School Psychology Review
1997, Vol. 26, No. 2, pp. 176–188

The Base Rate Problem and Its Consequences for Interpreting Children's Ability Profiles

Joseph J. Glutting
University of Delaware

Paul A. McDermott
University of Pennsylvania

Marley M. Watkins
Pennsylvania State University

Joseph C. Kush
Duquesne University

Timothy R. Konold
University of Virginia

Abstract: Base rates refer to the proportion of a population that falls within a diagnostic category, either identifying an exceptionality (e.g., learning disability [LD], emotional disturbance [ED], or simply representing "normal" variation. This article familiarizes readers with the importance and meaning of base rates. It presents several univariate and multivariate base-rate procedures useful for identifying unusual IQ subtest variation. It compares the various base-rate procedures with the statistical significance-testing approach routinely used by psychologists. The mathematical superiority of one base-rate procedure is highlighted (i.e., the *nonlinear multivariate base-rate method*), and its practical and scientific benefits are discussed. The nonlinear multivariate base-rate method is used to address the more important question of whether subtest analysis has validity for differential decision making. Specifically, the nonlinear multivariate method is employed to determine whether children with LD (*N* = 925) and ED (*N* = 100) are more likely to show unusual subtest patterns than children from the normative sample of the Wechsler Intelligence Scale for Children-Third Edition (*N* = 2,200). Results are discussed and recommendations are provided for improving future research on subtest analysis.

A golden anniversary is about to take place in the field of individual intelligence testing. The precipitating event occurred in 1949 when subtests were introduced on the newly created Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949). Since then, literally hundreds of publications have promoted the analysis of children's subtest scores. This legacy of interpretive relevance continues to be reflected in textbooks on intelligence testing wherein, despite the recent presentation of some notable limitations and caveats, page-after-page remain devoted to the identification of unusual subtest patterns and the generation of inferential hypotheses about them (cf. Kamphaus, 1993; Kaufman, 1994; Sattler, 1992).

The current article has several purposes. One is to familiarize readers with the term *base rates*. Another is to present one univariate, and two multivariate, base-rate methods for identifying unusual IQ subtest profiles. These procedures will be contrasted against the statistical significance-testing approach commonly employed by psychologists. Serious limitations will be identified with (a) the statistical significance-testing approach, (b) the univariate base-rate approach, and (c) the linear multivariate base-rate approach. An alternative method of interpretation will then be presented. This procedure employs *nonlinear multivariate base rates* as the mechanism for identifying unusual subtest profiles. This article will discuss the advantages of comparing subtest scores to these nonlinear multivariate taxonomies to obtain accurate base rates, and it will present a 10 subtest taxonomy developed recently for the standardization sample of the Wechsler Intelli-

Address all correspondence concerning this article to Joseph J. Glutting, University of Delaware, School of Education, Department of Educational Studies, Willard Hall Education Building, Newark, DE 19716-2901.

gence Scale for Children-Third Edition (WISC-III; Wechsler, 1991). Most importantly, the *validity* of interpretations based on ability profiles will be investigated by comparing subtest scores to the WISC-III taxonomy using a large sample of children (N = 1,025) identified as having learning disabilities (LD) or emotional disturbance (ED). The article concludes with recommendations for improving future research on ability testing.

Drawbacks of Statistical Significance Testing and Univariate Base Rates

Historically, about the same time that profile analysis was becoming popular with the WISC, measurement specialists working from a different perspective, recognized that questions about profile variation were best addressed through *nonlinear methods* of statistical analysis (Cattell, 1949; Horst, 1941; Mosel & Roberts, 1954; Osgood & Suci, 1952). These procedures, supported by research, were never incorporated into the ability-testing literature. During the decades that followed, the predominant research strategy was to investigate ability profiles using either *linear-univariate* or *linear-multivariate* methodologies. Likewise, almost without exception, practitioners adopted the two linear-univariate methods recommended by authoritative sources on ability assessment (Kamphaus, 1993; Kaufman, 1994; Sattler, 1992; Wechsler, 1991). The first consists of examining statistical significance levels between one or more sets of subtest scores. The second documents variations in univariate base rates.

A number of publications have addressed similarities and differences between statistical significance testing and univariate base rates (Cahan, 1986; Glutting, McDermott, Prifitera, & McGrath, 1994; Silverstein, 1993; Stone, 1991). Establishing the *statistical significance* of a score discrepancy is important because it greatly enhances the probability that the difference is not merely due to chance. However, statistically significant differences can be quite common and ordinary. They simply reflect the distinct, but natural, variation of test scores and are not necessarily a reason for concern.

By way of example, consider the situation Glutting and his colleagues present for the WISC-III (Glutting, Konold, McDermott, Kush, & Watkins, in press). They examined the number of children from the WISC-III standardization sample (N = 2,200) who showed at least one statistically significant subtest deviation. Scores from the 10 mandatory subtests were compared one at a time to children's personal means (optional WISC-III subtests were excluded). Statistically significant deviations were determined by tabled $p < .05$ critical values identified in the WISC-III manual (see Table B.3, p. 264). The analysis was restricted to the delineation of weaknesses (i.e., children showing subtest scores significantly below their own mean). The number of strengths was not investigated. Results showed that 42.7% of the children had at least one statistically significant subtest weakness. Thus, when clinicians use statistical significance as an interpretive guideline, they are willing to identify some sort of learning problem on the WISC-III, or generate an hypothesis, for more than 40% of the children in the United States.

The implications of base rates are of special interest in diagnostic assessment, where base rates refer to the frequency, or percentage, of a population that falls within a particular diagnostic category (Cureton, 1957; Meehl & Rosen, 1954; Wiggins, 1973). For instance, the high base rate of "exceptional" subtest profiles identified by statistical significance testing is a problem that has begun to be recognized in textbooks on intelligence testing (Kamphaus, 1993; Kaufman, 1994; Sattler, 1992). The common response is to encourage psychologists to compare and contrast subtest scores to distributions of *univariate base rates*. The analyses customarily begin by subtracting a child's lowest subtest score from his or her highest subtest score. The resulting difference is compared to cumulative percentages reported for the test's standardization sample, and a decision is made whether the obtained discrepancy shows an unusual (i.e., infrequent) base rate. The procedure is univariate because only one difference is derived, even though two subtest scores are used.

Unfortunately, the univariate base-rate approach suffers from a number of limitations. First, its analyses do not account for the strength or pattern of correlations among subtest scores. As a result, some comparisons are prone to showing larger (or smaller) differences as a consequence of the magnitude of association between the subtests being analyzed. Second, the methods are univariate. Only one difference score is compared to the appropriate distributional statistics (i.e., standardization sample mean and standard deviation). The comparison must then be repeated as necessary (e.g., between individual subtest scores and the average Verbal or Performance Scale score). Third, *profiles* are nonlinear, multivariate entities and they are quite unlike individual subtest scores or linear composites formed from groups of subtest scores. The net effect is that univariate base rates *distort* the true frequency of score differences in much the same way as that shown for statistical significance testing.

Multivariate Methodologies

In reality, *all* univariate methods are inadequate to analyze groups of subtest scores because profile analysis requires *multiple* dependent comparisons. As indicated at the outset of this article, measurement specialists have recognized for nearly 5 decades that profiles are integrated sets of test scores that require appropriate hypotheses and statistical treatments (Cattell, 1949; Horst, 1941; Mosel & Roberts, 1954). Two classes of multivariate methods can be used to examine profiles. Cattell (1949) referred to the procedures as either *R* or *Q* analysis. Both account for correlations among subtest scores. Moreover, because the procedures are multivariate, they are capable of completing multiple comparisons *simultaneously* — the typical situation that occurs during psychodiagnostic appraisals. Multivariate methods also better honor multidifferentiated views of intelligence as well as the full network of relationships that exist among such abilities (Sternberg, 1984). Likewise, they better account for the *true* (i.e., multivariate) base rate of score differences in the population.

R analysis is founded on the linear variation of test scores. However, by their nature, subtest profiles are doubly defined according to level (position toward the upper, central, or lower region of the ability continuum) and shape (the pattern of peaks and valleys across subtest scores). *R* analysis is insensitive to differences in *both* profile level and shape. *Q* analysis, on the other hand, respects both types of variation and is better able to address nonlinear, configural hypotheses (Cattell, Coulter, & Tsujioka, 1966; Tatsuoka, 1974).

Applications of Nonlinear Multivariate Methodology

Beginning in the last decade *Q* methodology was used to group children according to the level and shape of their ability scores. *Normative* taxonomies of the most common subtest profiles have been developed for standardization samples from a number of individually administered IQ tests, including the WISC-R, WAIS-R, WPPSI, K-ABC, and DAS (respectively, McDermott, Glutting, Jones, & Noonan, 1989; McDermott, Glutting, Jones, Watkins, & Kush, 1989; Glutting & McDermott, 1990; Glutting, McGrath, Kamphaus, & McDermott, 1992; Holland & McDermott, in press). The principal advantage of comparing subtest scores to these normative taxonomies is that they constitute a mathematically superior method of identification when a given subtest profile is clinically unusual and atypical of the most common, mutlivariate patterns of intellectual abilities.

Taxonomies for the WISC-III

We previously derived two normative taxonomies, comprising either 10 or 12 sets of subtest scores, for the standardization sample of the WISC-III (respectively, Glutting et al., in press; Glutting, McDermott, & Konold, 1997). Table 1 provides mean subtest scores and corresponding IQs for the 10 subtest taxonomy. The eight most common, or "core" types are arranged by descending order of FSIQs, and names are assigned on

Table 1
Mean Subtest Score Patterns and Associated Deviation IQs for the WISC-III 10 Subtest Taxonomy

Profile type number	Mean subtest score[a]										Mean deviation quotient[b]			
	PC	IN	CD	SM	PA	AR	BD	VO	OA	CM	FSIQ	VIQ	PIQ	Name and description
1	13	14	13	14	13	14	15	14	14	14	126	124	124	High ability
2	13	13	10	12	12	12	12	12	12	12	114	113	112	Above average ability
3	10	12	13	12	10	12	10	12	10	13	109	112	104	Above average ability & VIQ > PIQ
4	10	9	13	10	12	10	11	9	11	10	103	97	108	Average ability & PIQ > VIQ
5	10	11	8	11	9	10	10	10	10	10	99	102	96	Average ability & VIQ > PIQ
6	9	7	9	7	9	8	10	7	10	7	89	85	96	Below average ability & PIQ > VIQ
7	7	8	9	8	8	8	6	9	7	9	88	92	85	Below average ability
8	6	5	7	5	6	6	5	5	6	6	73	75	76	Low ability

Note. *N* = 2,2200. Tabled values are rounded to nearest whole number for convenient presentation. WISC-III = Wechsler Intelligence Scale for Children-Third Edition; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; FSIQ = Full Scale IQ; VIQ = Verbal Scale IQ; PIQ = Performance Scale IQ. The data in this table are copyrighted © 1995 by The Psychological Corporation. For permission to reproduce, transform, or otherwise adapt these data, contact The Psychological Corporation.

[a] The population standard score *M* = 10 and *SD* = 3 for each age group.

[b] Deviation quotients are conventional IQ equivalents specific to each age group with the population *M* = 100 and *SD* = 15.

the basis of this variation plus outstanding VIQ/PIQ contrasts.

A perusal of the core types reveals that general ability level is their most distinguishing characteristic. In addition, nearly all of the profile types tend to display score differences within general ability levels. For instance, profile types 3 and 5 are defined not only by general ability, but also by the presence of more severe VIQ > PIQ discrepancies than would normally be expected.

It is important to note that "severe" IQ differences in the core profile types were determined by cut scores derived across the WISC-III normative sample, whereby VIQ-PIQ differences > 22 points comprise 3% of VIQ > PIQ differences and PIQ-VIQ differences > 24 points comprise 3% of PIQ > VIQ differences. The 3% criterion approximates differences nearly two standard deviations above and below the population mean respectively and is consistent with the standard established by McDermott, Glutting, Jones, and Noonan (1989).

Therefore, for profile type 3, one would expect 3.0% of the children in this type to show a 22-point VIQ > PIQ difference. However, in actuality, 6.3% of the children exhibited a severe (i.e., 22-point) VIQ > PIQ discrepancy and no child exhibited a severe (i.e., 24-point) PIQ > VIQ discrepancy. Profile type 5 shows a similar outcome. Instead of the expected 3.0%, 5.6% of the children in this profile type exhibited severe VIQ > PIQ discrepancies. Conversely, profile types 4 and 6 show more PIQ > VIQ discrepancies, and profile type 7 is characterized by fewer PIQ > VIQ discrepancies. Interesting also is that deviations for Arithmetic and Coding often coincide directionally within ability levels, for example, when the two subtests covary to indicate relatively greater aptitude (profile types 7 and 8) or lesser aptitude (profile types 1 and 2).

Optional subtests from the WISC-III were not used in the analyses just above. However, as mentioned earlier, we also previously developed a 12 subtest taxonomy for the WISC-III (Glutting et al., 1997). This taxonomy included scores from the optional Digit Span and Symbol Search subtests. Interestingly, inclusion of the two other subtests had the effect of identifying profile variation associated with a third and fourth factor beyond variation associated with the FSIQ, VIQ, and PIQ. In other words, certain profile types in the 12 subtest taxonomy were not only defined by FSIQ variation, but also by variation associated with the Freedom from Distractibility and Processing Speed factors.

Implications for Practice: How to Best Identify Unusual Subtest Profiles

Several procedures can be used to compare subtest scores to the WISC-III taxonomy. The simplest is based on generalized distance theory (D^2) (Osgood & Suci, 1952), and it is the method recommended for everyday decision making. It begins by comparing a child's subtest scores to the three core types closest to his or her general ability level. If the sum of the squared differences for a child's profile is ≥ 98 for *each* comparison, the profile may be interpreted as being uncommon. By contrast, if any of the sums is < 98, the profile cannot be considered uncommon.

Glutting et al. (in press) provide a case example for readers interested in using the 10 subtest WISC-III taxonomy. Likewise, a case example is presented for the 12 subtest WISC-III taxonomy (Glutting et al., 1997). These earlier papers provide explicit, step-by-step computations on how to use generalized distance theory to make diagnostic decisions. The papers also contrast results from the generalized distance approach with those obtained using either the statistical significance-testing approach or the univariate base-rate approach. Lastly, each paper shows, in detail, the specific methodology and rationale used to uncover the 10 and 12 subtest taxonomies.

The Validity Issue

The base-rate problem is resolved when subtest scores are compared to a core profile taxonomy. However, by themselves, the comparisons do not address the more fundamental issue of whether subtest analysis is *valid*.

Psychologists receive extensive training in how to make sense of the information gathered during an examination. *Hypothesis generation* is the primary mechanism used to derive plausible interpretations. This process is creative and speculative. It seeks to develop informed guesses and working conjectures about the psychological functioning of children according to the score patterns they receive on diagnostic tests. For instance, a psychologist might infer that a child with a WISC-III Performance > Verbal Scale difference suffers from an expressive language disorder, or alternatively, that the child is more adept at processing visually presented material.

There is a flip side to hypothesis generation. It is *hypothesis testing.* This process is factual, scientific, and data driven. The purpose of hypothesis testing is to support, or disconfirm, the validity of inferences derived during hypothesis generation. Hypothesis generation and hypothesis testing are complimentary endeavors. Each is essential to differential decision making. The problem is that we know far more about how to develop interpretive hypotheses than we do about their validity.

Multiple sources of evidence can be used to validate score interpretations (Messick, 1989). However, in diagnostic assessment, two types of evidence are primary. Diagnostic, score-based interpretations become valid to the extent they (a) are associated with a viable *treatment* for individuals suffering from a disorder, or (b) accurately *predict* a high probability that an individual will contract a problem or disorder (Cromwell, Blashfield, & Strauss, 1975; Glutting et al., 1992; Gough, 1971; McDermott, 1981).

For some unknown reason, psychologists have come to believe that treatment validity is the most important evidence for intelligence tests. This situation is unfortunate because it occurred at the sake of prediction. Prediction is valuable in its own right because we may never be able to remediate all of the negative circumstances that can impact children's growth and well being. Moreover, with the exception of findings for global ability, treatment validity remains very much in doubt for more differentiated ability profiles, with research consistently demonstrating few positive outcomes for *multiple* aptitude by treatment interactions (Cronbach & Snow, 1977; Heller, Holtzman, & Messick, 1982; Ysseldyke & Christenson, 1988).

The predictive validity of WISC-III subtest profiles will now be investigated using a sample of children previously identified as having either LD or ED. Research on the utility of subtest analysis is most often directed to children experiencing LD or ED (Kavale & Forness, 1984; Mueller, Dennis, & Short, 1986). Therefore, if the profiles of children with LD or ED are found to be probabilistically (i.e., predictively) similar to the WISC-III taxonomy, it must be concluded that the profiles represent undistinctive variants of normal abilities and are not open to the generation of hypotheses about cognitive strengths or weaknesses. Alternatively, if the profiles deviate substantially from the WISC-III taxonomy, the outcome would provide empirical support for the continued interpretation of subtest profiles.

Method

Participants

The sample comprised students enrolled in special education programs in the states of Arizona, Delaware, New Jersey, Pennsylvania, Texas, and Virginia. Each child received a comprehensive psychological evaluation and was selected for study according to two criteria: (a) cognitive assessment, which included the 10 mandatory subtests of the WISC-III (supplementary subtest scores were excluded), and (b) a diagnosis of LD or ED. Only a small number of the sampled children were classified with mild ($N = 41$) or moderate ($N = 3$) mental retardation. Therefore, they were excluded.

The selection criteria identified a total of 1,025 participants. Of this total, membership was 925 in the LD group and another 100 in the ED group. The average age was 12 years, 5 months ($SD = 2.6$ years). Gender distribution was 69% male and 31% female. Ethnicity was 50% Anglo, 10% Black, 23% Hispanic, 16% American Indian, and 1% Other. Socioeconomic background data were unavailable.

Table 2
Prevalence of Special Education Groups Failing to Fit a WISC-III Core Profile Type

Group	Special education samples		WISC-III normative sample	z	p[a]
Emotionally disturbed	6.0%	vs.	5.4%	0.25	ns
Learning disabled	7.4%	vs.	5.4%	2.09	.05

Note. N = 2,200 for WISC-III normative sample; N = 100 for emotionally disturbed sample; N = 925 for learning disabled sample.

[a] Identification of significant prevalence trends is based on tests of the standard error of proportional differences corrected for the number of simultaneous statistical contrasts by the Bonferroni method.

Procedure

Generalized distance theory (D^2) offers the most convenient mechanism for the discovery of unusual subtest patterns; however, it is somewhat imprecise. Instead, similarity of children's WISC-III profiles to the eight core profile types was assessed by the $r_p(k)$ group similarity coefficient because it better accounts for correlations among variables than D^2, and it is the more accurate of the two methods in returning children to their correct core type and/or identifying unusual subtest profiles (Tatsuoka, 1974; Tatsuoka & Lohnes, 1988). A coefficient of <.16 was applied to identify children classified as LD or ED who failed to fit a core type (i.e., showed an unusual subtest profile).[1] Selection rates for the $r_p(k)$ method were determined by the 5.4% prevalence criterion established previously for the WISC-III norm group (Glutting et al., in press). Prevalence trends were calculated between children in special education who failed to fit a core type and those from the WISC-III normative sample using two-tailed tests of the standard error of proportional differences corrected for the number of contrasts (Ferguson, 1981).

Results

The first analysis indicated that children with ED do not exhibit unusual profiles more often than the population at large (see Table 2, upper part). On the other hand, the second analysis would seem to uphold longstanding beliefs about the diagnostic richness of subtest scores (see Table 2, lower part). This comparison revealed that children with LD are significantly more likely to display tell-tale patterns of specific abilities than children from the WISC-III standardization sample (p < .05).

However, as we have cautioned throughout this article, statistically significant differences can be misleading. A magnitude of effect statistic was used to overcome the problem. Specifically, Cohen's (1988) coefficient h was calculated between the proportion of children from the LD sample who showed unusual subtest profiles (7.4%) versus those from the WISC-III standardization sample (5.4%). The obtained h (.085) constitutes an extremely small effect size: "small" is defined as any $h \leq .20$ (Cohen, 1988). In other words, based on the obtained h of .085, profiles between the two groups show a 93.3% degree of overlap. (See Cohen, 1988, p. 184, for directions for calculating degree of overlap.)

The inconsequential group difference can be better understood from a more practical perspective. For every 100 children psychologists classify as LD, only two will display an unusual subtest profile more often than that expected for the U.S. population (i.e., 7.4% for LD vs. 5.4% for the WISC-III standardization sample; 7.4% – 5.4% = 2%). Thus, results from both the ED and LD comparisons discourage subtest analysis and raise serious concerns about whether multidifferentiated constructions of intelligence possess as much validity as that obtainable from more general, or even unitary, constructions.

Discussion

The circuses of P. T. Barnum were extremely popular. One reasons for their widespread appeal is that they had something for nearly everyone. Paul Meehl (1956) is credited with identifying a *Barnum effect* in personality assessment. It occurs when psychologists generate interpretive hypotheses from profiles that have high base rates of occurrence in the population. (See Furnham & Schofield, 1987, for a literature review and analysis.)

Like that for personality assessment, interpretations are often attached to ability profiles that are commonplace and ordinary. The present study investigated Barnum effects by employing a large data set obtained across multiple states. Barnum effects were evaluated by comparing the subtest scores of children with LD and ED to core profile types for the WISC-III. The advantage of this method over other procedures for identifying unusual profiles is that the core types supply nonlinear, multivariate base rates against which subtest scores can be compared.

Results showed, in essence, that children with LD and ED were no more likely to exhibit exceptional subtest configurations than children in general. The present investigation expanded the original Barnum effect definition to research on children's ability profiles. It revealed that subtest scores from the WISC-III failed to identify educational or psychological problems more often than levels available from common, multivariate base rates.

We previously enumerated several methodological problems which operate to negate, or equivocate, most of the research of children's ability profiles (McDermott, Fantuzzo, & Glutting, 1990; McDermott, Fantuzzo, Glutting, Watkins, & Baggaley, 1992). Among these factors 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, and it is a limitation that even undergraduate textbooks on research methodology warn against. The consequence of self-selection is that it unduly increases the probability of discovering group differences.

The current outcomes are all the more disturbing when viewed in light of the circularity problem. The WISC-III was used as part of the assessment battery to place children into the LD and ED groups and as the mechanism for identifying unusual profiles. Nonetheless, even with this unfair advantage, irregular WISC-III subtest scatter was no more likely to occur for children classified as LD or ED than it was for the greater population of children in the U.S.

Some Other Practical Implications for Daily Practice

In this article, we introduced the base-rate problem to the literature on children's ability testing, and drawing upon the personality studies of Meehl (1956), we labeled the phenomenon a Barnum effect. Results from an empirical investigation also established the practical and scientific consequences of Barnum effects by showing how they operate to confound outcomes in validity studies of subtest analysis. But of even greater relevance and dismay, Barnum effects raise the possibility that we may all be hurting children as a consequence of interpreting common and ordinary ability profiles as being rare in occurrence and as having diagnostic import and meaning.

Recommendations for Improving the Future of Ability Testing

The purpose of this section is to present suggestions for advancing the quality of research on children's ability variation. Despite nearly 50 years of study, research on subtest analysis continues to be plagued by methodological pitfalls and deficiencies (McDermott et al., 1990; McDermott et al., 1992). However, to move beyond these shortcomings, we must first refocus some of our attention from past research practices and personal preferences to the planning and production of more methodologically pertinent inquiry.

Specifically, we recommend that future research employ concomitant use of (a) *predictive methodology* (i.e., longitudinal research designs) and (b) *heterogeneous* *samples* (i.e., samples comprising children from special education *and* regular education). Implementation of these two procedures is the only effective remedy to the circular reasoning, base-rate problems, and host of other methodological limitations, that haunt current inquiry. To our knowledge, no investigation used the two procedures with children's subtest scores, and it is for this reason, we reiterate our earlier advice to "Just say no to subtest analysis" (McDermott et al., 1990). In other words, psychologists should refrain from speculations about the relative strengths and weaknesses in subtest profiles until methodologically sound inquiry offers preponderant and convincing evidence on their behalf.

The first caution compels us to remind readers of a second caution that we made elsewhere (Glutting et al., 1992; Glutting et al., 1997). This admonition is directed to psychologists who will ignore more recent research and persist in generating configural hypotheses according to the peaks and valleys of children's subtest scores. Psychologists who elect to differentially interpret subtest profiles, and who do so without comparing them to a core profile taxonomy, run a serious risk of mistaking common ability patterns as being rare and noteworthy. Such practice can only convolute decision making and it is unlikely to help children.

One study was extraordinary with respect to its methodological rigor. Moffitt and Silva (1987) examined unusual VIQ-PIQ differences (i.e., those with a base rate, or prevalence ≤ 10%) on the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974). Their sample was a large, unselected cohort followed longitudinally from birth (N = 925). The WISC-R was administered at ages 7, 9, and 11. Results showed children with unusual VIQ < PIQ discrepancies were more likely to develop reading problems. However, contrary to popular expectations, no significant effects emerged across a multitude of other outcomes, including spelling and mathematics achievement, etiological and health factors (e.g., pre-, peri-, and post-natal data regarding low birth weight for gestational age, low Apgar score, illnesses and accidental injuries during childhood, etc.), motor development, several indicators of neuropsychological sequelae, and parent and teacher reports of behavior problems. Equally important, no significant differences were evident on *any* outcome variable for children with unusual PIQ < VIQ discrepancies.

Subtest scores were not recorded in the Moffitt and Silva study (P. A. Silva, personal communication, May 14, 1992). Nonetheless, their investigation demonstrates that it is possible to employ longitudinal research designs and appropriate samples. Cost is the principal obstacle to this form of inquiry. For instance, we established that 5.4% of the WISC-III standardization sample shows unusual subtest profiles (Glutting et al., 1997). If 1,000 children were tested at random from the general population, and their subtest scores compared to the WISC-III taxonomy, approximately 50 would have unusual profiles (1,000 × .054 = 54). A sample of 50 is adequate for comparison purposes, but there is no guarantee that a sample this large would materialize when 1,000 children are actually tested. The prevalence for unusual profiles could be made somewhat more lenient, as Moffitt and Silva (1987) did when they used a 10% base rate for VIQ-PIQ discrepancies. In such an instance, 1,000 cases would yield approximately 100 children with unusual profiles (1,000 × .10 = 100). Thereafter, it would be possible to *wait*, to compare and contrast this group on important outcome variables to children without exceptional subtest profiles.

Quasi-Experimental Option

IQ tests are published to make a profit. Preferred sources of research support (e.g., the federal government, Spencer and Ford Foundations, etc.) are unlikely to finance any type of inquiry that might increase the sales of "for profit" tests. Test companies are also unlikely to finance the type of research we advocate because it generates no direct revenues. Therefore, we suggest another less satisfactory, but more practical option to testing large, random samples from the general population and following them longitudinally.

Triennial re-evaluations are currently mandated for children in special education. Furthermore, the current study demonstrated that children in special education (i.e., those with LD and ED) are no more likely to show unusual subtest profiles than the population at large. Given that unusual profiles are as prevalent among children in special education, it would be comparatively easier to test, and thereafter follow, large cohorts of these children. Groups with, and without, unusual profiles would be identified at the time of initial evaluation. Comparisons on important criteria would then be made during mandatory re-evaluations.

The proposed coupling of longitudinal research designs with available samples from special education constitutes a quasi-experimental approach to investigating subtest profiles. Unfortunately, the use of available samples causes a loss of randomness important to experimental discovery, and results from nonrandom, available samples are subject to regression and interaction effects (Campbell & Stanley, 1966). However, our proposed quasi-experimental strategy has the benefit of being more pragmatic and cost effective than testing random groups whose background characteristics need to approximate that of the U.S. population.

External validity, or generality, is the single greatest liability of quasi-experimentation (Campbell & Stanley, 1966). Therefore, it would be difficult to generalize results from the proposed quasi-experimental studies of children's ability profiles to the universe of children not attending special education. The overall effect is that the findings may hold only for that unique group of children selected for special education in the first place.

Efficacy of Group vs. Idiographic Data

Group methods are required to implement the research strategies presented in the preceding section. However, employing group data to study children's ability profiles is not without criticism. Each person represents a unique, intricate constellation of psychological functioning. Clinical assessment is characterized by the fact that only one person is tested at a time, and the tests themselves are selected to provide information helpful to that specific individual (cf. American Psychological Association, *Standards for educational and psychological testing*, 1985, p. 45). Given the personalized nature of clinical assessments, some professionals believe that research findings from group data may not apply to individuals. Specifically, according to this view, claims about the utility of subtest profiles are more meaningfully answered by directing inquiry to idiographic case-by-case analysis (see Kaufman, 1994, p. 36; O'Neill, 1993, chap. 4).

We would like to respond to the issue by paraphrasing Meehl (1986). If outcomes from group studies cannot be applied to individuals, there would be no point in conducting randomized trials to determine the validity of various medical techniques. A case in point is the polio vaccine experiments whose successful results are employed with individuals everyday — just as group findings from *all* medical studies are inevitably transferred to specific people. Hence, it typically is the case that results from group data provide excellent insights into the functioning of individuals.

In conclusion, our position regarding the merit of group versus idiographic data might be different if the constructs under consideration were amorphous, singularly specific to a given context, and/or difficult to measure (see Meehl, 1986, for a discussion). Perhaps then an idiographic orientation would be superior. However, profile analysis begins with variables that, by definition, are measured in rank order and distributed under asymptotic normal probabilities. Moreover, the hypotheses associated with subtest profiles are prognostic and testable. Thus, to infer under this latter set of circumstances that results from group data cannot be transferred to individuals is *just wrong* and tantamount to asserting that the science of Galilean multiple-case replication should give way to Aristotelian single-case anecdote.

References

American Psychological Association. (1985). *Standards for educational and psychological testing*. Washington, DC· Author.

Cahan, S. (1986). Significance testing of subtest score differences: The rules of the game. *Journal of Psychoeducational Assessment, 4*, 273–280.

Campbell, D. T., & Stanley, J. C. (1966). *Experimental and quasi-experimental designs for research.* Chicago: Rand McNally.

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

Cattell, R. B., Coulter, M. A., & Tsujioka, B. (1966). The taxonomic recognition of types and functional emergents. In R. B. Cattell (Ed.), *Handbook of multivariate experimental psychology* (pp. 288–329). Chicago: Rand McNally.

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

Cromwell, R. L., Blashfield, R. K., & Strauss, J. S. (1975). Criteria for classification systems. In S. Hobbs (Ed.), *Issues in the classification of children* (Vol. 1, pp. 425). San Francisco: Jossey-Bass.

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

Cronbach, L. J., & Snow, R. E. (1977). *Aptitudes and instructional methods: A handbook for research on interactions.* New York: Irvington.

Cureton, E. E. (1957). Recipe for a cookbook. *Psychological Bulletin, 54*, 494–497.

Ferguson, G. A. (1981). *Statistical analysis in psychology & education* (5th ed.). New York: McGraw-Hill.

Furnham, A., & Schofield, S. (1987). Accepting personality test feedback: A review of the Barnum effect. *Current Psychological Research & Review, 6*, 162–178.

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

Glutting, J. J., & McDermott, P. A. (1990). Patterns and prevalence of core profile types in the WPPSI standardization sample. *School Psychology Review, 19*, 471–491.

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

Glutting, J. J., McDermott, P. A., Prifitera, A., & McGrath, E. A. (1994). Core profile types for the WISC-III and WIAT: Their development and application in identifying multivariate IQ-achievement discrepancies. *School Psychology Review, 23*, 610–639.

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.

Gough, H. (1971). Some reflections on the meaning of psychodiagnosis. *American Psychologist, 26*, 160–167.

Heller, K. A., Holtzman, W. H., & Messick, S. (Eds.). (1982). *Placing children in special education: A strategy for equity.* Washington, DC: National Academy Press.

Holland, A. M., & McDermott, P. A. (in press). Discovering core profile types in the school-age standardization sample of the Differential Ability Scales. *Journal of Psychoeducational Assessment.*

Horst, P. (1941). The prediction of personal adjustment. Social Science Research Council Bulletin (No. 48). New York: Author.

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

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

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.

McDermott, P. A. (1981). Sources of error in the psychoeducational diagnosis of children. *Journal of School Psychology, 19*, 31–44.

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

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

McDermott, P. A., Glutting, J. J., Jones, J. N., & Noonan, J. V. (1989). Typology and prevailing composition of core profiles in the WAIS-R standardization sample. *Psychological Assessment, 1*, 118–125.

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

Meehl, P. E. (1956). Wanted — A good cookbook. *American Psychologist, 11*, 262–272.

Meehl, P. E. (1986). Diagnostic taxa as open concepts: Metatheoretical and statistical questions about reliability and construct validity in the grand strategy of nosological revision. In T. Millon & G. L. Klerman (Eds.), *Contemporary directions in psychopathology* (pp. 215–231). New York: Guilford.

Meehl, P. E., & Rosen, A. (1955). Antecedent probability and the efficiency of psychometric signs, patterns, and 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., & Silva, P. A. (1987). WISC-R verbal and performance IQ discrepancy in an unselected cohort: Clinical significance and longitudinal stability. *Journal of Consulting and Clinical Psychology, 55*, 768–774.

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

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

O'Neill, A. M. (1993). *Clinical inference: How to draw meaningful conclusions from tests.* Brandon, VT: Clinical Psychology Publishing.

Osgood, C. E., & Suci, G. J. (1952). A measure of relation determined by both mean differences and profile information. *Psychological Bulletin, 49*, 251–262.

Sattler, J. M. (1992). *Assessment of children* (3rd ed. rev.). San Diego, CA: Author.

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

Sternberg, R. J. (1984). The Kaufman Assessment Battery for Children: An information-processing analysis and critique. *Journal of Special Education, 18*, 269–278.

Stone, B. J. (1991). Significance testing of the difference vs. the frequency of the difference: What's the significance? I don't know! National Association of School Psychologists, *Communiqué, 20*(4), 26.

Tatsuoka, M. M. (1974). *Classification procedures: Profile similarity.* Champaign, IL: Institute for Personality and Ability Testing.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis* (2nd ed.). New York: Macmillan.

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

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

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

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

Ysseldyke, J. E., & Christenson, S. L. (1988). Linking assessment to intervention. In J. L. Graden, J. E. Zins, & M. J. Curtis (Eds.), *Alternative educational delivery systems: Enhancing instructional options for students* (pp. 91–109). Washington, DC: National Association of School Psychologists.

Footnote

[1]Copies of computer programs used to calculate generalized distance (*D*) and $r_p(k)$ may be obtained from the senior author. Both operate in SPSS and can be applied to any sample. The programs read subtest standard scores from a data file, match children to the WISC-III core types, and print either (*D*) or $r_p(k)$ values for each child (one for each of the 8 core types). The programs identify children who fail to fit a core type. They also can be modified to meet specific purposes.

Joseph J. Glutting, PhD, is a Professor in School Psychology at the University of Delaware. His research interests include the interpretation of results from individually-administered tests of ability, achievement, and personality.

Paul A. McDermott, PhD, is Professor of Measurement at the University of Pennsylvania. He has published extensively in the areas of psychoeducational assessment and measurement.

Marley W. Watkins, PhD, is an Associate Professor in School Psychology at Pennsylvania State University. He is a Diplomate in School Psychology, American Board of Professional Psychology, and his research interests include diagnostic assessment, the development of microcomputer interpretation programs, and computer assisted instruction.

Joseph C. Kush, PhD, is an Assistant Professor in school psychology at Duquesne University . His research interests include cognitive and intellectual theory and assessment.

Timothy R. Konold, PhD, is an Assistant Professor in the Department of Leadership, Foundations, and Policy at the Curry School of Education, University of Virginia. His research interests include the application of psychometric theory and principles to the educational and psychological assessment of children and adults.