November 1991

Validity of intelligence test scores in the definition of learning disability: A critical analysis

D. J. Francis

K. A. Espy
University of Nebraska-Lincoln, kespy2@unl.edu

B. P. Rourke

J. M. Fletcher

CHAPTER 2

Validity of Intelligence Test Scores in the Definition of Learning Disability: A Critical Analysis

David J. Francis, Kimberly A. Espy, Byron P. Rourke, and Jack M. Fletcher

The relationship of intelligence test performance to learning deficiency is a longstanding issue affecting treatment and research on learning-disabled children. Despite many questions concerning the use of intelligence tests for classifying disabled learners, these tests have become entrenched in every form of work with these children (Kaufman, 1979; Sattler, 1988).

In general, intelligence test scores are used to separate children with generalized impairments of learning (e.g., those who are mentally deficient) from children who have more isolated forms of learning impairment (e.g., those who are learning disabled). Additional distinctions are sometimes made among learning-disabled children in an attempt to separate or classify those children reading at levels appropriate for their measured intellectual potential from those reading below their intellectual potential. Most prominent has been Rutter and Yule's (1975) distinction between "general reading backwardness" and "specific reading retardation."

Although most research involving these types of distinctions concerns children who are deficient in reading, the notion readily generalizes to other academic problems (e.g., arithmetic) and neurobehavioral disorders (e.g., attention deficit–hyperactivity disorder). Virtually any definition of learning disability used for policy (e.g., Kavanaugh & Gray, 1986) employs the concept of intelligence as an index of learning potential. This notion is even more firmly embedded in current definitions used for research on learning disabilities.

The widespread employment of conceptions of intelligence in the definition of disabled learners belies the many problems associated with this practice. Some of these problems are conceptual, whereas others are psychometric/statistical. Unfortunately, well-reasoned argument has been unsuccessful in altering current conceptions, and empirical data have not been adequately employed in addressing these important issues.

Conceptual Problems with the Use of IQ Tests

The conceptual problems underlying the use of IQ test scores with learning-disabled children largely involve the notion that such scores are indices of learning potential. When IQ scores are used as an index of potential, the underlying assumptions revolve around the notion that there is a measurable constant that can be labeled "potential." Historically, this hypothesis is influenced by the idea of a generalized intelligence factor ("g") that represents some type of innate, biologically derived factor that sets upper limits on ability attainment (Spearman, 1923). When this limit is not attained, either constitutional or environmental explanations are postulated for the discrepancies. These ideas and their role in definitions of childhood neurobehavioral disorders can be found in Still (1902). Similar notions were the basis of concepts such as "minimal brain injury" (Strauss & Lehtinen, 1947) and were epitomized in policy-based definitions of "minimal brain dysfunction" and "specific learning disability" (Satz & Fletcher, 1980).

As Rutter (1978) suggested, definitions of learning disability that use IQ tests to index potential are vague and poorly operationalized. Taylor, Fletcher, and Satz (1984) summarized many of the problems with the use of IQ tests to measure potential. These problems included the multifactorial nature of composite scores of intellectual functioning. In other words, an IQ score is a summary of several aspects of cognitive functioning. Some aspects are correlated with reading ability (e.g., vocabulary), whereas other aspects have little relationship to reading skill (e.g., puzzle assembly). To the extent that an IQ score is related to reading ability, the IQ score will likely reflect the severity of the reading disorder. Median correlations are approximately .70 (Kaufman, 1979; Sattler, 1988). Lower scores on IQ tests may merely reflect the pervasiveness of cognitive impairment as opposed to an upper limit on cognitive ability. Similarly, some children with lower IQ scores have reading levels in excess of their measured intelligence. All this phenomenon indicates is that the IQ test does not measure some skill that is related to reading proficiency, not that the child is an "overachiever." For example, no intelligence test of which we are aware measures phonological segmentation skills, which are highly related to decoding skills in reading (Rosner & Simon, 1971).

Finally, the use of IQ test scores as an index of learning potential represents a complex causal network in which the joint influences of reading proficiency on IQ and vice versa are difficult to disentangle (Doehring, 1978). It may be that lower IQ scores result from rather than "cause" reading deficiency. These conceptual problems should be considered carefully when IQ scores are used as a primary defining characteristic of learning disability (Fletcher & Morris, 1986).

Psychometric Problems with the Use of IQ Scores

In addition to conceptual problems, the use of IQ test scores to define learning disabilities raises serious psychometric considerations. These problems have been discussed, but they have not seriously been considered in various policy statements and "official" definitions.

What Test?

One obvious problem involves the IQ test score to be used (Morris, 1988). For example, if the Wechsler Intelligence Scale for Children (Wechsler, 1949) is used, should the index of potential be the Verbal IQ, Performance IQ, or Full Scale IQ? What if another IQ test is used (e.g., Stanford–Binet, Fourth Edition, or Kaufman Assessment Battery for Children)? Current implementations of policy at the level of the school lead to unclear arguments among practitioners about various composites and test scores for an individual child. For a child with marked discrepancies in abilities, composite scores that average these discrepancies may mask true potential in the academic area.

What Cut-Off?

A related problem is where to place the cut-off for deciding minimal levels of "average" intelligence. As Morris (1988) suggested, there is little empirical evidence favoring cut-offs of 70, 80, or 90—yet all these scores find their way into empirical studies of disabled learners. Other approaches use some type of discrepancy between achievement and intelligence, but again, the extent of discrepancy necessary to yield a positive diagnosis is vague. Why should a criterion of 1 standard deviation be used as opposed to a criterion of 2 standard deviations?

Age-Based Criteria

When discrepancy-based definitions have been used, the extent of discrepancy has been expressed relative to age and IQ. The problems with age-based criteria are well known (Fletcher & Morris, 1986; Reynolds, 1984). Basically, age-based discrepancies represent unstandardized metrics that vary across age. For example, a child reading 2 years below age level at age 9 is more seriously impaired than is a 16-year-old reading 2 years below age level. Indeed, the level of reading skill representative of most 14-year-olds corresponds with the average literacy level of the United States and Canada. Phillips and Clarizio (1988) recommended this approach because it is possible to compare the academic performance of children who span different grades. In addition, achievement profiles are more easily translatable to educational recommendations. Finally, grade equivalents account for changes in within-grade variability across varying grade levels.

However, Spreen (1976) criticized grade-equivalent discrepancy methodology because dispersion of achievement test scores increases with advancing age (Reynolds, 1984; Salvia & Yesseldyke, 1981). Thus, an older child who performs 2 years below grade level may have a learning disability of similar severity as a younger child whose performance lags by 1 year. Regression between grade and test score is not equivalent across grades or even school subjects (Reynolds, 1984). In addition, Reynolds (1984) illustrated that grade-equivalent difference scores are based on the assumption that a constant rate of learning occurs across the entire school year. Many achievement tests are not administered at each month in the year but rather are extrapolated from both ends of the scale (Salvia & Yesseldyke, 1981). By relying on grade-equivalent discrepancies, small differences in achievement may be exaggerated (Reynolds, 1984).

Discrepancy Scores

IQ-based discrepancies are equally problematic. Two different approaches can be used. One approach simply establishes IQ and achievement cut-offs. If a child has "average" intelligence and reading scores that are below (for example) the 25th percentile, he/she can be considered reading-disabled. Another approach defines discrepancy according to relative levels of IQ and achievement. For example, a child who is reading 1 standard deviation below measured intelligence is considered eligible for special education in Texas and many other states and provinces. Why 1 standard deviation is used, as opposed to 1.5 or 2 standard deviations (comparable with other states and provinces), probably depends on issues involving policy (i.e., funding) and not on clinical characteristics of the child. Equally serious is the failure of these types of definitions to correct for regression artifact (see below).

Phillips and Clarizio (1988) cautioned that standard scores may possess many of the pitfalls associated with grade-equivalent scores. In particular, scaled score differences may not represent equal intervals because of the method of test construction. In fact, the Wide Range Achievement Test (Jastak & Jastak, 1965) derives both standard scores and centiles from grade equivalents converted directly from raw scores. Moreover, the assumption of normality of scores within age or grade groups may not be tenable. By using standard scores, normality may be forced, regardless of the underlying nature of the distribution of scores.

Lastly, Phillips and Clarizio (1988) note that, when differences between standard scores were used to define groups, unusual growth patterns of academic achievement occurred. For example, the Woodcock–Johnson Psycho-Educational Battery (Woodcock & Johnson, 1977) requires an elementary school child to gain 17.7 scaled score points to remain below the 10th centile, 19.2 points to continue to be in the average range (50th centile), and 20.2 points to stay in the superior range (90th centile). However, in high school, the necessary performance pattern changes. An increase of 3.5 scaled score points must have occurred for a student to have remained below average. Students who achieve at least above the 50th centile need to increase their score by at least 2.3 points, and those who had previously demonstrated superior performance had to gain at least 1.7 points to maintain such standing. Moreover, the performance pattern necessary to maintain centile ranking varied as a function of the test employed to assess achievement. Clearly, these findings warrant caution in interpretation of research based on standard score discrepancy definitions. Furthermore, such results necessitate both awareness of and familiarity with the properties of the individual test to be administered.

Regression to the Mean

In addition to objections to appropriate measurement differences, using discrepancy criteria based on comparisons of IQ and achievement scores raises statistical issues involving regression to the mean. If a difference score is formed on the basis of two measures that are neither perfectly correlated nor independent, the resulting distribution differs from the simple subtraction of the two component distributions (Cone & Wilson, 1981). If two measures are moderately correlated and an individual scores above the mean on the first test, on the average, the individual will not be expected to perform at that level or better on the second measure. Because performance on each test is not an independent event, the measures are correlated. Regression toward the mean occurs, in that it is more likely that the score the individual receives on the second test will be closer to the group mean than was the first test score. Correspondingly, the same effect occurs if performance is below the mean. In that case, the second test score is expected to be higher (more toward the mean) than the first score (McLeod, 1979; Yule, 1978).

Reynolds (1984) criticized the use of such discrepancy criteria precisely because regression artifacts are often ignored, especially given the relatively high intercorrelation between IQ and reading achievement. He found that using such comparisons leads to (1) an overidentification of higher-IQ-score children as disabled (in that some difference between IQ and achievement scores is expected) and (2) and underidentification of children with lower IQ scores (because the achievement score is expected to surpass low IQ).

In addition to the misallocation of special education services that accompanies psychometrically imprecise definitions, the use of uncorrected discrepancy criteria may be discriminatory. Often those children who perform more poorly on IQ measures come from educationally deprived environments, may be racially different from traditional reading-disabled samples, or may be of lower socioeconomic status. It is precisely these children who may require appropriate services to achieve successful reading outcomes.

The second approach to the operationalization of standard score discrepancy definitions is the use of regression procedures to predict the actual level of achievement that would be expected on the basis of age or intelligence, hence correcting for regression to the mean. Actual achievement is then compared to the predicted achievement score. If this difference exceeds what would be predicted given normal variation, a designation of reading disabled is indicated.

Rutter and Yule (1975) used regression-based definitional criteria to delineate a group of readers whose achievement was not commensurate with IQ and age (specific reading retardation; SRR). However, Siegel and Heaven (1986) criticized the use of IQ scores to predict reading ability. They argued that by using IQ as an estimate of potential, the same problems occur as when exclusionary definitions are used to define disabled readers (Fletcher & Morris, 1986). Briefly, this argument states that because reading and IQ test scores are positively correlated, the predicted reading score would be biased (in this case depressed) relative to a prediction based on an independent indication of learning potential. Although regression-based approaches do address the problem of regression artifacts that result when using raw score methods, such an approach is in no way a panacea. In general, the issues raised by the use of standard scores and grade-equivalent scores remain, including the comparability of measurement interval over time and the issue of basing assessments of potential on IQ scores.

Specific versus Backward Readers

Even when IQ and achievement scores are corrected for regression, it is not clear that children with discrepancies in IQ and achievement have more specific disabilities than do poor achievers whose IQ scores are not discrepant. Rutter and Yule (1975) defined children with achievement problems according to whether they were "backward readers," who read at IQ-appropriate levels, or "specific reading retarded," who have reading scores below expected levels according to their IQ scores. These designations were based on data derived from children between the ages of 9 and 11 on the Isle of Wight who were then reassessed at the age of 14. Children who scored 2 standard deviations or more below the group mean on nonverbal intelligence and reading attainment measures were subsequently administered a short form of the Wechsler Intelligence Scale for Children and the Neale Analysis of Reading. SRR was defined as an observed difference in reading achievement that was at least 2 standard errors below predicted levels via multiple regression analysis using age and IQ as predictors. "Backward" readers were those children whose reading was deficient on the basis of age alone, regardless of intelligence.

Rutter and Yule (1975) observed that those children identified as SRR differed from the backward readers along a series of measures including educational prognosis. First, there was a greater prevalence of males in the SRR group (76.7%) as compared to the backward readers (54.4%). In addition, the backward readers presented with a greater incidence of overt neurological dysfunction: 11.4% had evidence of "hard" neurological signs such as cerebral palsy, and 25.3% demonstrated evidence of "soft" neurological abnormalities such as developmental delay. None of the SRR children demonstrated "hard" neurological signs, and ony 18.6% showed the possibility of "soft" signs. Moreover, the backward readers were rated as more clumsy, showed more coordinational and constructional difficulties, and were more likely to demonstrate right-left confusion than were SRR children. Motor impersistence and choreiform movements were also more commonly found in the backward reader group. However, both SRR and backward reading groups had a similar incidence of family history of language problems, delayed speech milestones, and poor articulation.

When these children with reading difficulties were assessed as 14-year-olds, educational achievement varied as a function of skill area assessed. The SRR children demonstrated poorer spelling and reading skills than did the backward readers, indicating that they had fallen relatively further behind their age-matched peers. The arithmetic performance of the SRR children improved relative to the backward readers, but it remained significantly below grade levels. Rutter and Yule (1975) concluded that SRR is a relatively distinct disorder that can be summarized as a deficit that is peculiar to language, whereas backward readers demonstrated multiple difficulties in intellectual, neurological, and language areas.

These findings have not been uniformly replicated. Rodgers (1983) did not find evidence for a bimodal distribution of achievement in a large sample of 10-year-olds from Great Britain and Northern Ireland. The actual prevalence of disabled children with a greater than 2 standard deviation difference between actual achievement and that predicted by IQ was 2.29% compared to a predicted prevalence of 2.28%. Rodgers (1983) concluded that the distribution of reading achievement was distributed normally.

Other studies have addressed the original Rutter and Yule (1975) findings. Silva, McGhee, and Williams (1985) assessed 952 children from Dunedin, New Zealand, at 7 and 9 years of age. Children were divided into SRR and backward reader groups in a procedure identical to that used by Rutter and Yule (1975). They found that 74.4% of the backward readers were male, whereas 87.5% of the SRR group were male. Only the backward readers had significantly more neurological abnormalities than either the SRR or normal reader group. Furthermore, backward readers also demonstrated more motor difficulties than did any other group, although the SRR children showed more motor impairment than did normal readers. In contrast to Rutter and Yule (1975), Silva et al. (1985) also found significant differences between the two disabled reading groups on language measures, with the SRR group outperforming the backward readers. The SRR group did, however, achieve below levels of normal readers. The educational attainment was similar for both disabled groups in reading and spelling. Yet, the SRR group performed significantly better on arithmetic measures than did the backward readers, although they continued to achieve below levels of the normal reader group.

Jorm, Share, MacLean, and Matthews (1986) delineated groups of SRR and backward readers among 453 Australian children using the methods described above. They identified 14 retarded readers and 25 backward readers who were subsequently followed during the first three grades. In kindergarten these children were administered a neuropsychological battery consisting of diverse language, motor, and sensory measures. At grade 2 the children were given standardized achievement tests and were classified into diagnostic reading groups using procedures similar to those of Rutter and Yule (1975).

The SRR group differed significantly from backward readers in name writing and reading, letter copying, syntax, receptive vocabulary, sentence memory, and motor impairment. Backward readers differed from normal readers in almost all areas assessed, except motor impersistence, impulsivity, and pseudoword learning. The SRR group differed from normals only on specific language and early literacy skills such as name writing, recognition discrimination, picture and color naming, phoneme segmentation, and finger localization. Jorm et al. (1986) concluded that although the SRR and backward reader groups appeared similar in terms of reading ability, their cognitive competencies differed. SRR children had specific difficulties with language skills, whereas the backward readers had more global difficulties. Finally, they concluded that a common etiology of reading difficulty cannot be assumed for the two groups of disabled readers. It was more likely that the SRR group had academic difficulties because they were prevented from learning because of encoding difficulties (i.e., they were developmentally deviant). Backward readers demonstrated intact, yet reduced, general abilities and thus were considered to learn by a slower, yet normal, process (i.e., they were developmentally delayed).

van der Wissel and Zegers (1985) reviewed the Isle of Wight studies. Observing that there may have been a ceiling effect on the reading test employed, they performed simulation studies that they interpreted as showing that the so-called "hump" in the distribution of achievement test scores is a product of this ceiling effect as well as differences in gender ratios for the groups of backward and SRR children.

Yule (1985) responded to this study by noting that the definition of specific reading retardation used by van der Wissel and Zegers was based solely on a division of IQ scores. The original definitions used by Rutter and Yule (1975) were based across the IQ distribution. Yule also noted that van der Wissel and Zegers (1985) misinterpreted the nature of the reading test and the gender differences. Yule (1985) also disavowed any attempt to link IQ and reading achievement in a casual fashion, stating the "we are not arguing that IQ causes reading disorder, but that the use of our classification identified meaningful subgroups of poor readers" (p. 12).

Fletcher and Morris (1986) noted that the distinction between specific and backward reading is a classification hypothesis that should be subjected to systematic empirical investigation. Finding differences in neurobehavioral characteristics and educational progress supports the viability of the classification. However, there has been little uniformity in such findings.

In an earlier study that did not use regression-based definitions, Taylor, Satz, and Friel (1979) selected children with reading problems (Wide Range Achievement Test Reading below the 30th percentile) according to whether they met exclusionary definitions. They operationally defined the criteria provided by the World Federation of Neurology (Critchley, 1970) by specifying that a diagnosis of dyslexia could only occur in children with a Peabody Picture Vocabulary Test IQ greater than 89, average or above average socioeconomic status as rated by teachers, and an absence of neurological, sensory, or emotional difficulties as noted by teacher or parent. Those children who met the exclusionary criteria and who exhibited deficient reading skills were labeled "dyslexic" disabled readers, and those who did not (i.e., low socioeconomic status or IQ and low reading ability) were labeled "nondyslexic" disabled readers.

These two groups of disabled readers were compared against two groups of normal readers across seven different areas: neuropsychological and academic test performance, severity of reading problems, reversal and/or letter confusion, parental reading proficiency, neurological exam, and personality. Results showed no significant differences in any of the areas assessed between "dyslexic" and "nondyslexic" disabled reading groups of children. The two reading-disabled groups did, however, differ significantly from normal readers on measures in all seven domains. Furthermore, when IQ and socioeconomic variables were controlled, differences continued to be robust between disabled and nondisabled readers.

More recently, Share, McGhee, McKenzie, Williams, and Silva (1987) found no differences in prognosis between generally backward and specific reading-disabled children between 7 and 9 years of age. In addition, they noted that the types of differences found between children in this study and the Isle of Wight study had questionable causal relationships with reading impairment. Share et al. (1987) conlcuded that "on the basis of the data discussed here, there appears to be no firm evidence to support the validity of the distinctions between specific reading retardation and general reading backwardness" (p. 42).

In an investigation based on the Connecticut Longitudinal Study (Shaywitz & Shaywitz, 1988), Shaywitz, Shaywitz, Barnes, and Fletcher (1986) compared the influence of various definitions on the selection of children as learning disabled in an epidemiologic sample of school children in Connecticut. Although variations in the use of IQ indices and definitions resulted in different children being identified as learning disabled, few differences in cognitive ability were apparent among children grouped as learning disabled according to various definitions. There were also few differences among children defined as learning disabled whose scores were discrepant or not discrepant from IQ.

Limitations of Previous Studies

The varying findings of these studies undoubtedly reflect differences in samples and instruments. However, at this point, the major question is not so much whether there is a "hump" in the distribution of IQ–reading scores. Rather, the critical question is whether distinctions are valid between disabled readers whose reading ability is consistent with as opposed to inconsistent with measured intelligence.

The answer to this question appears to depend on how groups are defined, representing a classification problem (Fletcher, Francis, & Morris, 1988). It appears that when a more rigorous definition is used to form groups of disabled readers, group differences often emerge. Specifically, children who meet regression-based discrepancy criteria may be impaired on specific language measures when compared to normal readers (Jorm et al., 1986). Correspondingly, those children whose reading is incompatible with levels estimated by age but is consistent with that predicted by IQ are found to have global difficulties in functioning that span motor, neurological, and language domains.

It is not unexpected that smaller or nonsignificant differences were found by Taylor et al. (1979), because the IQ construct played a less prominent and clearly delineated role in the exclusionary selection criteria employed in that study. When IQ differences between disabled and nondisabled readers were controlled, skill differences between these groups defined by exclusionary criteria remained robust. However, IQ effects were neither considered nor controlled when comparing dyslexic and nondyslexic disabled reading groups, even though the dyslexic disabled childern were clearly of higher IQ. Such IQ differences could potentially mask inferior performance of the dyslexic disabled group. In addition, Taylor et al. (1979) used results from receptive vocabulary level as an estimate of IQ. Consequently, groups defined in this manner may not be comparable to those from research employing a more global measure to estimate intelligence. However, one advantage of the Taylor et al. (1979) study is the large sample size. Other comparisons of backward readers and specifically disabled readers have been hampered by small samples of disabled children derived from large epidemiologic samples.

An alternative approach to these issues is to use a large sample of clinically impaired children. A within-group approach will not address prevalence issues, but it can be used to address the validity of various definitions. In the remainder of this chapter, we discuss a series of three studies addressing the validity of discrepancy-based definitions of reading disability in a large cohort of learning-disabled children.

Comparisons of Various Definitions of Learning Disability

Sample

The children for this study were obtained from a data base of over 2,500 cases representing children referred for evaluation of learning disability in Windsor, Ontario. Each child received a comprehensive neuropsychological evaluation (Rourke, 1981; Rourke, Fisk, & Strang, 1986) along with the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949), and the Wide Range Achievement Test (WRAT; Jastak & Jastak, 1965). For this study, children were selected who ranged in age from 9 years to 14 years with WISC Full Scale IQ (FSIQ) scores above 70. These children were free of sensory, acquired neurological, and other problems traditionally used as exclusionary criteria. Application of these criteria resulted in a total sample of 1,069 children. The sample, 74% of whom were male, were predominantly white, middle-class children, who averaged 11 years, 4 months in age. The mean WRAT Reading standard score was 89.3 (*SD* = 14.8) with a mean WISC FSIQ of 98.5 (*SD* = 10.5).

Definitions

Two different definitions were used to identify children as reading disabled based on the word recognition score from the WRAT and the WISC FSIQ. The first definition employed a cutting-score approach that did not correct for regression artifact. Children were defined as reading disabled if their FSIQ exceeded 79 and their WRAT Reading standard score was below 93. In addition, children were categorized according to whether reading scores were consistent with or inconsistent with FSIQ using a criterion of 15 points: A child was considered "discrepant" if the WRAT score was less than the FSIQ by at least 15 points. This definition corresponds directly with criteria commonly used to define eligibility for special education services as a child with reading disability. Liberal criteria in terms of relative severity of word-recognition deficit and IQ were used in the hope of capturing the largest possible sample unbiased to selection variables.

Joint application of both definitions to all children resulted in four reading groups: normal readers (children not impaired according to either criterion); children impaired under both definitions (low achieving and discrepant); children who were low achievers (below 93) but not discrepant; and children whose reading was discrepant with IQ but exceeded a standard score of 93. In the above definitions, IQ–achievement discrepancy was based on observed standard score differences. Alternatively, this discrepancy can be based on regression formulas as previously discussed, where the discrepancy is between observed and predicted achievement.

Comparison Variables

To address whether differences in ability structure exist among groups formed with different definitional criteria, low-achieving children who were discrepant and not discrepant under the two definitions were compared on a set of tests derived from a modification of the Halstead-Reitan Neuropsychological Battery for Children (HRB; Rourke et al., 1986). These measures constitute a representative sample of neuropsychological skills and abilities and are ordinarily administered in a comprehensive evaluation of children with learning disabilities (Rourke, 1981). The linguistic and auditory-perceptual measures are especially sensitive to the reliable discrimination of children with reading disability from non-disabled children and from children with other types of learning disabilities (Rourke, 1978, 1981; Rourke et al., 1986).

Ten tests from the modified HRB were used. These tests are presented in Table 2.1 along with a summary of the constructs measured by each task. These constructs were defined according to maximum-likelihood factor analyses of the test battery in this sample recently completed by our group. It is apparent that these tests measure a variety of abilities frequently impaired in children with reading disabilities, including language, perceptual, and motor skills.

Comparison of Definitions: Study 1

Fletcher et al. (1989) provided a comparison of children in the sample grouped according to the joint application of the two definitions. Both unadjusted and regression-based approaches to defining IQ–achievement discrepancy were used separately. Joint application of raw score and discrepancy criteria produced four groups: children who scored above or below 92 on WRAT Reading and who had reading scores (regardless of

TABLE 2.1. Modified Halstead-Reitan Neuropsychological Tests by Factor Structure

Test	Factor
1. Category Test	Executive functions, spatial relations
2. Speech-Sounds Perception Test	General language, acoustic language
3. Auditory Closure Test	General language, acoustic language
4. Sentence Memory Test	General language, acoustic language
5. Verbal Fluency Test	General language, acoustic language
6. Finger-Tapping Test	Simple motor
7. Grooved Pegboard Test	Eye-hand coordination, spatial relations
8. Tactual Performance Test	Spatial relations, executive function, eye-hand
9. Trail Making Test, Parts A and B	Executive function
10. Target Test	Spatial relations, eye-hand

level) that were discrepant or not discrepant with WISC FSIQ. To simplify discussion, unadjusted comparisons of discrepancies between observed IQ and achievement are described as "uncorrected" discrepancies. In contrast, "regression-based" discrepancies (i.e., differences between observed and predicted achievement) are described as "corrected" discrepancies because they adjust for the IQ–achievement correlation.

Tables 2.2 and 2.3 summarize classifications of the 1,069 children for definitions uncorrected (Table 2.2) or corrected for (Table 2.3) the correlation of WRAT Reading and WISC FSIQ. For the uncorrected definition, Table 2.2 shows that there is a small group of children ($n = 36$) with reading standard scores greater than 92 whose reading score is at least 15 points below their FSIQ. The other children are distributed fairly evenly across the 2 × 2 matrix. About 34% have reading standard scores below 90 that are at least 15 points below their FSIQ scores, with 30% regarded as not impaired in reading. Some 32% of the children have poor reading, but, because of their FSIQ, would not qualify for special education services.

Table 2.3 presents the resultant 2 × 2 matrix when regression artifact is accounted for in the definition of an IQ–achievement discrepancy. It is apparent that the distribution of children across the four categories is different from that in Table 2.2. More children are identified as discrepant and fewer as nondiscrepant. In terms of overlap between the two definitions, 70 children (7%) with reading standard scores below 93 become eligible for services using the regression-based definition who were not eligible under the cut-off score definition. In general, these children scored lower on FSIQ ($M = 87.7$; $SD = 4.5$) and WRAT Reading ($M = 75.5$; $SD = 3.4$) than did the group of discrepant readers identified in the first analysis. However, 22 (2%) children who had reading standard scores below 93 and 15-point discrepancies between IQ and reading were no longer eligible under the regression-based definition. These children had a mean WISC FSIQ of 106.2 ($SD = 3.0$) and mean WRAT Reading standard score of 89.7 ($SD = 3.0$). Of the 36 higher-but-discrepant-achievement children eligible under the uncorrected discrepancy score criterion,

TABLE 2.2. Means and Standard Deviations for Full Scale IQ, and Reading Standard Scores of Children Categorized According to Reading Standard Scores and Raw Discrepancies

	Reading standard score	
	≤ 92	> 92

Discrepant	$N = 360$ (34%)
	FSIQ[a] = 101.8 (8.7)
	RdSS[b] = 78.4 (8.0)
		$N = 36$ (4%)
		FSIQ = 114.9 (5.3)
		RdSS = 96.1 (5.1)

Not discrepant	$N = 347$ (32%)
	FSIQ = 90.6 (6.0)
	RdSS = 83.9 (5.7)
		$N = 326$ (30%)
		FSIQ = 101.3 (10.9)
		RdSS = 106.4 (12.2)

[a] FSIQ, Full Scale IQ on WISC.

[b] RdSS, Reading standard score on WRAT.

TABLE 2.3. Means and Standard Deviations for Full Scale IQ, and Reading Standard Scores for Children Categorized According to Reading Standard Scores and Regression-Based Discrepancies

	Reading standard score	
	≤ 92	> 92

Discrepant	$N = 408$ (38%)
	FSIQ[a] = 99.2 (9.8)
	RdSS[b] = 77.3 (7.2)
		$N = 5$ (1%)
		FSIQ = 99.2 (9.8)
		RdSS = 90.2 (9.1)

Not discrepant	$N = 299$ (28%)
	FSIQ = 92.4 (7.1)
	RdSS = 86.2 (4.2)
		$N = 357$ (33%)
		FSIQ = 102.5 (11.0)
		RdSS = 105.6 (12.0)

[a] FSIQ, Full Scale IQ on WISC.

[b] RdSS, Reading standard score on WRAT.

only five remain eligible when the definition takes into account regression effects. Thus, the regression-based criteria make 70 "low-achieving readers" eligible but eliminate eligibility for 22 children with word recognition scores below 93, because these scores are within the range expected given their IQ. These criteria also eliminate eligibility for 31 children who exhibit age-appropriate word recognition scores (*M* = 97.1; *SD* = 3.6) and higher FSIQ (*M* = 114.6; *SD* = 4.2).

Overlap: Study 2

One problem with the Fletcher et al. (1989) study is the failure to account for overlap in the classifications in Tables 2.2 and 2.3. In other words, some children meet (or do not meet) both discrepancy-based definitions, whereas other children meet criteria set forth by only one discrepancy-based definition. To address the issue of overlap, Espy, Francis, Fletcher, and Rourke (1989) compared three groups of children who were "disabled readers" according to various discrepancy-based definitions: (1) both raw score and regression-based discrepancy definitions (IBOTH); (2) only raw score definitions (IRAW); and (3) only regression-based definition (IREG). All of these children met discrepancy-based definitions. Children who did not meet a discrepancy-based definition were placed into a single group regardless of reading level on the assumption that these children describe a continuum of reading impairment.

When the sample was divided in this fashion, 291 children met both definitions (IBOTH), 105 met only uncorrected definitions (IRAW), and 22 met only regression-based definitions (IREG). There were 651 children who met neither discrepancy-based definition of reading disability. Note that of these 651 children, 325 (49.9%) could be considered as low-achieving (WRAT Reading ≤ 92) children.

The mean WISC Verbal IQ and Performance IQ scores, and WRAT Reading, Spelling, and Arithmetic scores of these four groups are presented in Table 2.4. Among the more striking findings illustrated in this table are the differences in Full Scale IQ across the four groups. The IRAW group has the highest FSIQ (*M* = 106.7), whereas the IREG group has the lowest FSIQ (*M* = 86.0). In the three disabled groups, achievement and WISC scores fluctuate similarly, reflecting the relationship of IQ scores and WRAT achievement scores. Given this finding and the standard error of measurement associated with the WISC and the WRAT, another implication is that there is substantial skill and ability overlap among the three disabled groups. Such IQ differences among the three groups are a natural consequence of the two definitions and the fact that IQ and achievement are correlated.

TABLE 2.4. Scores on WISC and WRAT Variables for Children Categorized According to Raw and Regression-Based Discrepancies

	Group[a]							
Variable	ND (*N* = 651)		IBOTH (*N* = 291)		IRAW (*N* = 105)		IREG (*N* = 22)	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
WISC								
Verbal IQ	93.7	10.9	94.1	9.3	99.8	9.3	83.4	5.8
Performance IQ	99.4	11.6	109.8	12.1	113.1	7.9	91.4	5.8
Full Scale IQ	93.1	10.2	101.7	9.7	106.7	7.1	86.0	2.8
WRAT								
Reading	39.0	27.6	8.1	7.4	25.6	13.3	3.8	1.2
Spelling	25.6	22.5	6.4	6.6	16.5	11.7	3.6	1.9
Arithmetic	20.7	15.4	13.6	12.1	23.3	13.4	7.5	4.0

[a] ND, not discrepant; IBOTH, discrepant using raw score and regression-based criteria; IRAW, discrepant using raw score criteria only; IREG, discrepant using regression-based criteria only.

ABILITY PROFILES

Espy et al. (1989) performed a series of analyses to examine the magnitude of group differences and the role of IQ scores as differentiators of the four groups. To facilitate these comparisons, groups were compared on the set of 10 neuropsychological variables used as external validation measures (see Table 2.1). These comparisons were treated as a set of classification hypotheses. If it is valid to classify children as LD or not using either type of discrepancy-based definition, then robust differences on these external variables should emerge. Espy et al. (1989) performed these comparisons using general multivariate analysis of variance (MANOVA). However, an alternative to MANOVA in this situation is profile analysis. In the remainder of this chapter we provide an introduction to profile analysis as an alternative to MANOVA through an extended demonstration involving the four reading groups and the 10 neuropsychological measures referred to earlier.

Profile Analysis: Study 3

The most frequently asked questions in neuropsychological research often involve comparisons of two or more groups on multiple measures of neuropsychological functioning. There are several ways to address such questions in a statistical manner. The least informative approach, and the one most difficult to justify statistically (Huberty & Morris, 1989), is that of isolated multiple univariate comparisons. Yet this approach is chosen more frequently than any other, in part because researchers find multivariate alternatives difficult to carry out and interpret. Profile analysis (PA) represents an attractive multivariate alternative for neuropsychologists because PA directly compares "patterns" of group test performance and is easily performed and interpreted. PA accomplishes this pattern comparison by separating differences among groups and differences among measures into three statistically independent pieces of information. These pieces of information are referred to as the dimensions of (1) shape, (2) elevation, and (3) flatness. The remainder of this chapter describes these three questions addressed in PA, and it compares PA with traditional univariate and multivariate alternatives for examining mean group differences. Finally data from the preceding discussion of reading disability definitions are used to demonstrate PA.

Much of the neuropsychological research focuses on the comparison of performance patterns in two or more groups. Indeed, in the 5 years from 1983 to 1988, no fewer than 83 of the articles in the *Journal of Clinical and Experimental Neuropsychology* (about 50%) involved such a comparison as the primary research question. PA is a conceptually simple procedure that directly examines differences in performance patterns. Surprisingly, PA has not seen widespread application in neuropsychology. In fact, only one of the 83 articles mentioned above formally applied PA, whereas 54 used some form of univariate statistic. Although these points were made previously (Francis, Fletcher, & Davidson, 1988), the potential applications of PA in neuropsychology so greatly outnumber the actual applications in the literature that we feel these points are worth repeating.

THE THREE DIMENSIONS OF PROFILE ANALYSIS

To facilitate the following presentation and discussion of PA, consider the patterns of means displayed in Figure 2.1. Each circle in Figure 2.1 corresponds to the mean *T*-score for one of four reading groups on 1 of 10 neuropsychological measures. The four reading groups were determined by the adjusted and unadjusted discrepancy-based criteria discussed previously. For each group, adjacent means have been connected by a straight line only to increase the visual impression of an ability "profile"; the lines should not be taken to imply that the horizontal dimension in the figure is continuous.

Roughly speaking, the variables have been ordered along the horizontal axis such that motor measures fall further to the left end, with spatial and verbal measures represented progressively further to the right.

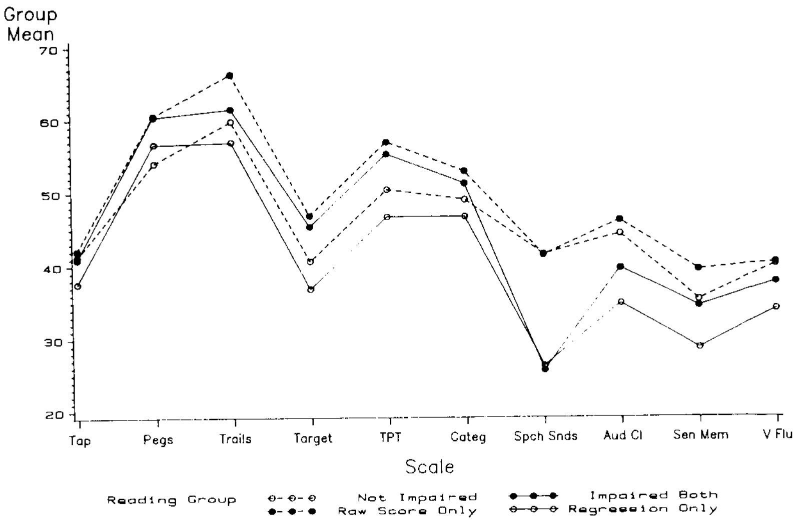

FIGURE 2.1. Mean profiles for four reading groups using 10 measures of neuropsychological performance. Criteria for forming reading groups are provided in the text. All measures have been transformed to *T*-scores ($M = 50$) using published norms and scaled such that higher scores indicate better performance. Abbreviations for measures are as follows: Tap, Finger Tapping, average *T*-score for left and right hands; Pegs, Grooved Pegboard, average *T*-score for left and right hands; Trails, average *T*-score for Trail Making A and B; Target, Target Test; TPT, Tactual Performance Test, average *T*-score for times using left, right, and both hands; Categ, Category Test, total score; Spch Snds, Speech-Sounds Perception; Aud Cl, Auditory Closure; Sen Mem, Sentence Memory; V Flu, Verbal Fluency.

This ordering is strictly arbitrary and was chosen as a matter of convenience to facilitate interpretation of the profiles in Figure 2.1. We return to the issue of variable ordering in due course.

In examining any set of group profiles, such as that of Figure 2.1, three possible questions come to mind. First, we must consider whether the population mean profiles are similarly shaped in the sense that the lines joining adjacent means are parallel for all groups. This is referred to as the "pattern," "shape," or "parallelism" hypothesis. An examination of Figure 2.1 suggests that the hypothesis of parallel profiles is likely to be rejected for these groups on these measures, especially in light of the large sample sizes in three of the four groups.

If one were to conclude that the profiles were parallel, it then becomes reasonable to consider two other hypotheses. First, do the profiles differ in level? This question is commonly referred to as the "levels" or "elevation" hypothesis. As an example of parallel profiles differing in elevation, mean profiles on the four primary verbal scales from the WISC (Wechsler, 1949) are presented in Figure 2.2 for three of the four reading groups. Clearly, in Figure 2.2, the most striking difference among the three profiles is their respective elevation. Second, if one were to conclude that the profiles were parallel, regardless of any possible differences in elevation, it also becomes reasonable to consider whether the population means for the different measures are equal. If the population means for all measures making up the profile were equal, then clearly the average profile (averaging across groups) would be flat. Not surprisingly, this final test is referred to as the "flatness" hypothesis.

From the foregoing discussion, it is clear that PA is quite different from the general MANOVA. MANOVA has been termed an unstructured multivariate analysis (Hand & Taylor, 1987) because MANOVA simply seeks to determine the way in which to combine a set of measures such that group differences on the measures are maximized. PA, in contrast, represents one form of structured multivariate analysis because PA seeks to determine answers to three specific questions concerning differences among groups and measures (Hand & Taylor, 1987), each of which can be conceptualized as a specific set of contrasts. Let us consider each question in somewhat greater detail in order to demonstrate how each can be conceptualized as a set of contrasts among groups and measures.

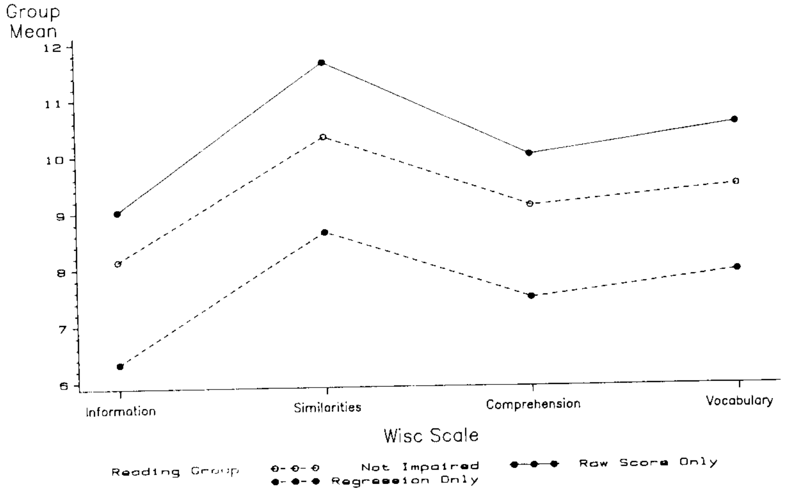

FIGURE 2.2. Mean profiles for three of four reading groups using primary verbal subscales from the WISC. Criteria for forming reading groups are provided in the text. Groups and measures were chosen to demonstrate the appearance of parallel profiles that differ in elevation.

PARALLELISM: THE TEST OF EQUAL SHAPE

The primary question in PA is the question of shape. If profiles do not have the same shape, then questions about elevation and flatness are somewhat meaningless because their answers would depend on which subset of measures or groups was being considered. The precedence of shape over elevation and flatness is analogous to the primary status afforded interactions over main effects in the analysis of variance (ANOVA)—main effects being unambiguous only in the absence of interactions. Actually, PA is directly analogous to the split-plot analysis of variance, representing designs with at least one between- and one within-subjects factor, a point we return to in a moment.

There are two equally legitimate ways to conceptualize the question of shape:

1. Are differences between measures the same for all groups?
2. Are differences between groups the same on all measures?

Consider again the profiles of Figure 2.1. In examining these, the first way of framing the shape question amounts to asking if the line segments joining any adjacent pair of measures have the same slope for all groups. For example, do the four line segments joining the means for Finger Tapping and Grooved Pegboard have equal slope? Similarly, are the slopes of the line segments joining Trails and Grooved Pegboard equal? If the line segments joining a given pair of measures have the same slope in all groups, then, clearly, the group profiles are parallel between those two measures. With p measures it is possible to compute $p - 1$ such linearly independent slopes for each group. For the entire group profiles to be parallel, each of these $p - 1$ slopes would have to be equal for all groups, although it is not necessary for all $p - 1$ slopes to be equal to one another. In the present example, there would be nine such slopes in each group, and it would appear that at least one group differs from the others on all but two of the nine slopes. Of course, we will want to know if these deviations from parallelism are statistically significant or if they could reasonably have resulted from random fluctuations in the data.

In truth, we refer to slopes because it is easy to think of slopes being parallel. However, the line segment slopes referred to are equivalent to differences between the means of adjacent measures. Hence, if we simply compute the difference between adjacent measures for each group, variability between groups in these computed mean differences indicates a lack of parallelism in group profiles. When the number of measures in the profile is three or more, the test of parallelism will involve two or more difference scores, and hence the test will be multivariate in nature.

Using the data presented in Figure 2.1 for the four reading groups, the test of parallelism yields a significant multivariate $F(27,3177) = 10.88$, $p < .0001$. Hence, for these groups we would reject the hypothesis that the profiles are parallel, and we would begin the process of further delineating just how the group profiles differ. Such follow-up questions may involve repeating the profile analysis for a subset of the groups or examining the flatness hypothesis for individual groups. For example, we may desire to know if profiles are parallel for the IQ-adjusted (regression) only and the unadjusted only groups or if the unimpaired group has a flat profile. Alternatively, we may want to know about group differences on specific measures (e.g., do the four groups differ in mean Finger Tapping?). In short, follow-up analyses may entail within-group analyses, between-groups contrasts on individual measures, or interactions between contrasts involving measures and contrasts involving groups, such as repeating the profile analysis for a subset of the groups.

ELEVATION: THE TEST OF EQUAL MEAN PROFILE HEIGHT AVERAGING OVER MEASURES

The question of equal profile elevation is only addressed if we are willing to conclude that the profiles have the same shape. Clearly, if the profiles do not have the same shape, then it is ambiguous to discuss differences in profile elevation because these differences vary at different points in the profile. The test of equal elevation is the most straightforward of the three tests because it represents a simple between-groups comparison on a single variable. That variable is the group grand mean for the set of measures, that is, the average for each group across the set of p measures. This process of aggregating over the p measures is justified because the existence of parallel profiles implies that differences between groups in profile elevation are the same on all measures. Therefore, averaging across the measures to obtain the group grand means will yield the best estimate of the magnitude of the differences among the groups in profile elevation.

Carrying out a PA on the data in Figure 2.2, we would not reject the parallelism hypothesis, $F(6,1548) = 0.6465$, $p = .69$. However, clear differences exist in mean profile elevation, $F(2,775) = 24.24$, $p < .0001$. At this point we might wish to examine these elevation differences further by performing contrasts between the groups (e.g., pairwise comparisons between the unadjusted only group and the unimpaired group). Regardless of any follow-up tests that might be performed to examine the elevation hypothesis, it is also possible to examine the question of profile flatness.

FLATNESS: THE TEST OF EQUAL MEAN PERFORMANCE ON ALL MEASURES AVERAGING ACROSS GROUPS

The flatness question is also addressed only in the absence of profile shape differences. Similar to the shape question, flatness is tested by obtaining the mean difference between pairs of adjacent measures. However, in contrast to the shape hypothesis, where slope segments are formed separately for each group, in evaluating the flatness hypothesis, these slope segments are formed by first averaging across groups to obtain the average performance on each measure. Differences are then formed between the grand means for adjacent measures. If the average profile is flat, then the set of $p - 1$ such differences will not be statistically different for 0. As in the test of shape, the test of flatness is a multivariate test whenever p is three or more. Using the data displayed in Figure 2.2 once again, the test of flatness is statistically significant, $F(3,773) = 53.0927$, $p < .0001$, indicating that the average profile among these three groups is *not* flat. This result is not surprising given the large sample sizes in these groups and a visual inspection of Figure 2.2. Having rejected the overall flatness hypothesis, we may wish to perform a more restrictive test of flatness by considering whether the average profile is flat for a subset of the measures (e.g., Information and Vocabulary).

RELATIONS TO MANOVA AND REPEATED-MEASURES ANOVA

As suggested previously, the standard MANOVA represents an unstructured multivariate hypothesis because it fails to distinguish between two types of information about group differences. Specifically, MANOVA combines between-groups differences in profile shape and between groups differences in average profile height into a single effect—group differences on the set of measures. PA separates this information into statistically independent pieces of Elevation and Shape. In addition, PA provides a test of the Flatness of the combined group profile.

The relationship of PA to repeated-measures ANOVA is even more straightforward. Consider a repeated-measures design with one between-subjects factor and one within-subjects factor. Such a design is frequently employed in neuropsychological investigations and is often referred to as a split-plot design. The Elevation hypothesis of PA is exactly the test of the between-subjects main effect in a split-plot design. The Flatness hypothesis of PA equals the test of the main effect of the within-subjects factor. Finally, the Shape hypothesis of PA equals the test of interaction between the within- and between- subjects factors in the split-plot design.

The direct correspondence between PA and the split-plot repeated-measures ANOVA brings up several important considerations. First, readers may or may not be aware that two general approaches exist for analyzing repeated-measures data, regardless of whether the design includes a between-subjects factor. These two general approaches are referred to as the "univariate" and "multivariate" approaches to repeated-measures ANOVA. Given the correspondence between PA and repeated measures, it is not surprising that both univariate and multivariate approaches can also be applied in PA, with the factors governing choice of method being identical in both PA and repeated-measures ANOVA.

We have focused our description on the multivariate approach to PA because we feel this approach is more easily justified statistically in light of the fact that assumptions for the analysis are more likely to be met. We advocate the same approach in analyzing standard repeated-measures designs, and the reader is strongly cautioned that the univariate approach should not be employed in PA or repeated-measures ANOVA unless adjusted for statistical dependence among observations (Maxwell & Arvey, 1982; O'Brien & Kaiser, 1985).

A second point that requires some clarification given the correspondence between PA and repeated-measures ANOVA is the importance of variable ordering in the profiles. The descriptions of the Shape and Flatness hypotheses given above seem to suggest that variable ordering is important. After all, differences are computed between adjacent measures. Obviously, the magnitudes of these differences will depend on which measures are chosen to be adjacent to one another. Yet we have stated that the Shape and Flatness hypotheses of PA correspond to the interaction of the between- and within-subjects factors, and the within-subjects main effect, respectively, in a split-plot design. Furthermore, statistical significance of these latter effects is not altered by reordering the levels of the within-subjects factor at the time of analysis. In fact, the same can be shown to be true for PA—ordering of the measures in the profile will not affect the overall tests of significance, that is, the overall tests of shape and flatness. Clearly, ordering is unimportant for the elevation hypothesis because all measures are averaged to form the grand mean for each group.

In essence, there is a fixed amount of variability attributable to these hypotheses. This total can always be fully explained by a set of $p - 1$ linearly independent single-degree-of-freedom effects (or contrasts). In other words, any set of $p - 1$ linearly independent differences between the p means will completely account for this total effect. Changing the set of single-degree-of-freedom contrasts leaves the total effect unchanged, provided a complete set of $p - 1$ linearly independent contrasts is specified. The set of differences between adjacent measures in a profile is just such a set of $p - 1$ linearly independent contrasts. Put simply, the ordering of variables in a PA is irrelevant to the overall tests of shape, elevation, and flatness.

ASSUMPTIONS

The assumptions underlying the multivariate approach to PA are the same as those for MANOVA: (1) measures must follow a multivariate normal distribution; (2) observations on different subjects must be independent of one another; and (3) variance/covariance matrices must be equal across groups. In addition to these standard MANOVA assumptions, PA requires commensurability of measures. If variables are measured on different metrics (i.e., variables are not commensurable), then shape differences may be an artifact of scale.

This point is most easily understood when one considers the second formulation provided above for the shape hypothesis: namely, parallel profiles imply that differences between groups are equal for all measures. Consider an obvious case where measures are not commensurable. Suppose, for example, that an experimenter has data for two groups on two measures. The measures are high school grade point average (GPA) and junior year Scholastic Aptitude Test (SAT) scores. The range of possible mean differences between groups in GPA is +4 to −4, whereas mean differences between groups in SAT scores have a much wider possible range. One would not expect GPA differences between groups to equal SAT differences in magnitude. Consequently, in their original metric, it would make little sense to conduct a PA using GPA and SAT as profile measures because shape differences would reflect artifactual scale differences.

A viable solution when profile measures are not commensurable in their original metrics is to transform the scores to z- or T-scores prior to profile analysis. This yields commensurate scales; however, the metric of the data analysis is changed, and the researcher must determine if the new metric is an acceptable one. We feel that z- or T-score transformation in neuropsychological work is reasonable because most meaures lack a clearly defined metric or one that has more inherent value than the standard deviation metric of T-scores. It is important that any such transformation make use of external information for standardization.

A less obvious example of noncommensurability can be provided by neuropsychological data. Many measures collected by neuropsychologists are scored in a positive direction. Other measures are scored negatively insofar as higher scores indicate poorer performance. The WISC subscales are obvious examples of positively scored scales, whereas the Trail Making Tests, Grooved Pegboard, and time scores of the Tactual Performance Test are examples of negatively scored scales. A PA using both positively and negatively scored scales will almost certainly lead to rejection of the parallelism hypotheses. In this situation, parallel profiles would imply that the best performing group on a positively scored measure is equally the poorest performing group on a negatively scored measure. Obviously, for positively and negatively scored measures to be included in a single PA, one set of measures must be transformed so that the scoring is reversed. For the analyses presented in Figure 2.1, all measures have been scored so that higher scores indicate better performance.

The most important point of the preceding discussion is that PA results are not invariant under transformation of the data. Hence, the choice of metric for analysis must be defensible, and all measures must be scaled similarly. It is our opinion that transformation to normal scores will work well in most neuropsychological applications for reasons just cited and because of the interest in neuropsychological research with distinguishing normative from nonnormative performance.

CONCLUDING COMMENTS ON PROFILE ANALYSIS

The preceding treatment of PA was intentionally nontechnical and relatively narrow in that it dealt only with the primary questions of shape, elevation, and flatness. Interested readers will find more detailed treatment of PA in standard multivariate texts such as Bernstein, Garbin, and Teng (1988), Stevens (1986), Harris (1985), and Morrison (1990), and in the general literature on research methods (e.g., Maxwell & Arvey, 1982). Of the textbooks mentioned, Harris (1985) and Morrison (1990) offer the most complete coverage of PA, but they are also the most mathematically demanding. Stevens (1986) is the most introductory, dealing only with statistical comparison of group profiles as we have done here. Bernstein et al. (1988) is less mathematically demanding than Harris (1985) and Morrison (1990), but it does not deal with statistical inference in the application of PA. Rather, Bernstein et al. presents an overview and discussion of various measures of profile similarity as part of a more general treatment of classification methods. This material is relevant for measuring similarity between group profiles, between an individual's profile and that of a group, or between the profiles of two individuals. The relevance of this latter aspect of PA to research in clinical neuropsychology has been ably discussed and demonstrated by Chelune and his colleagues (Chelune, Heaton, Lehman, & Robinson, 1979; Chelune & Moehle, 1986; Lehman, Chelune, & Heaton, 1979). For that reason, and because this aspect of PA primarily concerns the classification of individuals into known groups, we have chosen to focus on the statistical comparison of mean profiles as an alternative to MANOVA, reflecting the focus on group definition in the preceding section of the chapter.

Conclusions Regarding Discrepancy-Based Definitions of Reading Disability

In the preceding section, we saw that four groups of children meeting different criteria for reading disability differ significantly from one another in their patterns of neuropsychological test performance. When examining the scales from the Halstead–Reitan Battery (Reitan & Davison, 1974) and other measures, clear differences in profile shape emerged. Although visual inspection of profiles suggests that these differences are small in magnitude, it must be kept in mind that these ability profiles are displayed in a standard score metric. With this fact in mind, the most striking characteristic of these profiles is the markedly poor performance of the two regression-based groups on measures heavily influenced by phonological analysis, such as Speech-Sounds Perception and Auditory Closure. As such, these results are reminiscent of Jorm et al. (1986). A second striking characteristic of these profiles is the generally superior performance of the group meeting only the unadjusted discrepancy score criterion. On examination of verbal skill profiles for the regression only, unadjusted score only, and unimpaired groups, clear differences in profile elevation emerged, with the unadjusted score only group showing clearly superior verbal skills. It is not surprising that the unadjusted score only group outperforms the remaining groups on these verbal scales from the WISC (Wechsler, 1949). The very nature of the IQ–achievement discrepancy score criterion dictates that high-IQ children will be identified by this definition when the discrepancy is not adjusted for the correlation between IQ and achievement. It should also be noted that the group showing the poorest performance in these verbal scale profiles is not eligible for special services in many states and provinces because those jurisdictions do not adjust for the correlation between achievement and IQ.

Much work remains to be done in delineating ability differences between and among groups meeting specific criteria for definitions of learning disability. The data presented here suggest that agencies responsible for dealing with the implications of such definitions need to recast their respective nets in establishing diagnostic criteria for determining eligibility for services. There are many reasons that might be put forward for eliminating discrepancy-based criteria for determining eligibility for services. However, it is clear that, if discrepancy-based criteria are to be used, these criteria should adjust for the correlation between IQ and achievement.

Acknowledgments

Supported in part by NICHD Grant HD-21888, Psycholinguistic and Biological Mechanisms in Dyslexia. Portions of this chapter were presented at the 1988 meeting of the American Education Research Association, New Orleans, LA, April 6, 1988.

References

Bernstein, I. H., Garbin, C. P., & Teng, G. K. (1988). *Applied multivariate analysis.* New York: Springer.

Chelune, G. J., Heaton, R. K., Lehman, R. A., & Robinson, A. (1979). Level versus pattern of neuropsychological performance among schizophrenic and diffusely brain-damaged patients. *Journal of Consulting and Clincial Psychology, 47,* 155–163.

Chelune, G. J., & Moehle, K. A. (1986). Neuropsychological assessment and everyday functioning. In. D. Wedding, A. M. Horton, Jr., & J. Webster (Eds.), *The neuropsychology handbook: Behavioral and clinical perspectives* (pp. 489–525). New York: Springer.

Cone, T. E., & Wilson, L. R. (1981). Quantifying a severe discrepancy: A critical analysis. *Learning Disability Quarterly, 4,* 359–371.

Critchley, M. (1970). *The dyslexic child.* Springfield, IL: Charles C. Thomas.

Doehring, D. G. (1978). On the tangled web of behavioral research on developmental dyslexia. In A. L. Benton & D. Pearl (Eds.), *Dyslexia: An appraisal of current research* (pp. 125–135). New York: Oxford University Press.

Espy, K. A., Francis, D. J., Fletcher, J. M., & Rourke, B. P. (1989). Implications of raw cut-off score and regression-based definitions of reading disability. *Journal of Clinical and Experimental Neuropsychology, 11,* 31–32.

Fletcher, J. M., Espy, K. A., Francis, D. J., Davidson, K. E., Rourke, B. P., & Shaywitz, S. E. (1989). Comparisons of cut-off and regression-based definitions of reading disabilities. *Journal of Learning Disabilities, 22,* 334–338.

Fletcher, J. M., Francis, D. J., & Morris, R. (1988). Methodological issues in neuropsychology: Classification, measurement, and the comparison of non-equivalent groups. In F. Boller & J. Grafman (Eds.), *Handbook of neuropsychology* (Vol. I, pp. 83–110). Amsterdam: Elsevier.

Fletcher, J. M., & Morris, R. (1986). Classification of disabled learning: Beyond exclusionary definitions. In S. Ceci (Ed.), *Handbook of cognitive, social, and neuropsychological aspects of learning disabilities* (Vol. I, pp. 55–80). New York: Lawrence Erlbaum.

Francis, D. J., Fletcher, J. M., & Davidson, K. C. (1988, February). *Profile analysis and the study of patterns of neuropsychological performance.* Paper presented at the annual meeting of the International Neuropsychological Society, New Orleans, LA.

Hand, D. J., & Taylor, C. C. (1987). *Multivariate analysis of variance and repeated measures: A practical approach for behavioral scientists.* London: Chapman & Hall.

Harris, R. J. (1985). *Multivariate statistics* (2nd ed.). Orlando: Academic Press.

Huberty, C. J., & Morris, J. D. (1989). Multivariate analysis versus multiple univariate analyses. *Psychological Bulletin, 105,* 302–308.

Jastak, J. L., & Jastak, S. R. (1965). *Wide Range Achievement Test.* Wilmington, DE: Guidance Associates.

Jorm, A. F., Share, D. L., MacLean, R., & Matthews, R. (1986). Cognitive factors at school entry predictive of specific reading retardation and general reading backwardness: A research note. *Journal of Child Psychology and Psychiatry, 27,* 45–54.

Kaufman, A. S. (1979). *Intelligent testing with the WISC-R.* New York: Academic Press.

Kavanaugh, D., & Gray, D. (Eds.) (1986). *Biobehavioral measures of dyslexia.* Parkton, MD: York Press.

Lehman, R. A., Chelune, G. J., & Heaton, R. K. (1979). Level and variability of performance on neuropsychological tests. *Journal of Clinical Psychology, 35,* 358–363.

Maxwell, S. E. & Arvey, R. D. (1982). Small sample profile analysis with many variables. *Psychological Bulletin, 92,* 778–785.

McLeod, J. (1979). Educational underachievement: Toward a defensible psychometric definition. *Journal of Learning Disabilities, 12,* 42–50.

Morris, R. D. (1988). Classification of learning disabilities: Old problems and new approaches. *Journal of Consulting and Clinical Psychology, 56,* 789–794.

Morris, R., & Fletcher, J. M. (1988). Classification in neuropsychology: A theoretical framework and research paradigm. *Journal of Clinical and Experimental Neuropsychology, 10,* 640–658.

Morrison, D. F. (1990). *Multivariate statistical methods* (3rd ed.). New York: McGraw-Hill.

O'Brien, R. G., & Kaiser, M. D. (1985). MANOVA method for analyzing repeated measures designs: An extensive primer. *Psychological Bulletin, 97,* 316–333.

Phillips, S. E., & Clarizio, H. F. (1988). Limitations of standard scores in individual achievement testing. *Educational Measurement: Issues and Practice, 7,* 8–15.

Reitan, R. M., & Davison, L. A. (Eds.). (1974). *Clinical neuropsychology: Current status and applications.* New York: John Wiley & Sons.

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

Rodgers, B. (1983). The identification and prevalence of specific reading retardation. *British Journal of Education Psychology, 53,* 369–373.

Rosner, J., & Simon, D. P. (1971). The auditory analysis test: An initial report. *Journal of Learning Disabilities, 4,* 40–48.

Rourke, B. P. (1978). Reading, spelling, arithmetic disabilities: A neuropsychologic perspective. In H. R. Myklebust (Ed.), *Progress in learning disabilities* (Vol. 4, pp. 97–120). New York: Grune & Stratton.

Rourke, B. P. (1981). Neuropsychological assessment of children with learning disabilities. In S. B. Filskov & T. J. Boll (Eds.), *Handbook of clinical neuropsychology* (pp. 453–478). New York: Wiley-Interscience.

Rourke, B. P., Fisk, J. L., & Strang, J. D. (1986). *Neuropsychological assessment of children.* New York: Guilford Press.

Rutter, M. (1978). Prevalence and types of learning disabilities. In A. L. Benton & D. Pearl (Eds.), *Dyslexia: An appraisal of current research* (pp. 3–29). New York: Oxford University Press.

Rutter, M., & Yule, W. (1975). The concept of specific reading retardation. *Journal of Child Psychology and Psychiatry, 16,* 181–197.

Salvia, J., & Ysseldyke, J. (1981). *Assessment in special and remedial education* (2nd ed.). Boston: Houghton Mifflin.

Sattler, J. M. (1988). *Assessment of children's intelligence and special abilities* (2nd ed.). Philadelphia: W. B. Saunders.

Satz, P., & Fletcher, J. M. (1980). Minimal brain dysfunctions: An appraisal of research concepts and methods. In H. E. Rie & E. D. Rie (Eds.), *Handbook of minimal brain dysfunctions: A critical review* (pp. 669-714). New York: John Wiley & Sons.

Share, D. L., McGhee, R., McKenzie, D., Williams, S., & Silva, P. D. (1987). Further evidence relating to the distinction between specific reading retardation and general reading backwardness. *British Journal of Developmental Psychology, 5,* 35-44.

Shaywitz, S. E., & Shaywitz, B. A. (1988). Afftention deficit disorder: Current perspectives. In J. F. Kavanagh & T. J. Truss (Eds.), *Learning disabilities* (pp. 369-523). Parkton, MD: York Press.

Shaywitz, S. E., Shaywitz, B. A., Barnes, M., & Fletcher, J. M. (1986, October). *Prevalence of dyslexia in a epidemiological sample.* Paper presented at the meeting of the Child Neurology Society, Halifax, Nova Scotia, Canada.

Siegel, L. S., & Heaven, R. K. (1986). Categorization of learning disabilities. In S. Ceci (Ed.), *Handbook of cognitive, social and neuropsychological aspects of learning disabilities* (Vol. I, pp. 95-121). Hillsdale, NJ: Lawrence Erlbaum.

Silva, P. A., McGhee, R., & Williams, S. (1985). Some characteristics of nine year old boys with general reading backwardness or specific reading retardation. *Journal of Child Psychology and Psychiatry, 20,* 407-421.

Spearman, C. E. (1923). *The nature of intelligence and the principles of cognition.* London: Macmillan.

Spreen, O. (1976). Neuropsychology of learning disabilities: Post conference review. In R. M. Knights & D. J. Bakker (Eds.), *The neuropsychology of learning disabilities* (pp. 445-468). Baltimore: University Park Press.

Stevens, J. (1986). *Applied multivariate statistics for the social sciences.* Hillsdale, NJ: Lawrence Erlbaum.

Still, G. F. (1902). Some abnormal psychological conditions in children. *Lancet, 1,* 1077-1082.

Strauss, A. A., & Lehtinen, L. E. (1947). *Psychopathology and education of the brain-injured child.* New York: Grune & Stratton.

Taylor, H. G., Fletcher, J. M., & Satz, P. (1984). Neuropsychological assessment of children. In G. Goldstein & M. Hersen (Eds.), *Handbook of psychological assessment* (pp. 211-234). New York: Pergamon Press.

Taylor, H. G., Satz, P., & Friel, J. (1979). Developmental dyslexia in relation to other childhood reading disorders: Significance and clinical utility. *Reading Research Quarterly, 15,* 84-101.

van der Wissel, A., & Zegers, F. E. (1985). Reading retardation revisited. *British Journal of Developmental Psychology, 3,* 3-9.

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

Woodcock, R. W., & Johnson, M. B. (1977). *Woodcock-Johnson Psycho-Educational Battery Achievement Test.* Hingham, MA: Teaching Resources Corporation.

Yule, W. (1978). Diagnosis: Developmental psychological assessment. In A. G. Kalverboer, H. M. van Praag, & J. Mendlewicz (Eds.), *Advances in biological psychiatry: Vol. 1, Minimal brain dysfunction: Fact or fiction* (pp. 1-49). Basel: S. Karger.

Yule, W. (1985). Response to van der Wissel and Zegers. *British Journal of Developmental Psychology, 3,* 11-13.

CHAPTER 3

Methodological and Statistical Issues in Cluster Analysis

John W. DeLuca, Kenneth M. Adams, and Byron P. Rourke

This chapter raises some important methodological and statistical issues with respect to the use of numerical taxometric methods, in particular cluster analysis. For example, misconceptions regarding cluster methodology, specifically the notion of "internal validity" are noted. Although determination of the reliability and validity of a cluster solution remains a paramount aspect of the classification process, there continues to exist some confusion surrounding definitions. In addition, there are several issues regarding the use of the two-stage cluster procedure that are vague, misleading, or simply in error. Such problems include the following: percentage relocated as a measure of "stability" and an indicator of "internal validity" (reliability); determination of the "best" starting position for the iterative relocation process; and clarification of what constitutes a true two-stage cluster procedure. In this chapter we review these critical issues as they pertain to the use of cluster analysis in classification research.

The role of classification research is growing rapidly within the field of neuropsychology. Nowhere is it more evident than in the application of cluster analytic techniques to the study of learning-disabled children. Several investigators (e.g., Del Dotto & Rourke, 1985; DeLuca, Rourke, & Del Dotto, Chapter 10, this volume; Fletcher & Morris, 1986; Fuerst, Fisk, & Rourke, 1989; Lyon, Stewart, & Freedman, 1982; Morris, Blashfield, & Satz, 1986) have identified various subtypes of children based on either academic, neuropsychological, or personality dimensions. The work of these and other authors has provided some indication of the reliability and validity of the resulting subtypes. However, the emphasis on this phase of the classification process in the general neuropsychological literature had been less than optimal.