial correlations will decrease as each factor is removed, until the indicators have been conditioned on all the real factors. When additional components are partialed out, the partial correlations will not decrease further and may even increase. Thus the appropriate number of components is indicated when the smallest, or minimum, average partial correlation is observed.

Until recently, there were no clear advantages to any one of these approaches for identifying the correct number of factors in a data set. Consequently, the choice of which criterion to use was largely a matter of convention or convenience. Factor analyses conducted with most ability tests employed several decision rules, typically adopting the Kaiser criterion, Cattell's scree test, and ML goodness-of-fit tests as the standards (e.g., Thorndike et al., 1986; Wechsler, 1991, 1997; Woodcock & Johnson, 1989). However, Monte Carlo studies conducted over the last 15 years convincingly demonstrate that MAP and HPA perform much better than the alternatives in recovering the correct number of factors (Velicer, Eaton, & Fava, 2000; Zwick & Velicer, 1986). The Kaiser criterion and ML goodness-of-fit test both tend to overestimate the number of factors. Cattell's scree test appears fairly accurate, but still less so than MAP or HPA.

The methodological findings in the preceding paragraph imply that published ability tests have probably overestimated the number of dimensions assessed by each battery. For example, the WISC-III purportedly measures four cognitive abilities, according to analyses published in the technical manual (Wechsler, 1991). However, other authorities dispute whether or not the Freedom from Distractibility factor emerges (e.g., Sattler, 1992). No published analysis to date has used HPA or MAP—the procedures that possess the best methodological support. When HPA is applied to the median correlations published in the WISC-III manual, results using all 13 subtests suggest that only two components should be interpreted, clearly corresponding to the Verbal Comprehension and Perceptual Organization Indexes (see Table 15.3). If only the 10 mandatory subtests are administered, then the WISC-III only measures one factor sufficiently to meet the HPA criterion.

The preceding analyses should not be interpreted as indicating that the Freedom from Distractibility and Processing Speed

TABLE 15.3. Horn's Parallel Analysis (HPA) of the WISC-III Median Correlations: Eigenvalues Listed by Component Number (n = 2,200)

	13 subtests		10 subtests	
Component	Observed	Avg. random	Observed	Avg. random
1	**5.63**	1.12	**4.97**	1.10
2	**1.25**	1.10	1.01	1.08
3	1.04	1.07	0.89	1.05
4	0.84	1.06	0.67	1.03
5	0.76	1.03	0.56	1.01

Note. "Avg. random" values based on the average of five random data sets. **Boldface** numbers exceed comparable eigenvalue for random data, thus meeting HPA criterion for retention.

factors do not exist. Instead, the outcomes make it reasonable to conclude that the WISC-III does not contain a sufficient number of subtests to adequately satisfy statistical criteria for interpretation of the Freedom from Distractibility and Processing Speed factors. The addition of subtests that identify these specific dimensions could increase the amount of covariance attributable to each factor (thus increasing the eigenvalue, leading to the retention of the factor when the augmented battery is used). However, results clearly indicate that the WISC-III is not long enough (i.e., does not contain a sufficient number of subtests) to meet statistical criteria for retaining more than one or two factors.

Similar to that for the WISC-III, the cross-battery approach has not applied either of the two optimal algorithms for determining the appropriate number of factors to retain. Without using these types of decision rule in an empirical analysis, the risk is that investigators will interpret factors that have not adequately been measured by the battery of indicators. Experts agree that overfactoring is less problematic than underfactoring (i.e., retaining too few dimensions), but that neither is desirable (Wood, Tataryn, & Gorsuch, 1996). In a clinical context, overfactoring leads practitioners to interpret aggregates of subtests as if they measured a more general construct—when in fact the communality between the subtests is not sufficient to measure the purported factor in large groups, let alone individuals. The HPA analysis of the WISC-III shows that it has been overfactored. The poor measurement of the Freedom from Distractibility and Processing Speed dimensions has probably contributed to the difficulty in establishing incremental validity for these constructs.

In summary, cross-battery approaches should carefully document which combinations of subtests are adequate to measure broad cognitive abilities. These analyses should rely on decision rules such as MAP or HPA, and not traditional criteria, because there are clear methodological advantages to these newer approaches. MAP or HPA analyses of the WISC-III suggest that it actually may be difficult to measure an ability adequately with only a pair of subtests (both the Freedom from Distractibility and Processing Speed factors contain only two subtests as indicators). Consequently, it becomes more crucial for advocates of cross-battery approaches to determine what subtest constellations are sufficient to measure each construct.

CONCLUSION

Cross-battery cognitive assessment is explicitly based upon the theories of Horn and Noll (1997) and Carroll (1993), and sees 10 broad second-order factors as more important for diagnosis and treatment than the higher order *g* factor. Since no single intelligence test adequately measures all ten broad cognitive factors hypothesized within the cross-battery model, subtests are extracted from a variety of cognitive tests and combined to create measures of the Gf-Gc factors. However, a number of theoretical and psychometric issues underlying cross-battery assessments have not been adequately addressed. Many of these technical impedi-

ments to the cross-battery approach derive from measurement issues created by pulling subtests from their standardized protocols and forming conglomerates of subtests with different reference groups. Many of these pitfalls could be avoided by standardizing subtests measuring the various Gf-Gc broad cognitive ability factors all within one test and sample. This is planned for the Stanford–Binet Intelligence Scales: Fifth Edition (measuring nine factors; see Youngstrom, Glutting, & Watkins, Chapter 10, this volume) and has been done for the WJ-III (measuring nine factors using 46 subtests).

However, even if these tests possess the desired factor structure and the same excellent psychometric qualities that have distinguished earlier editions of these instruments, important issues will still remain before multifactor assessment will be ready to contribute to clinical and psychoeducational assessment. First, the law of parsimony will require demonstrations that specific ability factors substantially outperform predictions based on omnibus, full-scale scores alone. Second, ipsative interpretation methods used with factors must be empirically demonstrated to be reliable and valid. Finally, the incremental validity of factor scores must translate into improved treatment, diagnosis, or educational interventions. These gains must be judged large enough—by policy makers and consumers, as well as practitioners—to justify the increased time and expense required for thorough multifactor assessment. Although research on assessment, including cross-battery methods, should continue, it should not prematurely be applied to make high-stakes diagnostic decisions about children.

REFERENCES

Accardo, P. J., & Whitman, B. Y. (1991). The misdiagnosis of the hyperactive child. In P. J. Accardo, T. A. Blondis, & B. Y. Whitman (Eds.), *Attention deficit disorders and hyperactivity in children* (pp. 1–21). New York: Marcel Dekker.

Adams, M. J. (1990). *Beginning to read: Thinking and learning about print.* Cambridge, MA: MIT Press.

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

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

Arizona Department of Education. (1992). *Standard reporting format for psychoeducational evaluation.* Phoenix: Author.

Bailey, K. D. (1994). *Typologies and taxonomies: An introduction to classification techniques* (Vol. 102). Thousand Oaks, CA: Sage.

Ball, J. D., Archer, R. P., & Imhof, E. A. (1994). Time requirements of psychological testing: A survey of practitioners. *Journal of Personality Assessment, 63,* 239–249.

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

Barkley, R. A. (1998). *Attention-deficit hyperactivity disorder: A handbook for diagnosis and treatment* (2nd ed.). New York: Guilford 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.

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.

Bracken, B. A. (1988). Ten psychometric reasons why similar tests produce dissimilar results. *Journal of School Psychology, 26,* 155–166.

Bracken, B. A., McCallum, R. S., & Crain, R. M. (1993). WISC-III subtest composite reliabilities and specificities: Interpretive aids. *Journal of Psychoeducational Assessment Monograph Series,* 22–34.

Briggs, S. R., & Cheek, J. M. (1986). The role of factor analysis in the development and evaluation of personality scales. *Journal of Personality, 51,* 106–148.

Brody, N. (1985). The validity of tests of intelligence. In B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements, and applications* (pp. 353–389). New York: Wiley.

Brody, N. (1997). Intelligence, schooling, and society. *American Psychologist, 52,* 1046–1050.

Brown, M. B., Swigart, M. L., Bolen, L. M., Hall, C. W., & Webster, R. T. (1998). Doctoral and nondoctoral practicing school psychologists: Are there differences? *Psychology in the Schools, 35,* 347–354.

Bus, A. G., & van Uzendoorn, M. H. (1999). Phonological awareness and early reading: A meta-analysis of experimental training studies. *Journal of Educational Psychology, 91,* 403–413.

Camara, W., Nathan, J., & Puente, A. (1998). *Psychological test usage in professional psychology: Report of the APA practice and science directorates.* Washington, DC: American Psychological Association.

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 Chil-

dren—Third Edition. *Psychological Assessment, 10,* 285–291.

Carnine, D. W., Silbert, J., & Kameenui, E. J. (1997). *Direct instruction reading* (3rd ed.). Upper Saddle River, NJ: Merrill.

Carroll, J. B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Carroll, J. B. (1995). On methodology in the study of cognitive abilities. *Multivariate Behavioral Research, 30,* 429–452.

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

Ceci, S. J. (1991). How much does schooling influence general intelligence and its cognitive components?: A reassessment of the evidence. *Developmental Psychology, 27,* 703–722.

Ceci, S. J., & Williams, W. M. (1997). Schooling, intelligence, and income. *American Psychologist, 52,* 1051–1058.

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

Cohen, J. (1959). The factorial structure of the WISC at ages 7–6, 10–6, and 13–6. *Journal of Consulting and Clinical Psychology, 23,* 285–299.

Cohen, J. (1990). Things I have learned (so far). *American Psychologist, 45,* 1304–1312.

Cohen, M., Becker, M. G., & Campbell, R. (1990). Relationships among four methods of assessment of children with attention deficit-hyperactivity disorder. *Journal of School Psychology, 28,* 189–202.

Cromwell, R. L., Blashfield, R. K., & Strauss, J. S. (1975). Criteria for classification systems. In N. Hobbs (Ed.), *Issues in the classification of children* (Vol. 1, pp. 4–25). San Francisco: Jossey-Bass.

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods.* New York: Irvington.

Curtis, M. J., Hunley, S. A., & Baker, A. C. (1996, March). *Demographics and professional practices in school psychology: A national perspective.* Paper presented at the annual meeting of the National Association of School Psychologists, Atlanta, GA.

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

Daniel, M. (1999, September). *Subtest order* [Online]. Available: http://www.home.att.net

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.

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.

Ehri, L. C., Nunes, S. R., Willows, D. M., Schuster, B. V., Yaghoub-Zadeh, Z., & Shanahan, T. (2001). Phonemic awareness instruction helps children learn to read: Evidence from the National Reading Panel's meta-analysis. *Reading Research Quarterly, 36,* 250–287.

Eisman, E. J., Dies, R. R., Finn, S. E., Eyde, L. D., Kay, G. G., Kubiszyn, T. W., Meyer, G. J., & Moreland, K. L. (1998). *Problems and limitations in the use of psychological assessment in contemporary health care delivery: Report of the Board of Professional Affairs Psychological Assessment Workgroup, Part II.* Washington, DC: American Psychological Association.

Elliott, C. D. (1990). *Differential Ability Scales.* 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.

Flanagan, D. P. (2000). *Giving subtests out of the framework of the standardization procedure* [Online]. Available: http://www.home.att.net

Flanagan, D. P., & McGrew, K. S. (1997). A cross-battery approach to assessing and interpreting cognitive abilities: Narrowing the gap between practice and cognitive science. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 314–325). New York: Guilford Press.

Flanagan, D. P., & McGrew, K. S. (1998). Interpreting intelligence tests from contemporary Gf-Gc theory: Joint confirmatory factor analysis of the WJ-R and KAIT in a non-white sample. *Journal of School Psychology, 36,* 151–182.

Flanagan, D. P., McGrew, K. S., & Ortiz, S. O. (2000). *The Wechsler intelligence scales and Gf-Gc theory: A contemporary approach to interpretation.* Boston: Allyn & Bacon.

Floyd, F. J., & Widaman, K. F. (1995). Factor analysis in the development and refinement of clinical assessment instruments. *Psychological Assessment, 7,* 286–299.

Flynn, J. R. (1999). 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.

Garb, H. N. (1998). *Studying the clinician.* Washington, DC: American Psychological Association.

Gardner, H. (1983). *Frames of mind: The theory of multiple intelligences.* New York: Basic Books.

Glutting, J. J., Adams, W., & Sheslow, D. (2000). *Wide Range Intelligence Test Manual.* Wilmington, DE: Wide Range.

Glutting, J. J., & McDermott, P. A. (1990a). Childhood learning potential as an alternative to traditional ability measures. *Psychological Assessment: A Journal of Consulting and Clinical Psychology, 2,* 398–403.

Glutting, J. J., & McDermott, P. A. (1990b). Principles and problems in learning potential. In C. R. Reynolds & R. W. Kamphaus (Eds.), *Handbook of psychological and educational assessment: Vol. 1. Intelligence and achievement* (pp. 296–347). New York: Guilford Press.

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

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

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

Glutting, J. J., Youngstrom, E. A., & McDermott, P. A. (2000). *Validity of the WISC-III processing speed factor in identifying children's achievement problems: An epidemiological study using the WISC-III/WIAT linking sample.* Manuscript submitted for publication.

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

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

Goodyear, P., & Hynd, G. W. (1992). Attention-deficit disorder with (ADD/H) and without (ADD/WO) hyperactivity: Behavioral and neuropsychological differentiation. *Journal of Clinical Child Psychology, 21,* 273–305.

Gorsuch, R. L. (1988). Exploratory factor analysis. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental research* (2nd ed., pp. 231–258). New York: Plenum Press.

Gough, H. (1971). Some reflections on the meaning of psychodiagnosis. *American Psychologist, 26,* 106–187.

Gregory, R. J. (1999). *Foundations of intellectual assessment: The WAIS-III and other tests in clinical practice.* Boston: Allyn & Bacon.

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 Review, 12,* 249–267.

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

Groth-Marnat, G. (1999). Financial efficacy of clinical assessment: Rational guidelines and issues for future research. *Journal of Clinical Psychology, 55,* 813–824.

Guilford, J. P. (1967). *The nature of human intelligence.* New York: McGraw-Hill.

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

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

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 & Bacon.

Hartigan, J. A., & Wigdor, A. K. (1989). *Fairness in employment testing.* Washington, DC: National Academy Press.

Herrnstein, R. J., & Murray, C. (1994). *The bell curve: Intelligence and class structure in American life.* New York: Free Press.

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

Horn, J. L. (1965). A rationale and test for the number of factors in factor analysis. *Psychometrika, 30,* 179–185.

Horn, J. L. (1988). Thinking abut human abilities. In J. R. Nesselroade & R. B. Cattell (Eds.), *Handbook of multivariate experimental psychology* (2nd ed., pp. 645–685). New York: Plenum Press.

Horn, J. L., & Noll, J. (1997). Human cognitive capabilities: Gf-Gc theory. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 53–91). New York: Guilford Press.

Hsiao, J. K., Bartko, J. J., & Potter, W. Z. (1989). Diagnosing diagnoses. *Archives of General Psychiatry, 46,* 664–667.

Hunt, E. (1995). The role of intelligence in modern society. *American Scientist, 83,* 356–368.

Hunter, J. E., & Hunter, R. F. (1984). Validity and utility of alternative predictions of job performance. *Psychological Bulletin, 96,* 72–98.

Ivnik, R. J., Smith, G. E., Malec, J. F., Kokmen, E., & Tangalos, E. G. (1994). Mayo cognitive factor scales: Distinguishing normal and clinical samples by profile variability. *Neuropsychology, 8,* 203–209.

Jencks, C. (1972). *Inequality: A reassessment of the effect of family and schooling in America.* New York: Basic Books.

Jensen, A. R. (1980). *Bias in mental testing.* New York: Free Press.

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

Jensen, P. S., Watanabe, H. K., & Richters, J. E. (1999). Who's up first?: Testing for order effects in structured interviews using a counterbalanced experimental design. *Journal of Abnormal Child Psychology, 27,* 439–445.

Jones, W. T. (1952). *A history of Western philosophy.* New York: Harcourt, Brace.

Kamphaus, R. W. (1993). *Clinical assessment of children's intelligence.* Boston: Allyn & Bacon.

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

Kaufman, A. S., & Kaufman, N. L. (1983). *K-ABC: Kaufman Assessment Battery for Children.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1993). *The Kaufman Adolescent and Adult Intelligence Test manual.* Circle Pines, MN: American Guidance Service.

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 Disabilities Quarterly, 7,* 136–156.

Kazdin, A. E. (1995). *Conduct disorders in childhood and adolescence* (2nd ed.). Thousand Oaks, CA: Sage.

Keith, T. Z. (1990). Confirmatory and hierarchical confirmatory analysis of the Differential Ability Scales. *Journal of Psychoeducational Assessment, 8,* 391–405.

Keith, T. Z. (1999). Effects of general and specific abilities on student achievement: Similarities and differences across ethnic groups. *School Psychology Quarterly, 14,* 239–262.

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.

Kessel, J. B., & Zimmerman, M. (1993). Reporting errors in studies of the diagnostic performance of self-administered questionnaires: Extent of the problem, recommendations for standardized presentation of results, and implications for the peer review process. *Psychological Assessment, 5,* 395–399.

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 annual meeting of the National Association of School Psychologists, Seattle, WA.

Kline, R. B. (1998). *Principles and practice of structural equation modeling.* New York: Guilford Press.

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

Kraemer, H. C. (1988). Assessment of 2 × 2 associations: Generalization of signal-detection methodology. *American Statistician, 42,* 37–49.

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.

Kranzler, J. H. (1997). What does the WISC-III measure?: Comments on the relationship between intelligence, working memory capacity, and information processing speed and efficiency. *School Psychology Quarterly, 12,* 110–116.

Lezak, M. D. (1995). *Neuropsychological assessment* (3rd ed.). New York: Oxford University Press.

Lichtenstein, R., & Fischetti, B. A. (1998). How long does a psychoeducational evaluation take?: An urban Connecticut study. *Professional Psychology: Research and Practice, 29,* 144–148.

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

Lubinski, D., & Benbow, C. P. (2000). States of excellence. *American Psychologist, 55,* 137–150.

Lund, A. R., Reschly, D. J., & Martin, L. M. C. (1998). School psychology personnel needs: Correlates of current patterns and historical trends. *School Psychology Review, 27,* 106–120.

Marsh, H. W., Hau, K. T., Balla, J. R., & Grayson, D. (1998). Is more ever too much?: The number of indicators per factor in confirmatory factor analysis. *Multivariate Behavioral Research, 33,* 181–220.

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, R. A. (1992). Illusion of meaning in the ipsative assessment of children's ability. *Journal of Special Education, 25,* 504–526.

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

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.

McGhee, R. (1993). Fluid and crystallized intelligence: Confirmatory factor analysis of the Differential Ability Scales, Detroit Tests of Learning Aptitude—3, and Woodcock–Johnson Psycho-Educational Battery—Revised. *Journal of Psychoeducational Assessment Monograph Series,* 20–38.

McGrew, K. S. (1997). Analysis of the major intelligence batteries according to a proposed comprehensive Gf-Gc framework. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests and issues* (pp. 151–179). New York: Guilford Press.

McGrew, K. S., & Flanagan, D. P. (1998). *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment.* Boston: Allyn & Bacon.

McGrew, K. S., Keith, T. Z., Flanagan, D. P., & Underwood, M. (1997). Beyond *g*: The impact of Gf-Gc specific cognitive ability research on the future use and interpretation of intelligence tests in the schools. *School Psychology Review, 26,* 189–201.

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

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, pat-

terns, or cutting scores. *Psychological Bulletin, 52,* 194–216.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement* (pp. 13–103). New York: Macmillan.

Moffitt, T. E., Gabrielli, W. F., Mednick, S. A., & Schulsinger, F. (1981). Socioeconomic status, IQ, and delinquency. *Journal of Abnormal Psychology, 90,* 152–156.

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.

Murphy, J. M., Berwick, D. M., Weinstein, M. C., Borus, J. F., Budman, S. H., & Klerman, G. L. (1987). Performance on screening and diagnostic tests. *Archives of General Psychiatry, 44,* 550–555.

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.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (3rd ed.). New York: McGraw-Hill.

Oh, H. J., & Glutting, J. J. (1999). An epidemiological-cohort study of the DAS process speed factor: How well does it identify concurrent achievement and behavior problems? *Journal of Psychoeducational Assessment, 17,* 362–275.

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.

Platt, J. R. (1964). Strong inference. *Science, 146,* 347–352.

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

Prifitera, A., Weiss, L. G., & Saklofske, D. H. (1998). The WISC-III in context. In A. Prifitera & D. H. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist–practitioner perspectives* (pp. 1–39). New York: Academic Press.

Psychological Corporation. (1999). *Wechsler Abbreviated Scale of Intelligence Manual.* San Antonio, TX: Author.

Ree, M. J., & Earles, J. A. (1993). Intelligence is the best predictor of job performance. *Current Directions in Psychological Science, 1,* 86–89.

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 validity 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 Press.

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. 425–439). Washington, DC: National Association of School Psychologists.

Reschly, D. J., & Wilson, M. S. (1997). Characteristics of school psychology graduate education: Implications for the entry-level discussion and doctoral-level specialty definition. *School Psychology Review, 26,* 74–92.

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

Roid, G. H., Prifitera, A., & Weiss, L. G. (1993). Replication of the WISC-III factor structure in an independent sample. *Journal of Psychoeducational Assessment Monograph Series,*

Rosnow, R. L., & Rosenthal, R. (1989). Statistical procedures and the justification of knowledge in psychological science. *American Psychologist, 44,* 1276–1284.

Rutter, M. (1989). Isle of Wight revisited: Twenty-five years of child psychiatric epidemiology. *Journal of the American Academy of Child and Adolescent Psychiatry, 28,* 833–653.

Salvia, J., & Ysseldyke, J. E. (1998). *Assessment* (5th ed.). Boston: Houghton Mifflin.

SAS Institute, Inc. (1990). *SAS/STAT User's Guide: Version 6, Fourth Edition, Volume 1, ACECLUS-FREQ.* Cary, NC: Author.

Sattler, J. M. (1992). *Assessment of children: Revised and updated third edition.* San Diego, CA: 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.

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

Sines, J. O. (1966). Actuarial methods in personality assessment. In B. A. Maher (Ed.), *Progress in experimental personality research* (pp. 133–193). New York: Academic Press.

SPSS. (1999). *SPSS Base 9.0 Applications Guide* (Version 9.0). Chicago: Author.

Stahl, S. A., & Murray, B. A. (1994). Defining phonological awareness and its relationship to early reading. *Journal of Educational Psychology, 86,* 221–234.

Sternberg, R. J. (1988). *The triarchic mind: A new theory of human intelligence.* New York: Viking Press.

Sternberg, R. J. (1996). Myths, countermyths, and truths about intelligence. *Educational Researcher, 25,* 11–16.

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

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.

Thomas, A. (2000). School psychology 2000: Salaries. *Communique, 28,* 32.

Thorndike, R. L. (1984). *Intelligence as information processing: The mind and computer.* Bloomington, IN: Center on Evaluation, Development, and Research.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). *Stanford–Binet Intelligence Scale: Fourth Edition.* Chicago: Riverside.

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 & Bacon.

Traub, R. E. (1991). *Reliability for the social sciences.* Newbury Park, CA: Sage.

Truch, S. (1993). *The WISC-III companion: A guide to interpretation and educational intervention.* Austin, TX: Pro-Ed.

Velicer, W. F. (1976). Determining the number of components from the matrix of partial correlations. *Psychometrika, 41,* 321–327.

Velicer, W. F., Eaton, C. A., & Fava, J. L. (2000). Construct explication through factor or component analysis: A review and evaluation of alternative procedures for determining the number of factors or components. In R. D. Goffin & E. Helmes (Eds.), *Problems and solutions in human assessment: A festschrift to Douglas Jackson at seventy* (pp. 1–89). Mahwah, NJ: Erlbaum.

Wagner, R. K. (1997). Intelligence, training, and employment. *American Psychologist, 52,* 1059–1069.

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

Washington Elementary School District. (n.d.). *Learning disabilities assessment model.* Phoenix, AZ: Author.

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., & 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). WISC-R subtest analysis of variance: The right way, the wrong way, or no way? *School Psychology Review, 23,* 640–651.

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

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

Watkins, M. W., & 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.*

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

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

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

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

Wiggins, J. S. (1973). *Personality and prediction: Principles of personality assessment.* Reading, MA: Addison-Wesley.

Williams, W. M., & Ceci, S. J. (1997). Are Americans becoming more or less alike?: Trends in race, class, and ability differences in intelligence. *American Psychologist, 52,* 1226–1235.

Wood, J. M., Tataryn, D. J., & Gorsuch, R. L. (1996). Effects of under- and overextraction on principal axis factor analysis with varimax rotation. *Psychological Methods, 1,* 354–365.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. *Journal of Psychoeducational Assessment, 8,* 231–258.

Woodcock, R. W., & Johnson, M. B. (1989). *Woodcock–Johnson Tests of Cognitive Ability—Revised.* Chicago: Riverside.

Woodcock, R. W., & Johnson, M. B. (2000). *Woodcock–Johnson Tests of Cognitive Ability—Third Edition.* Itasca, IL: Riverside.

Youngstrom, E. A., & Glutting, J. J. (2000). *Individual strengths and weaknesses on factor scores from the Differential Ability Scales: Validity in predicting concurrent achievement and behavioral criteria.* Manuscript submitted for publication.

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

Ziskin, J. (1995). *Coping with psychiatric and psychological testimony* (5th ed.). Los Angeles, CA: Law and Psychology Press.

Zwick, W. R., & Velicer, W. F. (1986). Comparison of five rules for determining the number of components to retain. *Psychological Bulletin, 99,* 432–442.