The Pennsylvania State University

The Graduate School

College of Education

LONG-TERM STABILITY OF MEMBERSHIP IN WISC-III SUBTEST AND FACTOR SCORE CORE PROFILE TAXONOMIES

A Thesis in

School Psychology

by

Ellen R. Borsuk

Submitted in Partial Fulfillment
of the Requirements
for the Degree of

Doctor of Philosophy

August 2005

The thesis of Ellen R. Borsuk was reviewed and approved* by the following:

Marley W. Watkins
Professor of Education
In Charge of Graduate Programs in School Psychology
Thesis Adviser
Chair of Committee

Barbara A. Schaefer
Associate Professor of Education

Pamela S. Wolfe
Associate Professor of Special Education

Janice C. Light
Professor of Communication Sciences and Disorders

*Signatures are on file in the Graduate School.

ABSTRACT

Although often applied in practice, cognitive subtest profile analysis has failed to achieve empirical support. Nonlinear multivariate profile analysis may have benefits over clinically based techniques, but the psychometric properties of these methods must be studied prior to their interpretation and use. The current study posed the following question: Is WISC-III cluster membership based on nonlinear multivariate subtest and factor profile analysis stable over a 3-year period? Membership stability to the subtest and factor taxonomies, including constancy of displaying an unusual profile, was based on data from 579 and 177 students, respectively. General and partial kappa coefficients either failed to reach statistical significance or indicated poor classification stability, with the exception of two profile types. It was concluded that, with two possible exceptions, profile-type membership to empirically derived subtest and factor WISC-III taxonomies cannot be used in educational decision-making. Directions for future research and limitations of this study were considered.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

ACKNOWLEDGMENTS

I would like to express my gratitude to my thesis adviser, Dr. Marley Watkins, the other members of my doctoral committee, Ronn Walvick, and Aaron Borsuk. Without the insight and support of the above mentioned people, this work would not have been possible.

Long-Term Stability of Membership in WISC-III Subtest and Factor Score Core Profile Taxonomies

Over 5.7 million students in the United States between the ages of 6 and 21 received special education services during the 2000-01 school year (U.S. Department of Education [USDOE], 2001). Many students benefit from special education. A review by Forness (2001), for example, identified a number of effective special education interventions (e.g., mnemonic strategies, direct instruction). On the other hand, students who qualify for special education services but who do not receive such support are at a disadvantage as they cannot benefit from those effective special education interventions. Further, those erroneously identified as qualifying for special education are rendered a serious disservice. For example, removal from the general education classroom in order to receive special education services is thought to interfere with instruction (Friend & Bursuck, 2002). Thus, the importance of making sound decisions for students with respect to qualification for services becomes obvious.

In addition to diagnosis, other educational decisions for students must also be made with utmost care. Instructional methods, materials used during teaching, and classroom environment all play an important role in student outcome and, thus, must be given careful consideration. For example, results of a meta-analysis conducted by the National Reading Panel (National Institute of Child Health and Human Development, 2000) revealed that phonemic awareness instruction was effective in improving phonemic awareness, reading, and spelling, and was most successful when coupled with certain methodological and environmental variables. Teaching phoneme manipulation with letters in an overt and systematic manner, emphasizing only up to two methods of

phoneme manipulation, and small-group instruction resulted in the largest effects. Certain instructional materials are also effective when teaching students. Rieth and Semmel's (1991) review of the literature called attention to the promise of the appropriate use of computer-assisted instruction in the classroom with students with whom teachers are experiencing difficulties. Finally, classroom environment can also influence educational outcome. For example, seating arrangement may have an effect on student and teacher behavior (e.g., Ridling, 1994). It is evident, then, that poor decisions regarding instructional techniques, teaching materials, and classroom environment can negatively impact student outcome.

Educational decisions made for students involving either eligibility for special education services or instructional planning (i.e., methods, materials, or environment) have important implications and, thus, should be made with care. These choices should be based on data, such as test scores, shown to be useful in the educational decision-making process. Given that reliability of test scores is necessary, though not sufficient, for valid interpretation and use of these scores (American Educational Research Association [AERA], American Psychological Association [APA], and National Council on Measurement in Education [NCME], 1999), investigation of score reliability, including stability over time, is crucial.

Wechsler Series Tests as a Frequently Employed Tool in Educational Decision-Making

Tests of intelligence are often an important component of the assessment conducted to make educational decisions for students. Further, one of the major roles of the school psychologist is that of assessor, and one of the major components of assessment is cognitive assessment (Alfonso & Pratt, 1997). Intelligence can be and has been defined in

a number of ways (Sattler, 2001). Although variations in the definition exist, often thought to be part of the construct of intelligence are "attributes such as *adaptation to the environment, basic mental processes,* and *higher-order thinking* (e.g., *reasoning, problem solving,* and *decision making*)" (Sattler, 2001, p. 135). For example, David Wechsler (1944) described intelligence as the ability "to act purposefully, to think rationally, and to deal effectively with his or her environment" (p. 3). In addition, Wechsler perceived intelligence as a collection of abilities, rather than as a single aptitude (Wechsler, 1991).

The Wechsler test series is often used by school psychologists to assess intellectual functioning (Kamphaus, Petoskey, & Rowe, 2000; Sparrow & Davis, 2000). Not only are Wechsler series tests favored among school psychologists, but they are an important component of the training and practice of clinical psychologists (Belter & Piotrowski, 2001; Watkins, Campbell, Nieberding, & Hallmark, 1995). Remarkably, it has been noted that millions of students have been given the Wechsler Intelligence Scale for Children-Third Edition (WISC-III; Wechsler, 1991) when being assessed to determine entitlement to special education (Watkins & Canivez, 2004).

A survey conducted by Alfonso, Oakland, LaRocca, and Spanakos (2000) found that the Wechsler series was frequently taught in school psychology training programs, partly due to perceived frequency of clinician use. Ninety-two percent of school psychology courses in individual cognitive assessment required students to complete one or more WISC-III protocols, and 90% had students complete written reports on this measure. The Wechsler series is the most commonly taught of the traditional tests in school psychology cognitive assessment classes, according to survey respondents (Alfonso et al.).

Training is thought to be a good indicator of future practice (Alfonso et al., 2000). Consistent with this prediction, school psychologists have a history of frequent use of tests belonging to the Wechsler series, including the Wechsler Intelligence Scale for Children-Revised (WISC-R; Wechsler, 1974), the Wechsler Adult Intelligence Scale-Revised (WAIS-R; Wechsler, 1981), and the WISC-III (Alfonso & Pratt, 1997). Additionally, results of a survey of 354 school psychologists indicated that the WISC-III was very commonly used, with 65% of respondents administering the instrument at least twice weekly on average (Pfeiffer, Reddy, Kletzel, Schmelzer, & Boyer, 2000). Given how widely established the Wechsler series have become in the fields of clinical and school psychology, it is likely that the fourth edition of the Wechsler Intelligence Scale for Children (WISC-IV; Wechsler, 2003a, 2003b) will enjoy continued popularity.

Stability of Global Wechsler Series Test Scores

Given the frequency of use of Wechsler series tests among psychologists for making crucial decisions about students, it is vital to determine whether clinicians are making sound decisions based on obtained scores from these measures. Because intelligence is thought to remain relatively stable over time for a child of at least 5 years of age (Sattler, 2001), professionals are inclined to make long term decisions for students based on test results. For example, the WISC-III was seen as being helpful for diagnostic purposes as well as for placement decisions by the school psychologists surveyed by Pfeiffer et al. (2000). In addition, educational decisions made for students have tended to be long-term given that, under the Individuals with Disabilities Education Act Amendments of 1997 (IDEA-97), students with identified disabilities could be re-evaluated as infrequently as once every 3 years. Combined with the tenet that reliability is a prerequisite for validity

(AERA, APA, & NCME, 1999), it is critical to determine whether scores from Wechsler series tests remain stable over time. Further, special emphasis should be given to WISC-III score stability given how widespread this measure became among school psychologists and, therefore, its likely continued popularity in the form of the WISC-IV.

A review by Canivez and Watkins (1998) revealed that test-retest reliability coefficients have been repeatedly found to be moderate to high during investigations of both short- and long-term stability of the Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) and WISC-R IQ scores. For example, correlation coefficients ranging from .74 to .84 were found between FSIQ scores at different age levels for an unselected birth cohort (n = 794) tested longitudinally on the WISC-R at ages 7, 9, 11, and 13 (Moffitt, Caspi, Harkness, & Silva, 1993).

A lesser amount of research has been conducted supporting the short- and long-term stability of WISC-III IQ and factor scores (Canivez & Watkins, 1998). As a result, Canivez and Watkins (1998) explored the long-term stability of WISC-III scores. Participants were 667 students with an average test-retest interval of 2.83 years. The majority of participants had disabilities. Results revealed that stability coefficients were in the upper .80s and lower .90s for Verbal IQ (VIQ), Performance IQ (PIQ), Full Scale IQ (FSIQ), Verbal Comprehension Index (VC), and Perceptual Organization Index (PO) scores. Mean IQ and index scores did not change significantly over time, with the exception of VIQ scores. The mean VIQ difference over time of only .64 points was determined not clinically meaningful (Canivez & Watkins, 1998). The authors concluded that long-term stability of these WISC-III scores sufficed for individual diagnostic purposes. However, they cautioned that examination of group means constitutes a

nomothetic outlook and that, despite group trends, individual scores may fluctuate significantly over time. In fact, IQ and index scores did not remain stable for many individual cases and only FSIQ scores were fairly stable for most students. Canivez and Watkins (1999) found analogous results across ethnicity (Caucasian, Hispanic/Latino, and Black/African American), gender, and age (6 to 13 years). Findings also remained constant across disability (learning disability [LD], serious emotional disability, and mental retardation), although slightly lower stability coefficients were found (*r*s mainly low to mid .80s with the exception of FSIQ, which ranged from high .80s to low .90s; Canivez & Watkins, 2001).

In addition to being stable, there is support for the utility of global, or overall, intelligence scores. For example, based on a review of the literature, Glutting, McDermott, Konold, Snelbaker, and Watkins (1998) concluded that there is strong evidence to support the use of global intelligence scores for making predictions regarding school achievement, occupational success, and other significant variables. Further, they are integral to contemporary diagnosis of LD and mental retardation (Reschly, 1997). On the other hand, global intelligence scores are not useful for intervention planning (Gresham & Witt, 1997).

Beyond Global Wechsler Scores: Popularity of Profile Analysis for Educational Decision-Making

WISC-III IQ scores and some of the index scores show diagnostically adequate stability coefficients as well as utility for predictive purposes; however, many clinicians go beyond these global scores and apply profile analysis to subtest and index scores in order to make intervention decisions. Sattler (2001) noted that "profile analysis aims to

describe the child's unique ability pattern and, in so doing, go beyond the information contained within the FSIQ" (p. 299). Profile analysis refers to the determination of cognitive strengths and weaknesses in order to come to decisions regarding diagnosis and treatment (Glutting et al., 1998). That is, practitioners use profile analysis of cognitive test scores in order to make eligibility decisions as well as to generate hypotheses about a child's cognitive skills and deficits that can be used to guide intervention. Watkins and Kush (1994) noted, however, that the absence of empirical support shifted the focus of profile analysis from diagnosis to identification of intellectual strengths and weaknesses, which can in turn be used to direct treatment.

About 89% of the sample of school psychologists surveyed by Pfeiffer et al. (2000) reported that they used index scores and/or subtest profile analysis. Further, when asked what they found to be most useful about the WISC-III, about 70% of respondents reported that they valued factor scores and/or profile analysis. This was the most popular response. Further, 29% of the sample found individual subtests to be useful. On the other hand, a minority of respondents (18%) perceived various aspects of profile analysis as depicted in the WISC-III manual to be undesirable.

Additionally, according to a survey by Alfonso et al. (2000), 89% of school psychology training programs used *Assessment of Children's Intelligence and Special Abilities* (3rd ed. Rev.; Sattler, 1992), and 29% used *Intelligent Testing with the WISC-III* (Kaufman, 1994) as texts for individual cognitive assessment courses. These texts promote profile analysis and offer guidelines for its application. For example, Sattler (1992) noted that although profile analysis with the WISC-R, Wechsler Preschool and Primary Scale of Intelligence (WPPSI; Wechsler, 1967), and WAIS-R is not useful for

making diagnostic decisions, it is still useful for assessing cognitive strengths and weaknesses and for prescribing treatment. Although Sattler's book has been updated (Sattler, 2001), it continues to promote similar guidelines for the WISC-III, Wechsler Preschool and Primary Scale of Intelligence-Revised (WPPSI-R; Wechsler, 1989), Wechsler Adult Intelligence Scale-Third Edition (WAIS-III; Wechsler, 1997), and other modern intelligence tests.

Similar to Sattler (1992, 2001), Kaufman's (1994) text encourages clinicians to make both short and long term educational decisions for students based, in part, on cognitive profile interpretations. Hypotheses derived from systematic interpretation of WISC-III results, in combination with other information, should lead clinicians to make decisions regarding instructional styles, teaching materials, and instructional environment (Kaufman). Through illustrative case studies, Kaufman demonstrated the integration of information derived from profile analysis of the WISC-III with background, achievement, and other relevant information in order to arrive at educational and behavioral recommendations, including incorporation of diagrams into instruction, placement in a structured learning environment, and gifted education.

Even the WISC-III manual supports the practice of profile analysis (Wechsler, 1991). The WISC-III manual implicitly endorses the use of profile analysis in making classification decisions by stating that "intersubtest *scatter* is the variability of an individual's scaled scores across the subtests. Such variability is frequently considered as diagnostically significant" (p. 177). Further, the WISC-III manual concurs with Kaufman (1994) by advocating the importance of integrating WISC-III scores with other applicable information, such as background information and performance on other tests, when

interpreting WISC-III results. Like Sattler (1992, 2001) and Kaufman, the WISC-III manual outlines procedures for conducting profile analysis; the WISC-IV continues to provide similar guidelines.

Due to the popularity of profile analysis with WISC-III scores, especially its use for long-term decision making for students, it is critical to determine whether profiles remain stable over time. If WISC-III profiles are not stable over time, clinicians who use profiles to make important diagnostic and treatment decisions may be making unsound educational choices for students. The nature of a profile and clinically based methods of profile analysis will be discussed first; the stability of nonlinear multivariate profile type membership will ultimately be considered.

Components of a Profile

Profiles can be defined as an examinee's set of scores on a given assessment occasion, such as an examinee's WISC-III scores, where the *elements* of the profile would be subtest scores, index scores, and the like (Livingston, Jennings, Reynolds, & Gray, 2003). A Profile has three dimensions: elevation, scatter, and shape (Cronbach & Gleser, 1953).

Elevation

Profile elevation is the level of an examinee's profile, or the mean element score (Cronbach & Gleser, 1953). In addition, the level of various subtests or other, more global scores can be considered in isolation. These scores are normative in that they are indicative of an examinee's performance compared to a standardization group.

Scatter

Scatter is a measure of dispersion. As such, traditional measures of dispersion including the range, variance, and standard deviation have often been used in the calculation of scatter. For example, scatter can be defined as the square root of the sum of square difference scores between each element score and the mean, a multiple of the standard deviation (Cronbach & Gleser, 1953). Similarly, Plake, Reynolds, and Gutkin (1981) suggested measuring scatter with the profile variability index (PVI). Calculation of the PVI involves inserting subtest scores into the formula used to calculate variance. A large value of PVI is indicative of significant scatter within the more global scale (McLean, Reynolds, & Kaufman, 1990). Comparison to base rates is thought to allow for interpretation of PVI scores (McLean et al.). Plake et al. advocated the use of the PVI because it incorporates information from all subtests into its calculation.

In addition, scatter is frequently operationalized by calculating the range between an examinee's highest and lowest subtest standard scores (Konold, Glutting, McDermott, Kush, & Watkins, 1999). This number is then compared to the percentage of students in the normative sample who have a difference of at least this magnitude in order to determine whether the examinee's discrepancy is rare.

Methods of computing scatter that diverge from traditional measures of dispersion have also been suggested. A common method for determining scatter is identification of the number of subtests that deviate from the mean by a predetermined quantity, such as 3 points (Watkins & Glutting, 2000). Statistical significance can also be used to identify subtests that differ from the mean (Sattler, 2001). However, the element mean is not always considered in calculation of scatter. For example, Konold et al. (1999) noted that

scatter analysis can be conducted by calculating whether the difference between scores is statistically significant. The magnitude of this difference can then be examined for its frequency within the general population (e.g., Kaufman & Lichtenberger, 2000).

Shape

In addition to elevation and scatter, information about shape can be gleaned from WISC-III profiles. The shape of a profile is the residual data in the profile once elevation and scatter information have been removed (Cronbach & Gleser, 1953). Shape can be described as an examinee's unique patterns of high and low element scores on a given test (Watkins & Glutting, 2000). Given that elements are deemed to be high or low relative to an examinee's own mean, shape measurement, represented by a series of scores indicating the number of standard score points between an examinee's mean and each subtest score, is ipsative. This is in contrast to normative measurement where a given score tells of an examinee's performance relative to a group. For example, two examinees will have the same ipsative score on the Coding subtest if they both scored 2 standard points above their respective means on this subtest (i.e., +2); however, the first student may have a mean subtest score of 15 and a Coding score of 17, while the second student has a mean score of 6 and a Coding score of 8. Their normative Coding scores are very different, while their ipsative scores on this subtest are identical. Table 1 displays two students' WISC-III subtest scores showing identical ipsative profiles, but widely discrepant normative scores.

Table 1

Ipsative and Normative WISC-III Subtest Scores for Two Students

Score type	Mean	PC	IN	CD	SM
		Student 1			
Normative	15	18	11	17	14
Ipsative	0	+3	-4	+2	-1
		Student 2			
Normative	6	9	2	8	5
Ipsative	0	+3	-4	+2	-1

Note. PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities.

Clinically Based Profile Analysis Methods with the WISC-III

Many methods of profile analysis are clinically based rather than empirically derived. Examples of clinically based methods can be found by examining popular systems of WISC-III profile analysis (Kaufman, 1994; Kaufman & Lichtenberger, 2000). These systems discuss the relevance of IQ, index, and subtest score scatter in the interpretation of results. The clinician is taught to first consider the more global scores, given their superior reliability; however, analysis of scatter within these global scores is

thought to be necessary in order to determine whether the global score in question represents a unified and, therefore, meaningful construct, or whether narrower scores (e.g., index scores) are more cohesive and, as such, better represent the examinee's ability (Kaufman & Lichtenberger). In the final steps of their WISC-III interpretive guidelines, Kaufman and Lichtenberger lead the clinician through determination of subtests that represent significant strengths or weaknesses. This final scatter analysis is followed by an interpretation of the profile shape.

Examination of the shape of a given WISC-III profile is thought to provide insight about the examinee's underlying set of abilities (Kaufman & Lichtenberger, 2000). Although not empirically based, over 75 subtest variation patterns across the WISC, WISC-R, and WISC-III have been described (Glutting, McDermott, & Konold, 1997). These patterns are frequently used by clinicians to generate hypotheses (Glutting, McDermott, & Konold). Similarly, Glutting et al. (1998) noted that over 100 subtest patterns and their interpretations exist for Wechsler series tests and other individual intelligence tests for children. For example, the presence of an ACID (characterized by poor scores on the **A**rithmetic, **C**oding, **I**nformation, and **D**igit Span subtests) or SCAD profile (poor performance on the **S**ymbol Search, **C**oding, **A**rithmetic, and **D**igit Span subtests) on the WISC-III is thought to provide insight on a child's intellectual abilities (Kaufman & Lichtenberger). Although these profiles were originally thought to be helpful in the diagnosis of LD, reviews of the literature have found that these profiles are not useful for differential diagnosis, even though they appear to be more prevalent in groups of children with LD and other disabilities (Kaufman & Lichtenberger; Watkins, 2003).

Subtest patterns thought to be amenable to interpretation have also appeared in the literature in the form of subtest recategorizations. That is, WISC-III subtests are rearranged, and are no longer classified into the IQ and index scores found in the WISC-III manual. One popular way to reorganize WISC-III subtest scores is Bannatyne's system (Bannatyne, 1968). Recategorization of the WISC-III subtests in this way is thought to provide the clinician with an awareness that would not be possible from examination of only the IQ, index, and subtest scores outlined in the WISC-III manual (Kaufman, 1994). This should enhance the examiner's understanding of student abilities. Similar to other subtest trends that have been described, the Bannatyne system is based on clinical experience and is not firmly grounded in research or theory.

Kaufman and Lichtenberger (2000) provided a table of abilities, such as attention span, long term memory, and social comprehension, thought to underlie various groups of WISC-III subtests, although they noted that this listing is not finite and that practitioners may add to it. They also listed abilities believed to underlie individual subtests. However, they emphasized that the clinician should consider several subtests together, and advocated interpreting subtest scores in isolation only as a last resort. There are guidelines given to make decisions about which abilities likely underlie the strengths and weaknesses evident in an examinee's profile (Kaufman & Lichtenberger). Again, these hypotheses of the correspondence between subtests and various abilities have little empirical support and, instead, are based on clinical experience. For example, Kamphaus (1998) stated that "most of the presumed abilities that are offered for WISC-III interpretation are just that: Presumptions that are not supported by a preponderance of

scientific evidence" (p. 45) and "the number of untested hypothesized abilities is far larger than the list of tested ones" (p. 45).

Fundamental Difficulties with Reliance on Clinically Based Profile Analysis Methods

Basing profile analysis techniques on clinical judgment methods is not scientifically sound. Practitioner judgment regarding diagnosis and treatment is subject to error (Davidow & Levinson, 1993; Faust, 1986; Spengler, Strohmer, Dixon, & Shivy, 1995). Clinicians often rely on heuristics that serve as cognitive shortcuts, but that may lead to erroneous conclusions. For example, reliance on the representative heuristic results in ignoring base rates and, instead, in considering only existing knowledge, which may not match reality. That is, a clinician may believe that children with disabilities tend to posses a certain characteristic. However, it may be that a high percentage of all children, both with and without disabilities, display this characteristic, but that the clinician rarely has contact with children who are not disabled. As such, a child displaying this characteristic may be erroneously diagnosed based on the clinician's judgment.

Further, it is recommended by supporters of clinical methods of profile analysis that clinicians integrate data obtained from formal testing with other relevant knowledge in a manner consistent with his or her theoretical perspective in order to arrive at educational decisions for students (Kaufman & Lichtenberger, 2000). That is, Kaufman and Lichtenberger advocated the importance of considering guidelines regarding profile shape in the context of other clinical information (i.e., background information, observations, results of other testing) when deciding whether to accept or reject hypotheses about students' abilities. Educational planning, based on generated hypotheses, should also not result from WISC-III results alone (Kaufman, 1994). Further,

Kaufman and Lichtenberger stated that "to best interpret any particular piece of clinical evidence, each psychologist must use his or her own theoretical framework" (p. 90). Unfortunately, this is neglectful of the research that speaks to the inability of practitioners to accurately integrate clinical data in order to arrive at meaningful results (e.g., Faust, 1986).

Additional Limitations of Clinically Based Profile Analysis Methods in the Interpretation of WISC-III Scores

There are many other difficulties associated with the use of clinically based profile analysis techniques. These include low reliability of subtest scores, significant scatter as a frequent occurrence, failure to employ multivariate techniques, difficulty with reliance on ipsative scores, and fundamental research errors.

Low Reliability of Subtest Scores

The reliability of cognitive subtest scores is weak. More global IQ scores demonstrate better reliability. For example, internal consistency reliability coefficients were calculated for the WISC-III normative sample subtest, factor, and IQ scores (Wechsler, 1991). The mean subtest reliability coefficient across age groups was found to range from .69 (Object Assembly) to .87 (Vocabulary and Block Design). In contrast, those for the IQ scores ranged from .91 to .96, and those for factor scores were found to be between .85 and .94.

In terms of test-retest reliability, the same trend was found (Wechsler, 1991). In a study with a short retest interval (range = 12 to 63 days; median = 23 days), stability coefficients for WISC-III subtests ranged from .57 (Mazes) to .89 (Vocabulary). Across composite scores, stability coefficients were found to range between .82 (FD) and .94

(VIQ and FSIQ). Some studies have investigated stability coefficients when a long period of time elapsed between test and retest. For example, Livingston et al. (2003) conducted research involving 60 participants with academic and behavior problems. Participants were administered the WISC-R twice with an average test-retest interval of 3.09 years. Test-retest reliability coefficients indicated that normative subtest score reliability coefficients (ranging from .53 to .76) were lower than index and IQ score reliability coefficients (ranging from .80 to .85).

Similar results were found by Canivez and Watkins (1998) using a sample of 667 students between kindergarten and Grade 11. Participants, most of who had disabilities, were tested twice with the WISC-III an average of 2.83 years apart. Canivez and Watkins (1998) found test-retest reliability coefficients for subtests ranged from .55 (Symbol Search) to .78 (Block Design). These values were lower than those found for WISC-III composite scores (excluding FD and PS), which ranged from .85 (VC) to .91 (FSIQ). The authors concluded that while temporal stability coefficients of composite scores are strong enough to be used to make diagnostic decisions for individuals, this is not true for subtest coefficients.

A correlation coefficient with a magnitude of a least .90 is recommended when making important decisions about students based on results of a given test (Salvia & Ysseldyke, 2001). As such, it is apparent that while composite intellectual measures may be useful when making important educational decisions, subtest scores are too unstable for this purpose.

Significant Scatter as a Frequent Occurrence

Although scatter is often established by determining the number of subtests whose scores diverge from the child's mean by either a statistically significant degree or by a specified amount, or by calculating whether the difference between scores is statistically significant, significant differences are a frequent occurrence and may not indicate that a child has a problem (Konold et al., 1999). It is important to recall that significance is influenced by sample size, uncontrolled variance, and amount of time elapsed between treatments (Glutting, McDermott, & Konold, 1997). So, for example, with 2,200 children in the WISC-III standardization sample, it is expected that significant differences will be found between scores (Glutting, McDermott, & Konold). Further, given the large sample size, the magnitude of the significant score difference may be trivial. In one study, 55.6% of the WISC-III standardization sample (n = 2,200) were found to have at least one subtest score that was statistically significantly lower ($p < .05$) than their individual mean score, and 46.7% had at least two subtest scores that were significantly different from their means in either a positive or negative direction (Glutting, McDermott, & Konold).

Failure to Employ Multivariate Techniques

Profile analysis calls for multiple simultaneous and interrelated comparisons (Konold et al., 1999). As discussed, profiles contain information related to elevation, scatter, and shape (Cronbach & Gleser, 1953). This information cannot be fully considered unless multiple dependent comparisons are conducted simultaneously. However, popular clinical profile analysis techniques do not employ these multiple comparisons. For example, when calculating the range between the highest and lowest subtest score only one difference score results; hence, this procedure is univariate

(Glutting, McDermott, Watkins, Kush, & Konold, 1997). Univariate techniques involve only one comparison at a time and are not appropriate for profile analysis (Glutting, McDermott, Watkins, et al.).

Difficulty with Use of Ipsative Scores

One of the biggest difficulties related to commonly employed scatter and shape measurement relates to the fact that obtained scores are ipsative. That is, scatter and shape scores are generally only interpretable in an intra-individual sense. For example, a given score on a specific subtest indicates varying amounts of deviation, or scatter, depending on the mean score of the examinee. Also, the shape of a profile is defined through determination of subtest scores as high or low depending on the examinee's mean subtest score (Watkins & Glutting, 2000). Each subtest is given an ipsatized score, or a score that communicates the examinee's performance on this particular subtest relative to his or her average performance. McDermott, Fantuzzo, Glutting, Watkins, and Baggaley (1992) noted that ipsatized scores have been found to be problematic in many respects.

McDermott et al. (1992) discussed the common misconception of ipsative scores being as theoretically and statistically sound as global, normative measures of ability. When transformation takes place from normative to ipsative scores, each person's scores are changed by their unique average score. That is, each examinee's score is adjusted by a different amount. As such, ipsative scores are not comparable between individuals, even though a main purpose of intelligence testing is to determine a person's ability in comparison to others. Normative scores are able to provide information on relative ability

as well as ability across profile scores. Hence, normative scores provide all the information contained in ipsative scores plus additional, relevant data.

Accordingly, when McDermott et al. (1992) compared ipsatized WISC-R scores with their normative counterparts, many important differences were found. Using the 2,200 children in the WISC-R standardization sample, dramatic decreases were seen in the average intercorrelation once subtest scores were ipsatized (from .42 to -.09). Similarly, the average correlation coefficient between general ability and subtest scores decreased from .69 to .02 once ipsatization took place. These near-zero correlation coefficients as well as the fact that most intercorrelations were negative resulted in skepticism regarding the usefulness of ipsatized scores in the measurement of general ability (McDermott et al., 1992). That is, as discussed by Konold et al. (1999), ipsatization of scores results in the correlation between subtest scores being disregarded.

Unlike normative scores, ipsative measures fail to take global intelligence into account. McDermott, Fantuzzo, and Glutting (1990) noted that ipsatization of scores has the effect of removing common variance, or *g*, in the case of intelligence measures. For example, ipsatization of students' scores resulted in the elimination of almost 60% of the WISC-R's reliable variance. Similarly, McDermott et al. (1992) reported a removal of about 55% of the reliable variance found in WISC-R test scores upon ipsatization. McDermott et al. (1992) noted that the "evidence indicates that ipsative scores do not measure the same constructs conveyed by conventional ability scores, and it is not known what constructs they do measure" (p. 511).

In terms of temporal stability, ipsative scores have even poorer reliability than do normative subtest scores (McDermott et al., 1992). That is, difference scores have lower

reliabilities than do the scores from which they were derived. For example, normative scores were calculated to be more stable than ipsative scores in both the short- (approximately 1 month) and long-term (approximately 3 years) (McDermott et al., 1992). Part of the WISC-R standardization sample (n = 97) was used for short-term analyses, and the long-term analyses involved 189 children receiving special education services. Short-term analyses resulted in average stability coefficients of .78 for normative scores, while average correlation coefficients for ipsative scores were .63 when subtest discrepancies were calculated from the mean of all 11 subtests, and .62 when discrepancies were computed from the verbal and performance means. Long-term analyses resulted in stability coefficients of .50, .37, and .28, respectively. Further, subtests with the most common variance suffered the largest reduction in stability upon ipsatization. When classificatory stability based on strengths and weaknesses (as defined by ipsative scores having an absolute value of at least 3 points) was investigated using the same samples, discouraging results were found across both the short-term (approximately 1 month) and long-term (approximately 3 years) test-retest intervals. That is, the probability was low that an individual would possess a given strength or weakness over time.

Similarly, the 60 participants with academic and behavior problems in Livingston et al.'s (2003) sample who were administered the WISC-R twice, an average of 3.09 years apart, received ipsative scores that had lower test-retest reliability than their normative scores. Ipsative score coefficients ranged from .29 to .58 with the majority being below .5, while subtest coefficients for normative scores were between .53 and .76. Results of this study showed that ipsative scores cannot be interpreted reliably.

Watkins and Canivez (2004) also found ipsative scores to be very unstable over time. Based on 76 WISC-III patterns found in the literature, Watkins and Canivez classified 579 students, most having an identified disability, according to the ipsative methods described by Kaufman and Lichtenberger (2000). All participants were readministered the WISC-III an average of 2.8 years later and classificatory stability was found using kappa (*k;* Cohen, 1960). Agreement mostly at chance levels was found across subtest and composite patterns. Shorter (≤ 2 years) and longer (> 2 years) test-retest intervals for the subtest patterns and intermediate IQ components did not influence results. Further, instability appeared to be specific to ipsative subtest analysis; classification decisions based on a global IQ cut-score of 70, achievement cut-scores of 85, and exceptionality category decisions remained comparatively stable over time.

Group Differences, Inverse Probabilities, and Circular Reasoning

Watkins, Glutting, and Youngstrom (in press) identified two fundamental errors that researchers who find support for the interpretation of profile analysis often commit. First, subtest profile determination often results from the analysis of statistically significant group differences. However, not only does this method fail to take the magnitude of the group difference into account, but being able to reliably differentiate at a group level is not necessarily transferable to an individual level. Diagnostic utility statistics must be calculated in order to evaluate a test's ability to differentiate between individuals (e.g., sensitivity, specificity, positive predictive power, negative predictive power, Receiver Operating Curve analysis). For example, Watkins (2000) reviewed four articles in a special issue of *School Psychology Quarterly* dedicated to profile analysis research and

concluded that "detailed analyses found all four cognitive profile reports lacking in terms of reliability, validity, or diagnostic utility" (p. 475).

The second problem discussed by Watkins et al. (in press) relates to erroneously equating inverse probabilities. It is critical to remember that the probability of having a certain diagnosis given a positive test is not the same as the probability of having a positive test given a specified diagnosis (McFall & Treat, 1999). The difficulty that can arise from incorrectly equating these probabilities is illustrated: Inaccurate LD diagnosis based on subtest profile analysis results when the probability of having a certain WISC-III subtest pattern given an LD diagnosis is equated with the probability of having an LD given a subtest configuration, even in the case when the former is much higher.

In addition to difficulties related to inverse probabilities, Glutting et al. (1998) noted that circular reasoning is also one of the main problems that plagues research on cognitive profile analysis. Circular reasoning refers to the use of subtest profiles from IQ tests when forming groups and again when determining profile characteristics of these same groups (Watkins, 2003). Watkins (2003) gives the example of using WISC-III scores in the diagnosis of LD in a sample of children and then using those same scores to determine subtest patterns found in this sample of children found to have LDs. This process increases the likelihood of finding differences between the LD group and other groups, greatly limiting generalizability of results (Glutting, McDermott, Watkins, et al., 1997).

Lack of Support for Diagnosis and Hypothesis Generation Resulting From Clinically Based WISC-III Profile Analysis

Given that clinically based profile analysis techniques are fraught with problems, it is not surprising that there is little evidence to support their use for diagnosis and hypothesis generation. For example, the instability of both subtest and ipsative scores highlights the danger of using these scores to make long term educational decisions.

Diagnosis

The first modern review to examine the utility of clinically based profile analysis techniques for diagnosis was done by Kavale and Foreness (1984). Kavale and Forness conducted a meta-analysis in order to clarify the research regarding the performance of students with LD on Wechsler series tests compared with that of other children. These authors identified 94 studies examining the performance of students with LD on the WISC, WISC-R, and WPPSI. Findings revealed that children with LD scored within the average range of intellectual ability on the FSIQ, Verbal Scale IQ, and Performance Scale IQ, although scores were slightly lower than those of children without disabilities. Further, results indicated that examination of Wechsler profile scatter, profile shape, factor scores, and scores resulting from recategorizations were not useful in differentiating students with LD from those without disabilities regardless of age and average FSIQ. Interestingly, those with LD were found to have less subtest scatter than students without disabilities, further negating the utility of scatter in the diagnosis of LD. Although there were some trends noted across subtest scores, there was much overlap with children without disabilities. Kavale and Forness concluded that "although WISC

profile and scatter analysis is not defensible for diagnosing LD, the WISC remains a valuable tool for *global* IQ assessment and should be restricted to this purpose" (p. 150).

More recent reviews have confirmed that applying clinically based subtest analysis to cognitive measures such as the WISC-III when diagnosing students lacks scientific basis (Watkins, 2003; Watkins et al., in press). Watkins et al. (in press) found that student diagnosis did not correspond to subtest scatter or to subtest patterns found in the literature. A review by Watkins (2003) reached similar conclusions regarding the inutility of making diagnostic decisions based on clinical profile analysis. Further, subtest analysis research and application was associated with a host of difficulties including poor subtest score reliability, use of ipsative measures, and reliance on group mean differences and inverse probabilities (Watkins, 2003). Watkins et al. (in press) concluded that "unmistakably, abundant scientific evidence and expert consensus recommend against the use of subtest profiles for the diagnosis of childhood learning and behavior disorders."

Thus, the literature has reached consensus regarding the lack of diagnostic significance of clinically based profile analysis and even proponents of clinical profile analysis recognize these limitations. For example, Kaufman and Lichtenberger (2000) stated that "the use of profiles with the WISC-III (and its precursors, the WISC and WISC-R) to define LD has been studied by many researchers…the consensus has been that although such profiles are consistently found in LD populations, these profiles alone cannot clearly distinguish normality from abnormality and identify special populations" (p. 203).

Hypothesis Generation

There has been little effort to review the utility of clinically based profile analysis for hypothesis generation (one notable exception is the review by Watkins [2003]). Unlike diagnosis, suggested methods of hypothesis generation lack falsifiability, or the possibility of being refuted, a necessary component of empirical research (Popper, 1959). That is, while an accepted classification system (e.g., IDEA-97) can be used to test whether specified methods of profile analysis are useful for diagnosis, no such criterion exists for hypothesis generation. For example, abilities thought to underlie groups of subtests were outlined by Kaufman and Lichtenberger (2000); however, they noted that "the lists of abilities and influences on the subtests are not exhaustive and may be added to by examiners" (p. 186). Thus, should results of profile analysis come to conclusions which are inconsistent with a child's performance, the examiner can simply choose an alternate explanation.

It is thus difficult to study the utility of clinically based profile analysis in generating hypotheses; however, the assumptions underlying hypothesis generation can be examined. One assumption is that identification of cognitive strengths and weaknesses are indicative of specific aptitudes, which can be translated into appropriate treatments. For example, a case report described by Kaufman and Lichtenberger (2000) recommended that one student be taught using meaningful physical manipulatives due to the 23 point discrepancy between her PO and VC score. The student's stronger non-verbal score was used as indication that she would benefit from instruction with manipulatives.

Basing intervention on a student's specific aptitudes, known as aptitude by treatment interaction (ATI), has not gained support in the literature. Although logical, Gresham and Witt (1997) noted that several reviews of the ATI literature have revealed that positive significant interactions could not be consistently shown. For example, Reschly (1997) found that aptitudes identified in ATI research could be grouped into three main types: modality (i.e., auditory, visual, and kinesthetic), cognitive (e.g., sequential and simultaneous), and neuropsychological (e.g., right and left hemisphere strengths and weaknesses). However, he concluded that interventions generated by following ATI logic do not result in higher levels of achievement. In fact, significant interactions in a direction opposite to theory have been found (e.g., Good, Vollmer, Creek, Katz, & Chowdhri, 1993). Gresham and Witt concluded that "based on the disappointing results of ATI studies using modality matching, cognitive style/processing, and neuropsychological assessment, there is little, if any, empirical support for prescribing different treatments based on the assessment of different aptitudes" (p. 253).

Given that the underlying assumption of hypothesis generation based on clinical profile analysis techniques have been invalidated (i.e., ATI research has not been supported by the literature), it is not surprising that research shows a lack of support for this procedure. For example, if hypothesis generation was a useful practice then uncovering students' specific aptitudes should correspond to their academic achievement. However, there are many studies that have found that clinically based profile analysis scores from Wechsler and other intelligence tests (e.g., ipsative subtest scores, normative subtest scores) were not predictive of achievement as measured by standardized achievement tests (e.g., Glutting et al., 1998; Kline, Snyder, Guilmette, & Castellanos,

1993; McDermott et al., 1992; Wechsler, 1991). In fact, cognitive profile information beyond aggregate normative elevation was not found to be useful in the prediction of academic achievement (e.g., Watkins & Glutting, 2000). Thus, it is global intellectual functioning scores that remain useful in making predictions about variables such as school achievement and occupational success (Sattler, 2001). A review by Watkins (2003) came to similar conclusions.

Nonlinear Multivariate Profile Analysis: An Empirical Approach

Advantages of Nonlinear Multivariate Profile Analysis Techniques Over Clinical Methods of Profile Analysis

It is clear that clinically based profile analysis techniques do not contribute appreciably to diagnostic decision making or valid hypothesis generation. However, an empirical approach to profile analysis may engender support for these purposes. Nonlinear multivariate profile analysis, an empirical method of profile analysis, has certain advantages over many clinically based approaches.

First, profiles, including those from the WISC-III, involve multiple scores that, together, yield information along three dimensions. Profile analysis methods should take this complexity into account. Nonlinear multivariate profile techniques take both level and shape of the profile into consideration simultaneously (Glutting, McDermott, Watkins, et al., 1997; Hair, Anderson, Tatham, & Black, 1998), unlike many clinical methods of profile analysis, which consider only a single profile dimension. That is, both linear (i.e., level) and nonlinear (i.e., shape) characteristics of the profile are considered when nonlinear multivariate techniques are employed (Glutting et al., 1998).

Consideration of both linear and nonlinear profile components is due to the fact that nonlinear multivariate profile analysis allows for the simultaneous examination of multiple subtest scores (Glutting, McDermott, & Konold, 1997). By its nature, profile analysis should involve simultaneous multiple dependent comparisons (Glutting, McDermott, Watkins, et al., 1997). Many clinical profile analysis techniques are univariate and univariate techniques do not allow for more than one comparison at a time. In contrast, nonlinear multivariate techniques can be used to conduct a number of comparisons simultaneously (Livingston et al., 2003). Further, Livingston et al. explained how the reliability of the set of profile scores taken together may present a more accurate picture than consideration of the reliability of scores in isolation, whether normative or ipsative. That is, taken in combination, subtest scores may be more reliable than individual element scores. This is important given that one of the disadvantages of clinically based profile analysis based on interpretation of single subtests scores is the low reliability of these scores.

Second, reliance on statistically significant differences between subtest scores and the mean or between scores in order to guide interpretation of cognitive results is problematic due to the fact that significant differences are common in the population and are not necessarily indicative of a problem (Konold et al., 1999). When using nonlinear multivariate profile analysis, this is no longer a concern. This is because profiles are considered for further interpretation if they are unusual compared to nonlinear multivariate base rates, or normative taxonomies of profiles. That is, unusual profiles thought to be of clinical interest can be defined as those that show a difference that is

significant when compared to all typical profiles (Glutting, McDermott, & Konold, 1997).

A third way in which nonlinear multivariate profile techniques may be beneficial relates to the fact that nonlinear multivariate subtest analysis does not rely on ipsative scores, as do many clinical profile analysis techniques. As discussed, there are many difficulties associated with using ipsative scores. For example, one limitation of ipsative scores is failure to take the correlations between subtest scores into account. On the other hand, nonlinear multivariate profile analysis accounts for the intercorrelations among subtests (Glutting, McDermott, & Konold, 1997).

Given the advantages of nonlinear multivariate profile analysis over clinically based techniques, this empirical method of profile analysis warrants further study. Specifically, perhaps empirical consideration of elevation, scatter, and shape together will result in a method of profile analysis that yields scores that can be reliably and validly interpreted with respect to diagnosis and hypothesis generation.

Cluster Analysis

Nonlinear multivariate profile analysis involves empirically grouping people into profile types based on score configurations that are commonly found in the population. One way of doing this is via a group of techniques known as cluster analysis. Cluster analysis refers to a set of multivariate techniques whose purpose is to group together items in a data set that are maximally alike on a pre-specified set of variables, while at the same time ensuring the greatest amount of difference between clusters (Hair et al., 1998). In this way the inherent structure underlying the data can be defined. That is, typically cluster analysis is used in the formation of taxonomies, or empirical classification

systems, such as the creation of a classification system of common cognitive profiles for the standardization sample of an intelligence test.

Throughout the process of cluster analysis many crucial decisions must be made by the researcher that can significantly affect the outcome. Further, there is no unambiguous set of rules to guide the researcher (Hair et al.) and cluster analysis will generate clusters whether or not an underlying cluster structure exists in the data (Speece, 1994-95). As such, researchers must ask questions regarding theory formulation, internal validity, and external validity (Speece).

It is important for the researcher to consider two assumptions underlying cluster analysis (Hair et al., 1998). First, cluster analysis depends on the researcher to obtain a sample that is representative of the population of interest, such as the standardization sample of certain intelligence tests, as no statistical procedure exists to generalize results of a cluster analysis from a sample to a population. Second, the researcher must realize that multicollinearity among variables can result in a disproportionate influence of those variables in the formation of clusters.

After identifying several difficulties with existing options in cluster analysis, McDermott (1998) developed a three-stage clustering method. Identified problems with contemporary applications of cluster analysis included a lack of accessibility to certain statistics and techniques relevant to cluster analysis as well as appropriate cluster reassignment not being included in hierarchical clustering algorithms. Multistage Euclidean grouping (MEG) takes these problems into account and incorporates certain best practice techniques in cluster analysis. Following score transformation into standard scores, two or more blocks of data are generated either randomly or according to a given

variable, such as age. Blocks must contain at least 100 cases, and preferably between 150 and 300 (McDermott).

The first stage of MEG involves applying an agglomerative hierarchical clustering algorithm to each data block (McDermott, 1998). An algorithm refers to a sequence of rules to be followed in order to convert a set of similarity measures between entities into a group of relatively homogeneous clusters (Borgen & Barnett, 1987). Clustering algorithms can be broadly categorized as hierarchical and nonhierarchical (Hair et al., 1998). Hierarchical methods are stepwise constructions of clusters that are either agglomerative or divisive. Agglomerative techniques are much more common than those that are divisive (Borgen & Barnett, 1987). Agglomerative procedures begin with each entity being one cluster (Hair et al.). In each step, the two most like clusters are combined, as defined by the algorithm, and the total number of clusters is reduced by one. Eventually all observations are part of the same cluster.

One example of an agglomerative hierarchical algorithm is Ward's (1963) method, a popular algorithm, especially in the behavioral sciences (Borgen & Barnett, 1987). Considered one of the best agglomerative methods, Ward's technique combines the two clusters that result in the smallest increase in within-cluster variance at each stage in the cluster analysis (Borgen & Barnett). In their review of cluster analysis, Aldenderfer and Blashfield (1984) summarized that Ward's method may be considered the technique of choice when full coverage is required and when clusters overlap, although research on the latter has yielded mixed findings. Ward's method has been incorporated into MEG applications (e.g., Konold et al., 1999).

In the second stage of MEG, higher order clustering is performed based on a similarity matrix generated from all the clusters derived during the first clustering stage (McDermott, 1998). Information about profile elevation, shape, and scatter is preserved in this matrix. Statistics summarizing characteristics of the resulting clusters and cluster solutions are provided following both the first and second stages of the analysis (McDermott). Results of various stopping rules are made available. Stopping rules are used to determine the final number of clusters in a solution when a hierarchical clustering technique is employed (Hair et al., 1998). Although more research in the area of stopping rules is needed (Borgen & Barnett, 1987; Milligan & Hirtle, 2003), MEG reports the results of stopping rules that have some empirical support, such as Mojena's (1977) popular stopping rule (e.g., Milligan & Cooper, 1985).

Using clusters produced in the second stage as starting points, the third stage involves *k*-means iterative partitioning (McDermott, 1998). Nonhierarchical clustering procedures, or *k*-means clustering, require prior knowledge of the number of clusters and the initial means of these clusters (Hair et al., 1998). All observations within a given distance of a starting point are then combined into one cluster. Observations may then be reassigned if they are more similar to another cluster. Nonhierarchical clustering is also known as iterative partitioning.

There are advantages and disadvantages to both hierarchical and nonhierarchical clustering procedures (Hair et al., 1998). The difficulty with hierarchical procedures is the fact that once assigned to a cluster, observations cannot be reassigned. Especially problematic may be the presence of outliers. Reassignment of observations is possible with nonhierarchical methods. However, this procedure requires the prespecification of

the number of clusters that will be formed as well as a method of identifying cluster starting points. The results of nonhierarchical cluster analysis are directly related to the selected starting points, and so the absence of a basis upon which to select these starting points renders the procedure of little use. On the other hand, when nonrandom starting points can be specified with confidence, the results are more robust against the presence of outliers, the distance measure employed, and the inclusion of extraneous variables.

Given the comparative advantages and disadvantages of each procedure, it can be advantageous to employ a combination of both methods by using the hierarchical procedure to identify the number of clusters as well as cluster starting points that will be applied to nonhierarchical clustering. This combination was incorporated into MEG (McDermott, 1998). That is, the mean profiles of second stage clusters were specified as staring points for third stage, iterative clustering.

Replication of an obtained cluster solution is one method of gathering evidence of validity for the proposed structure (Aldenderfer & Blashfield, 1984; Lorr, 1983; Milligan & Hirtle, 2003). That is, replication of a solution across samples from the same population supports generalizability of results (Aldenderfer & Blashfield). MEG includes built-in replications (McDermott, 1998). Specifically, MEG reports for each final cluster the percentage of first stage solutions in which it can be found. In this way, replicability of the cluster solution across the blocks of data is determined. Blocks comprising different age groups, for example, can ultimately be used to determine generalization of the final cluster solution across age groups. For example, when discussing their application of cluster analysis to the WPPSI standardization sample, Glutting and McDermott (1990a) noted that "groups of similar profiles…should be reasonably

replicable rather than chance mergers of different profiles" (p. 487) and "the typology was deemed to have at least some applicability if it independently replicated across each of the WPPSI's six age levels" (p. 487).

Taxonomies of Profiles from Commonly Used Cognitive Measures

Classifying participants empirically, using techniques such as cluster analysis, based on a given set of elements, such as the scores on a cognitive measure's subtests, leads to the creation of a taxonomy of profiles (Hair et al., 1998). These common profiles can be termed *core* profiles. Using MEG procedures, or a modification of these procedures, core profiles have been identified for the standardization samples of a number of cognitive tests, including the WISC-R (McDermott, Glutting, Jones, Watkins, & Kush, 1989), WPPSI (Glutting & McDermott, 1990a), McCarthy Scales of Children's Abilities (Glutting & McDermott, 1990b), WAIS-R (McDermott, Glutting, Jones, & Noonan, 1989), Kaufman Assessment Battery for Children (K-ABC; Kaufman & Kaufman, 1983a, 1983b) by Glutting, McGrath, Kamphaus, and McDermott (1992), Differential Ability Scales (DAS; Elliott, 1990) by Holland and McDermott (1996), and WISC-III (Donders, 1996; Glutting, McDermott, & Konold, 1997; Konold et al., 1999). Full coverage was required in every case in order for derived taxonomies to be representative of the population. In each instance, taxonomies contained relatively few core profiles, ranging across studies from 5 to 9. The major distinction among most of these profiles was in terms of elevation; however, many core profiles were not flat and were, instead, characterized by differences between composite scores. No profile was defined by scatter.

Given this study's focus, results of the WISC-III core subtest profile taxonomy based on 10 subtests, and those of the WISC-III core factor profile taxonomy will be described further.

WISC-III subtest profile taxonomy based on 10 mandatory subtests. A taxonomy based on the 10 mandatory WISC-III subtests was developed by Konold et al. (1999). A taxonomy based on only these 10 WISC-III subtests is useful given that students were infrequently administered the supplementary Digit Span and Symbol Search subtests and also, given that these two subtests do not contribute to global IQ scores (Konold et al.). The 2,200 students in the WISC-III standardization sample were the participants.

Cluster analysis was conducted using MEG procedures (Konold et al., 1999). As with other taxonomies developed for tests of cognitive functioning, the resulting 8 core profiles in the WISC-III taxonomy based on 10 subtests were mainly distinguished by FSIQ level, but were also marked by VIQ/PIQ discrepancies. An abnormal discrepancy between VIQ and PIQ denoted a profile in which more than 3% of examinees (determined via tests of statistical significance) displayed a difference found only in 3% of the general population. Konold et al. also noted that Arithmetic and Coding subtests tended to covary. Table 2 displays characteristics of the 8 core profiles along with each one's mean FSIQ, and Figure 1 illustrates core profile level and shape.

Table 2

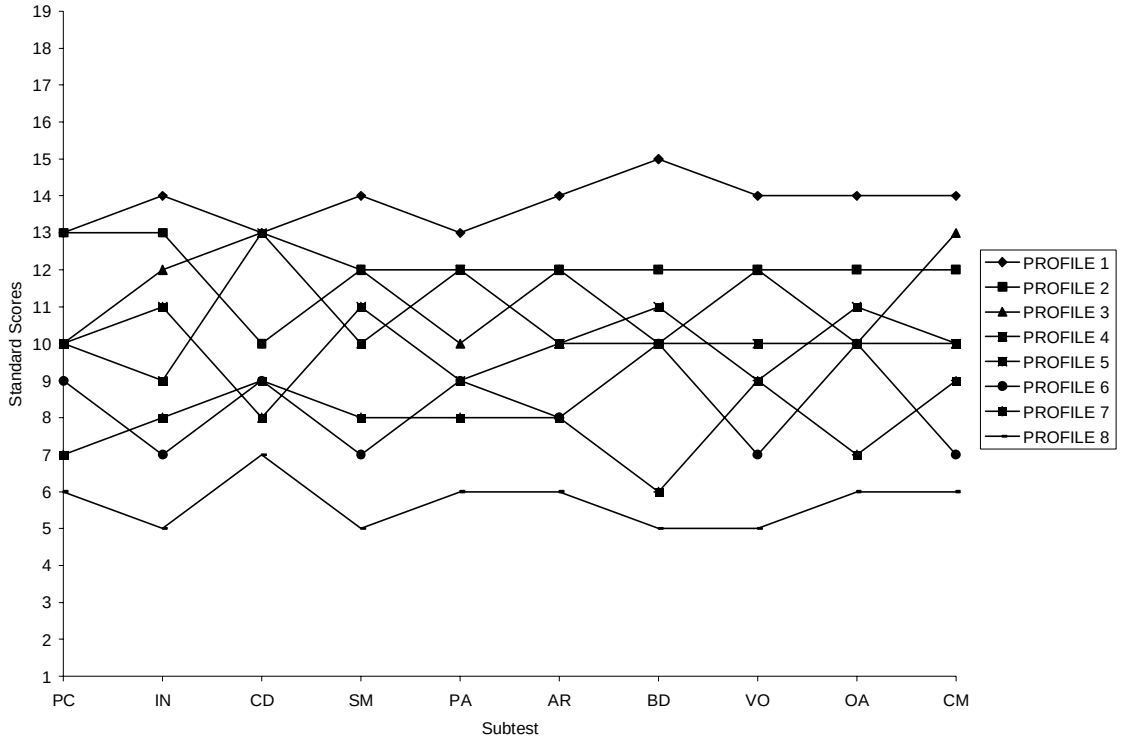
Description and Mean FSIQ of Core Profiles in the Taxonomy Developed Based on 10 WISC-III Subtest Scores

	Description	Mean FSIQ
Profile 1	High ability	126.2
Profile 2	Above average ability	113.9
Profile 3	Above average ability; VIQ > PIQ	108.5
Profile 4	Average ability; PIQ > VIQ	102.6
Profile 5	Average ability; VIQ > PIQ	99.1
Profile 6	Below average ability; PIQ > VIQ	89.3
Profile 7	Below average ability	87.6
Profile 8	Low ability	73.1

Note. Konold et al., 1999.

Figure 1

Core Profile Level and Shape for the WISC-III Taxonomy Based on 10 WISC-III Subtest Scores (Konold et al., 1999)



PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension.

WISC-III factor score taxonomy. Given the relatively low reliabilities of some WISC-III subtest scores, Donders (1996) developed a taxonomy based on the four WISC-III factor scores. Using the 2,200 children from the WISC-III standardization sample, a two-stage cluster analysis was performed. The first stage consisted of agglomerative hierarchical cluster analysis. The second stage of analysis consisted of a *k*-means clustering technique with starting points based on first stage results. The final solution

had 5 clusters. Of these, 3 clusters were predominantly distinguishable by level, while the other 2 revealed distinctions mainly based on shape, especially variation in the PS score. The two profiles defined by shape displayed differences between PS and the other three factor scores ranging from 13 to 17 points for one core profile, and from 9 to 12 for the other. Table 3 displays characteristics of the 5 core profiles along with each one's mean FSIQ, and Figure 2 illustrates core profile level and shape.

Table 3

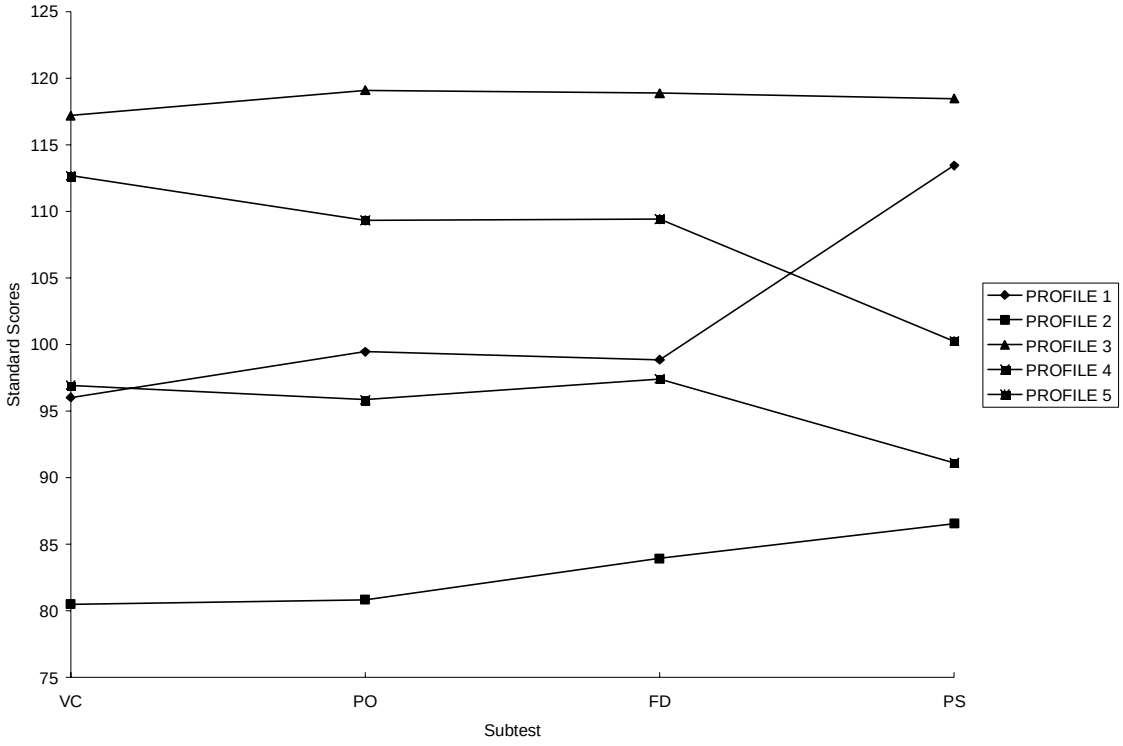
Description and Mean FSIQ of Core Profiles in the Taxonomy Developed Based on 4 WISC-III Factor Scores

	Description	Mean FSIQ
Profile 1	Average ability; relative strength on PS	98.9
Profile 2	Below average ability	78.3
Profile 3	Above average ability	121.4
Profile 4	Average ability	94.5
Profile 5	Above average ability; relative weakness on PS	111.0

Note. Donders, 1996.

Figure 2

Core Profile Level and Shape for the WISC-III Taxonomy Based on Four WISC-III Factor Scores (Donders, 1996)



VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractibility; PS = Processing Speed.

The WISC-III taxonomy derived for mandatory subtest scores (Konold et al., 1999) and the one generated for factor scores (Donders, 1996) exhibited a small number of profiles that were defined primarily by level, with some profiles defined by shape. These attributes conformed to a more general trend. That is, core profiles from many tests of intelligence revealed these characteristics across experimenters and methods. As such, results are promising in terms of internal structure, but must also be examined for

stability and external validity. The stability of profile membership should be investigated as evidence of reliability is a prerequisite for the valid interpretation and use of test scores (AERA, APA, & NCME, 1999). That is, if profile membership is not stable over time, then relying on nonlinear multivariate profile analysis to make decisions for students is unscientific and not defensible.

Temporal Stability of Multivariate Profiles

Some initial evidence exists to suggest that membership to empirically derived cognitive subtest profiles possesses some level of stability in the short- term (i.e., several weeks). Using a subsample of the WISC-R standardization sample (n = 303) that was administered 11 WISC-R subtests (Mazes was excluded) on two separate occasions with a retest interval generally ranging from 3 to 5 weeks, McDermott, Glutting, Jones, Watkins, and Kush (1989) examined WISC-R profile stability while controlling for practice effects measured over the short retest interval.

The taxonomy derived using scores from the WISC-R standardization sample (McDermott, Glutting, Jones, Watkins, & Kush, 1989) was used to determine students' initial profile memberships: These students had been part of the sample used to generate the taxonomy. Retest membership was established by first calculating Tatsuoka and Lohnes (1988, pp. 377-378) modifications of Cattell's (1949) coefficient of profile similarity (r_p), allowing for comparison of each profile to a set of core profiles. Second, an iterative cluster analytic technique was used to reassign profiles to more appropriate clusters. Both general and partial kappa coefficients (k_m; Fleiss, 1971) were calculated. While general k_m coefficients indicate overall classification agreement beyond chance,

partial k_m coefficients are computed to signify agreement beyond chance within each category (i.e., core profile type).

Considerable agreement was found between profile membership at both testing times (McDermott, Glutting, Jones, Watkins, & Kush, 1989). That is, overall agreement between test and retest classifications was determined to be 64.7%, and 57.5% beyond chance levels (i.e., general k_m = .575). Further, most partial k_m coefficients were found to be statistically significant ($p < .01$). Comparable results were found for overall short-term stability of profile-type membership for the WPPSI (general k_m = .216; Glutting & McDermott, 1990a), McCarthy Scales (general k_m = .728; Glutting & McDermott, 1990b), K-ABC (general k_m = .497; Glutting et al., 1992), and DAS (general k_m = .541; Holland & McDermott, 1996). In addition, partial k_m coefficients were again found to be statistically significant for McCarthy Scales' core profiles ($p < .0001$; Glutting & McDermott, 1990b) as well as for K-ABC core profiles and a group of unusual K-ABC profiles ($p < .001$; Glutting et al., 1992).

In contrast to findings for short-term profile stability, one study found that empirically derived cognitive subtest profiles may not be sufficiently stable for educational decision-making in the long-term (Livingston et al., 2003). Sixty students referred due to academic and behavior difficulties participated in this study. Participants were administered the WISC-R with an average retest interval of 3.09 years. At the first testing, the children had a mean age of 10.4 years and an average grade level of 4.1. Over 70% of the sample was male. Computation of a set of profile similarity measures, based on the average of the calculated similarity between each examinee's profile at the first and second testing occasion, revealed that IQ and index profiles were more stable than

subtest profiles. For example, the average r_p was found to be .62, .70, and .43 for IQ, index, and subtest score profiles, respectively. This pattern was also found when other measures of similarity were employed. Average subtest profile stability coefficients were determined by Livingston et al. to "indicate an unsatisfactory level of reliability" (p. 504).

However, the long-term stability of cognitive profiles must be further examined as Livingston et al. (2003) did not evaluate profile stability by comparison to a core taxonomy. Instead, each student's profile at Time 1 was compared to his or her profile at Time 2. However, it is possible that this group of referred children had unusual subtest profiles compared to core profile types and that their profiles were stable in the sense that they remained unusual over time. Further, the change in profiles found across time may not have been large enough for core profile type reassignment. For example, in developing core subtest taxonomies for the WISC-R and WISC-III, McDermott, Glutting, Jones, Watkins, and Kush (1989), Glutting, McDermott, and Konold (1997), and Konold et al. (1999) all set the a priori mean r_p between clusters to be < .40. Additionally, results showed that the range of average r_p scores across these studies was .20 to .33. That is, profiles in Livingston et al.'s study were more similar to one another at different points in time (average r_p = .43) than were core profile clusters based on the standardization sample cognitive scores. Another difficulty with Livingston et al.'s study was the small sample size. With only 60 participants it is quite possible that not all taxonomy categories were adequately represented, possibly resulting in misleading findings. Also, participants were students who were referred due to academic and behavior difficulties, making it even more probable that some taxonomy categories were not satisfactorily represented.

Purpose of Present Study

It is crucial that sound educational decisions, including those involving diagnosis, instructional methods, and other matters, be made on behalf of students. Data used to make decisions should possess temporal stability given that reliability is a prerequisite for validity (AERA, APA, & NCME, 1999) and given that educational choices based on these data may remain in effect for a duration spanning several years (IDEA-97; Kaufman, 1994; Kaufman & Lichtenberger, 2000). Profile analysis using the Wechsler series tests have become a popular method on which to base, in part, these educational decisions (Alfonso et al., 2000; Pfeiffer et al., 2000).

Global cognitive scores on Wechsler series tests, such as the WISC-III, have been shown to be stable over a time period of several years (e.g., Canivez & Watkins, 1998) and have been supported in their predictive ability (Sattler, 2001). On the other hand, clinically based methods of profile analysis, such as interpretation of shape based on guidelines with no empirical basis, are replete with difficulties (e.g., McDermott et al., 1992). It follows that the results of these analyses have not been shown to be helpful when making diagnostic or intervention decisions for students (e.g., Kavale & Forness, 1984; Kline et al., 1993; McDermott et al., 1992; Watkins & Glutting, 2000; Watkins & Worrell, 2000).

Nonlinear multivariate profile analysis, involving empirical determination of profile membership via its comparison to an empirically derived profile taxonomy (Glutting, McDermott, & Konold, 1997), is advantageous in several ways. For example, nonlinear multivariate profile analysis allows for the consideration of multiple subtest scores simultaneously (Glutting, McDermott, & Konold). Further, nonlinear multivariate profile

techniques take both level and shape of the profile into account at the same time (Glutting, McDermott, Watkins, et al., 1997; Hair et al., 1998). Perhaps empirical consideration of elevation, scatter, and shape together will result in a method of profile analysis that generates results that are useful in diagnosis and educational decision making.

Taxonomies of core profiles have already been developed for many intelligence tests including the WISC-III (Donders, 1996; Glutting, McDermott, & Konold, 1997; Konold et al., 1999). However, profiles must be stable for membership to be valid (AERA, APA, & NCME, 1999). Thus, it is important to determine whether students' profile-type membership remains stable in the long term (i.e., several years). That is, if a decision based on a profile is not stable over time, then the use of nonlinear multivariate profile analysis to make lasting educational decisions for students may lead to choices that are at best, ineffective, and at worst, harmful.

Given that there is virtually no research examining long-term empirical profile stability, the present study intends to explore the long-term (i.e., 3 year) stability of WISC-III nonlinear multivariate subtest and factor profile cluster membership. That is, the research question is: Is WISC-III cluster membership based on nonlinear multivariate subtest and factor profile analysis stable over a 3 year period?

Method

Participants

Participants in the present study consisted of two subsets of the sample studied by Canivez and Watkins (1998). The first subset of children had data available for all 10 WISC-III mandatory subtests (Sample 1). The other subset had information available for the factor scores (Sample 2).

Stability of membership to core subtest profile types was examined using scores of children from Sample 1. Students in Sample 1 had data available for all 10 WISC-III mandatory subtests (i.e., all subtests except Digit Span, Symbol Search, and Mazes). This criterion was chosen for two reasons. First, at the time when WISC-III administration was popular, students were infrequently administered the supplementary Digit Span, Symbol Search, and Mazes subtests (Konold et al., 1999). As such, results of this study are more generalizable because participants' WISC-III administration reflects what was most widely implemented. Second, a much larger sample size was possible when results of a 10-subtest administration rather than a 12-subtest administration were desired. Specifically, a sample size of 585 was attained, instead of a sample size of 177. A larger sample size is beneficial as obtained results are more likely to be generalizable (Gall, Borg, & Gall, 1996).

WISC-III test retest data for Sample 1 was reported by 107 school psychologists in 33 different states. On average, 5.47 cases were reported per psychologist, with a range from 1 to 24 and a standard deviation of 3.84. Table 4 displays the demographic characteristics of this sample.

Table 4

Gender, Race/Ethnicity, Disability, and Grade Level of Participants with Data Available for all 10 WISC-III Mandatory Subtests (Sample 1)

	n	*%*
Gender		
Boys	394	67.35
Girls	191	32.65
Race/Ethnicity		
White	447	76.41
Black	86	14.70
Hispanic	33	5.64
Native American	4	.68
Asian/Pacific	1	.17
Other	4	.68
Missing	10	1.71
Disability[a]		
Not disabled	18	3.08
Learning disability	368	62.91
Mental retardation	57	9.74
Emotional disability	42	7.18
Speech and language disability	16	2.74
Other disabilities	38	6.50

(table continues)

Table 4 (continued)

	n	%
Unspecified	46	7.86
Grade[b]		
Kindergarten	21	3.59
1	109	18.63
2	138	23.59
3	94	16.07
4	76	12.99
5	71	12.14
6	36	6.15
7	26	4.44
8	8	1.37
9	2	.34
Missing	4	.68

[a]Diagnoses made during first testing in accordance with state and federal guidelines. [b]Grades at time of first testing.

Sample 1 participants' average age was 9.16 years at Time 1 (range = 6.00 to 14.60; *SD* = 2.02) and 11.98 years at the Time 2 (range = 7.50 to 16.90; *SD* = 2.07). The mean amount of time between Time 1 and Time 2 was 2.82 years (*SD* = .54) and the range was .50 to 6.00 years. The test-retest interval was less than 1 year for only 1.20% of the sample.

In order to determine how well Sample 1 represented the population of children with disabilities, participants were compared to children aged 6 to 21 who received special

education services under IDEA-97 during the 2000-2001 school year (USDOE, 2001). Generally speaking, the two groups of students had similar characteristics. With the exception of Hispanic students being underrepresented in Sample 1, those receiving special education from the 50 states, the District of Columbia, and Puerto Rico were similar to Sample 1 in terms of race/ethnicity: 62.46% of the population of children with disabilities were White, 19.87% were Black, 14.49% were Hispanic, 1.32% were American Indian/Alaskan, and 1.86% were Asian/Pacific Islanders (USDOE). The same trend was seen when the 10 members of Sample 1 who were missing race/ethnicity information were disregarded.

The composition of Sample 1 was reasonably consistent with that of children receiving special education services from the 50 states and the District of Columbia, in terms of disability type: 49.94% of those receiving special education had LDs, 10.51% had mental retardation, 8.23% had an emotional disturbance, 18.97% had speech or language impairments, and 12.23% had other disabilities (USDOE, 2001). Although those with speech and language disabilities as well as those with *other* disabilities were underrepresented in the current sample, percentages of those with the other three disability types were fairly similar. A similar trend was seen when the 46 members of Sample 1 who were missing disability information were not included, although students with LDs now comprised 68.27% of Sample 1, overrepresenting this disability type.

In order to examine how representative Sample 1 was in terms of geographic location, the country was divided into the four regions outlined in the WISC-III manual: West, South, North Central, and Northeast. The population of students receiving special education from the 50 states and the District of Columbia were distributed across the

geographic regions as follows: 20.04% in the West, 36.51% in the South, 23.70% in the North Central region of the country, and 19.75% in the Northeast (USDOE, 2001). This was not unlike Sample 1 where 21.54% of participants were from the West, 35.73% were from the South, 31.45% were living in the North Central region, and 11.28% were in the Northeast. Those in the North Central region were slightly overrepresented in Sample 1, while participants from the Northeast were slightly underrepresented.

The second sample of students in the current study had information available for the four WISC-III factor scores (i.e., information was available for all 12 WISC-III subtests excluding Mazes). Scores from these students were employed in order to determine stability of membership to core factor profile types. Scores for all four factors were available for 177 students and were reported by 55 school psychologists in 26 different states. On average, 3.22 cases were reported by participants, with a range from 1 to 16 and a standard deviation of 2.68. Table 5 displays the demographic characteristics of this sample.

Table 5

Gender, Race/Ethnicity, Disability, and Grade Level of Participants with Data Available for all Four WISC-III Factor Scores (Sample 2)

	n	*%*
Gender		
Boys	121	68.36
Girls	56	31.64
Race/Ethnicity		
White	146	82.49
Black	16	9.04
Hispanic	12	6.78
Native American	2	1.13
Asian/Pacific	0	.00
Other	1	.56
Missing	0	.00
Disability[a]		
Not disabled	4	2.26
Learning disability	113	63.84
Mental retardation	16	9.04
Emotional disability	8	4.52
Speech and language disability	7	3.95
Other disabilities	14	7.91

(table continues)

Table 5 (continued)

	n	%
Unspecified	15	8.47
Grade[b]		
Kindergarten	5	2.82
1	38	21.47
2	50	28.25
3	30	16.95
4	21	11.86
5	24	13.56
6	5	2.82
7	3	1.69
8	0	.00
9	0	.00
Missing	1	.56

[a]Diagnoses made during first testing in accordance with state and federal guidelines. [b]Grades at time of first testing.

Sample 2 participants' average age was 8.88 years at Time 1 (range = 6.00 to 13.10; *SD* = 1.74). At Time 2, the average age was 11.72 years (range = 7.50 to 16.00; *SD* = 1.80). The mean amount of time between Time 1 and Time 2 was 2.84 years (*SD* = .48) and the range was .70 to 4.00 years, with only one participant having a retest interval under 1 year.

Sample 2 was somewhat less representative of the population of students with disabilities (USDOE, 2001) compared to Sample 1, which is not unexpected given its

much smaller sample size. However, overall, Sample 2 can be considered similar to the population of children receiving special education services in terms of race/ethnicity, disability type, and geographic location. Comparable trends to those noted for Sample 1 were found for race/ethnicity and disability type, with and without including the 15 students missing disability data (e.g., students with speech and language disabilities were underrepresented in Sample 2). In terms of geographic trends, while Sample 2 had similar proportions of students living in the western (23.16%) and northeastern (19.77%) parts of the country compared to the population of students receiving special education, southerners (25.99%) were slightly underrepresented and those from north central regions (31.07%) were slightly overrepresented.

Instrument

General Description of the WISC-III

In order to study long-term stability of empirical cluster membership, participants were administered the WISC-III at both Time 1 and Time 2, an average of 2.82 years later for Sample 1 and 2.84 years later for Sample 2. The WISC-III is an individually administered test of intelligence that is useful in assessment, diagnosis, and research (Wechsler, 1991). The WISC-III can be administered to children between the ages of 6 years, 0 months and 16 years, 11 months. All scores provided by the WISC-III are normative; that is, a child's scores indicate their performance relative to other children of the same age. Altogether, the WISC-III is comprised of 13 subtests, which can be organized into Verbal and Performance subtests.

There are six Verbal subtests: Information, Similarities, Arithmetic, Vocabulary, Comprehension, and Digit Span (Wechsler, 1991). The Information subtest consists of a

set of factual questions that are presented orally. In the Similarities subtest, the child is asked to identify the common element between word pairs that are presented orally. The Arithmetic subtest involves asking the child to mentally compute a series of math problems within a time limit. The examinee is asked to define a series of words presented orally in the Vocabulary subtest. The Comprehension subtest involves asking the child to answer a set of questions that are orally presented and that tap his or her understanding of common dilemmas or social matters. In Digit Span, children are asked to recall sets of increasingly long series of digits that are orally presented. They are then asked to repeat this activity, naming the digits in reverse order.

There are seven Performance subtests: Picture Completion, Coding, Picture Arrangement, Block Design, Object Assembly, Symbol Search, and Mazes (Wechsler, 1991). In Picture Completion the child is asked to identify a key part that is missing from each of a series of pictures representing everyday objects and sights. A time limit is imposed. The Coding subtest requires the examinee to fill in symbols that have been matched with a set of shapes or numbers, depending on the child's age. The child follows a key that shows which symbols correspond to which shapes or numbers and fill in the symbols either underneath the numbers or in the shapes, within a time limit. Picture Arrangement involves asking that the examinee assemble sets of cards with pictures on them that, when in the correct order, tell a story. Again, a time limit is imposed. Within a time limit, the child is required to arrange red-and-white blocks according to models displaying two-dimensional designs, in Block Design. In the Object Assembly subtest, the examinee arranges a set of puzzles within a time limit. The Symbol Search subtest entails the child searching for a specified target object or objects, depending on age,

within a search group. A series of these problems are presented to the child and a time limit is imposed. Finally, the Mazes subtest asks the child to solve a series of mazes of increasing difficulty within a time limit. All Verbal and Performance subtest scores have a mean of 10 and a standard deviation of 3.

A child's performance across all the subtests yields an overall, or Full Scale IQ (Wechsler, 1991). This score is computed based on a child's scores on the 10 mandatory subtests (i.e., all subtests except Symbol Search, Digit Span, and Mazes). In addition, both a Verbal (VIQ) and Performance (PIQ) composite score can be calculated based on scores from the 5 mandatory subtests found under each scale, respectively. These three composite scores can be considered estimates of the child's cognitive functioning. Scores have a mean of 100 and a standard deviation of 15.

Four factor scores can also be computed (Wechsler, 1991). The Verbal Comprehension index (VC) is composed of the Information, Similarities, Vocabulary, and Comprehension subtests. Picture Completion, Picture Arrangement, Block Design, and Object Assembly comprise the Perceptual Organization index (PO). Arithmetic and Digit Span make up the Freedom from Distractability index (FD) and, finally, the Processing Speed index (PS) score is based on a child's performance on the Coding and Symbol Search subtests. Like the FSIQ, VIQ, and PIQ, the factor scores have a mean of 100 and a standard deviation of 15.

WISC-III Standardization Sample

WISC-III scores are normative and are derived through comparison of an examinee's performance to the performance of a sample of children, known as a standardization sample (Wechsler, 1991). Stratified random sampling was used to

identify the WISC-III standardization sample and was employed in an effort to have a standardization sample that was representative of the population of the United States in terms of age, gender, race/ethnicity, geographic region, and parent education. The resulting standardization sample was similar to U.S. 1988 Census data for the chosen variables. A total of 2,200 children were included in the standardization sample, 200 children from each age group (100 male and 100 female) between the ages of 6 and 16. In addition, 7% of the standardization sample had disabilities or were receiving special services, and 5% were receiving gifted services. All children had an understanding of the English language and were able to speak English.

Reliability of WISC-III scores

A number of reliability studies were performed on the WISC-III (Wechsler, 1991). A joint committee selected by the AERA, APA, and NCME (1999) defined reliability as "the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable, and repeatable for an individual test taker" (p. 180).

Reliability coefficients were calculated for the subtest, factor, and IQ scores on the WISC-III (Wechsler, 1991). Subtest score internal consistency reliability coefficients were calculated by using the split-half reliability technique with the Spearman-Brown correction. The split-half reliability coefficient represents the correlation coefficient between an examinee's aggregate score on half of the test items and his or her score on the other half. The Spearman-Brown correction is instituted to adjust for the fact that only half the amount of available items are considered in the calculation of the split-half reliability coefficient; the reliability of a test score is proportional to the number of items

that comprise the test. For two subtests, Coding and Symbol Search, stability coefficients were calculated instead. This is due to the fact that Coding and Symbol Search are speeded tasks. Reliability of IQ and factor scores was calculated via a method for finding the reliability of test composites (Nunnally, 1978). For subtest scores, the average reliability across age groups ranged from .69 (Object Assembly) to .87 (Vocabulary and Block Design). The mean reliability coefficients across age groups for the IQ scores were .95, .91, and, .96 for VIQ, PIQ, and FSIQ, respectively, and, for factor scores, .94 (VC), .90 (PO), .87 (FD), and .85 (PS).

Another way that reliability data can be summarized is through calculation of a test-retest or stability coefficient, the correlation coefficient between examinees' scores on a given test administered at two separate points in time (AERA, APA, & NCME, 1999). A subsample of the standardization sample (n = 353) were participants in a study assessing the stability of WISC-III scores (Wechsler, 1991). The test-retest interval ranged from 12 to 63 days (median = 23 days). Children were drawn from the 6, 7, 10, 11, 14, and 15 year-old age groups. Stability coefficients for the subsample ranged from .57 (Mazes) to .89 (Vocabulary), and across composite scores ranged from .82 (FD) to .94 (VIQ and FSIQ). Stability coefficients were corrected for the standardization sample's variability. Obtained stability coefficients were deemed adequate by Wechsler (1991). Canivez and Watkins (1998) found similar results when they investigated the long-term stability of WISC-III scores (i.e., the average test-retest interval in their study was 2.83 years). Like Wechsler (1991), Canivez and Watkins (1998) found that stability coefficients for 667 students, most of who had disabilities, were in the upper .80s and lower .90s for VIQ, PIQ, FSIQ, VC, and PO scores. Like Wechsler's (1991) findings, FD, PS, and subtest

scores had lower test-retest reliability coefficients, ranging from .55 (Symbol Search) to .78 (Block Design) (Mazes was excluded).

Internal consistency, alternate-forms, and stability coefficients are all thought of as different kinds of generalizability coefficients (AERA, APA, & NCME, 1999). Reliability of test scores can also be described by the degree to which separate scorers are consistent (AERA, APA, & NCME). Inter-rater agreement can be obtained via calculation of a correlation coefficient between examinees' test scores generated by two or more different scorers. For most WISC-III subtests, inter-rater agreement coefficients averaged within the high .90s across age groups (Wechsler, 1991). For subtests where scoring requires more judgment, coefficients across age groups were found to be within the low .90s. Specifically, coefficients were as follows: .94 for Similarities, .92 for Vocabulary, .90 for Comprehension, and .92 for Mazes.

Finally, reliability data may be described by the standard error of measurement (AERA, APA, & NCME, 1999). The standard error of measurement represents the standard deviation of the distribution of an examinee's test scores on repeated test administrations, where conditions are exactly the same for each administration. Standard error of measurement ranged from 1.08 (Vocabulary) to 1.67 (Object assembly) for subtest averages across age groups, and from 3.20 (FSIQ) to 5.83 (PS) for composite means across the different age levels (Wechsler, 1991).

Evidence of Validity of WISC-III Scores

Validity can be defined as "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests" (AERA, APA, & NCME, 1999, p. 9). Validity is thought to be a unitary construct and, as such, there are no types

of validity. However, different sources of validity evidence can be gathered. Two ways in which evidence of validity can be gathered include evidence based on relations to other variables and evidence based on internal structure, (AERA, APA, & NCME). These forms of validity evidence have been gathered to support interpretation of WISC-III scores (Wechsler, 1991).

Evidence based on relations to other variables includes convergent as well as discriminant evidence of validity (AERA, APA, & NCME, 1999). Convergent evidence is collected when it is determined that test scores are related to criterion variables thought to measure similar constructs, while discriminant evidence is gathered when a low relationship is found between measures of different constructs. Convergent and discriminant evidence of validity is available for the WISC-III (Wechsler, 1991). Matrices of intercorrelations were developed for each age group between subtest, factor, and IQ scores. Findings indicated that Verbal subtests were more highly correlated with each other (i.e., convergent evidence) than they were with Performance subtests (i.e., discriminant evidence), and vice versa. Corrections were made for inflated correlation coefficients by removing from composite scores the subtest with which it was being correlated. On the other hand, correlation coefficients between composite scores were not corrected and, as such, some were inflated.

More convergent evidence for WISC-III score interpretation was gathered by examination of WISC-III scores with other measures of cognition. That is, WISC-III scores were found to be highly correlated with other measures of intelligence (Wechsler, 1991). Correlation coefficients between the WISC-R and WISC-III were .89 for FSIQ, .90 for VIQ, and .81 for PIQ. Correlation coefficients between subtest scores for these

two measures ranged from .42 (Picture Arrangement) to .80 (Information). Wechsler (1991) noted that these correlation coefficients were of sufficient magnitude to conclude that the two measures assess fundamentally the same construct. Similar results were obtained for a clinical sample composed of children with learning and reading disabilities, attention-deficit hyperactivity disorder, and depression or anxiety disorders (Wechsler, 1991). That is, correlation coefficients were .86 between FSIQ scores, .86 between VIQ scores, and .73 between PIQ scores.

Correlation coefficients between WISC-III and WAIS-R scores for 16-year-olds were also high, again suggesting that both tests measure analogous constructs (Wechsler, 1991). The correlation coefficient between FSIQ scores was .86, the one between VIQ scores was .90, and the one between PIQ scores was .80. The range of correlation coefficients between subtest scores was .35 (Picture Arrangement) to .85 (Vocabulary). Similarly, high correlation coefficients were found between WISC-III and WPPSI-R scores for a group of 6-year-olds (.85 FSIQ, .85 VIQ, .73 PIQ) (Wechsler, 1991). Finally, the correlation coefficient findings revealed similarities between the WISC-III and the DAS (Wechsler, 1991). For example, the correlation coefficient between the WISC-III FSIQ scores and the DAS General Conceptual Ability scores was .92. The correlation coefficient was .78 between these two sets of scores for a sample of children with LDs and attention deficit hyperactivity disorder.

Evidence based on test-criterion relationships is also encompassed by evidence based on relations to other variables (AERA, APA, & NCME, 1999). This source of evidence explores the extent to which test scores predict attainment on a relevant criterion variable. One example would be determination of the accuracy with which scores on an

intellectual measure predict achievement results, given that cognitive functioning has been determined to be a predictor of achievement (e.g., see review by Flanagan, Andrews, & Genshaft, 1997). Evidence based on the relation of WISC-III scores to a relevant criterion was gathered by calculating correlation coefficients between the WISC-III scores and a number of measures of achievement (Wechsler, 1991). The correlation coefficient was substantial (r = .73) between WISC-III FSIQ scores and the Total School Ability Index scores of the Otis-Lennon School Ability Test Sixth Edition, Form I (Otis & Lennon, 1989). Correlation coefficients between WISC-III and DAS achievement test scores were also examined. Correlation coefficients were found to be .59 between WISC-III FD scores and DAS Spelling scores, .55 between PS and Spelling scores, .61 between FD and Word Reading, and .57 between PS and Word Reading. Wechsler (1991) noted that similar coefficients were obtained between these measures of achievement and FSIQ and VIQ scores. Correlation coefficients between WISC-III scores and scores on the Wide Range Achievement Test-Revised (WRAT-R; Jastak & Wilkinson, 1984) were calculated for a group of children with LDs and attention deficit hyperactivity disorder. Correlation coefficients between FSIQ scores and Reading, Spelling, and Arithmetic scores were found to be .53, .28, and .58, respectively.

Correlation coefficients were also calculated between WISC-III and achievement scores for a sample of students where each student was administered one of five group administered achievement tests (Wechsler, 1991). Coefficients were high; for example, the coefficient between the FSIQ and achievement scores was .74. Finally, school grades given by teachers across four subject areas (math, English, reading, and spelling) were found to correlate moderately with FSIQ scores (r = .47). Overall, Wechsler (1991)

concluded that "the correlations with other measures of intellectual ability and academic achievement provide support for the construct validity of the instrument" (p. 216).

Evidence based on internal structure is another source of validity evidence (AERA, APA, & NCME, 1999). This form of validity evidence is gathered by examining whether the internal structure of a test matches the structure that is expected based on the presumed underlying construct. Results of an exploratory factor analysis with the total standardization sample as well as with four age group subsamples, which altogether included all 2,200 children, supported VIQ and PIQ scores as well as a four factor solution (VC, PO, FD, PS) (Wechsler, 1991). A cross-validation study was conducted by testing the stability of factor score coefficients across samples (Wechsler, 1991). Participants were 440 children randomly drawn from the standardization sample. Next, based on seven random samples of 352 children drawn with replacement from the remaining 1,760 cases, factor score coefficients were derived. Median correlation coefficients for each factor score among the seven sets of factor scores for the cross-validation sample were above .9. Thus, there was further evidence to support a four factor solution containing the VC, PO, FD, and PS factor scores given that this configuration was stable across samples. In addition, confirmatory factor analysis also supported the four factor solution (Wechsler, 1991).

The four factor structure was also found to be the best fit for a group (n = 167) of children with learning disabilities, reading disorders, or attention-deficit disorders, as determined through confirmatory factor analysis (Wechsler, 1991). The only difference was that Picture Arrangement loaded on three factors. Confirmatory factor analysis also found the four factor solution to be the best fit for a group (n = 141) of children of low

ability, some of who were mentally retarded. The fact that the four factors is a sound solution for children with various disabilities is important given that most participants in the present study have identified disabilities such as LDs and mental retardation.

Procedures

WISC-III data for the current study was obtained through the work of Canivez and Watkins (1998). That is, participants of this study represented a subset of the sample used by Canivez and Watkins (1998) in a study exploring the long-term stability of the WISC-III. Canivez and Watkins (1998) sent out a request to 2,000 school psychologists who were randomly chosen from among members of the National Association of School Psychologists. School psychologists contacted by Canivez and Watkins (1998) were asked to submit demographic information as well as test and retest data for students who were twice tested with the WISC-III. That is, data was requested for students who had been evaluated at two points in time, each time in order to determine special education eligibility. No other criteria were specified, such as number of cases to report, age of children whose scores were reported, and so on. There was no requirement that intellectual functioning be assessed by the same school psychologist at both points in time. Finally, confidentiality was maintained as students' names were not requested.

WISC-III test-retest data was received for a total of 667 students. These scores were reported by 114 school psychologists in 33 different states, yielding a response rate of 5.70%. On average, 5.85 cases were reported per psychologist, with a range from 1 to 25 and a standard deviation of 3.93.

Profile Similarity Measures

The purpose of this study was to determine the long-term stability of nonlinear multivariate WISC-III cluster membership. Measures of profile similarity can be used to assess the stability of core profile membership over time (Reynolds, 1997). In order to determine whether cognitive profile membership remained stable over time, participants' profiles were compared to core subtest or core factor profiles at both Time 1 and Time 2. A measure of profile similarity was used to measure the likeness between a profile and the core profiles in the taxonomies, and thus to establish profile membership.

Profile similarity measures are the result of "finding a method of specifying the degree of similarity between any two specific profiles" (Moffitt et al., 1993, p. 460). Many profile similarity techniques exist (Cronbach & Gleser, 1953), and they are most commonly applied to the results of cluster analysis (Livingston et al., 2003). Different similarity measures provide different information (Cronbach & Gleser).

Euclidean distance measures. One measure of profile similarity is *D* (Cronbach & Gleser, 1953; Osgood & Suci, 1952). *D* is a measure of dissimilarity that is based on the Euclidean distance between two profiles (Cronbach & Gleser). That is, the sum of squared differences between each pair of points from the two profiles is calculated (D^2) and the square root of this number is then found. D^2 can also be used as a profile similarity measure, although it was deemed less desirable than *D* given the inflated results that are produced for bigger differences (Cronbach & Gleser). Euclidean distance measures are sensitive to all three profile dimensions and do not have any restrictive assumptions (Livingston et al., 2003).

On the other hand, the number of profile elements as well as the metric of the scores affect the value of Euclidean distance measures and, as such, some researchers may decide to divide the results by the number of elements and to convert the scores to a shared metric (Livingston et al., 2003). Further, these measures cannot be easily interpreted. That is, the value of D and of D^2 increases from 0 as profiles are progressively less similar, but more precise guidelines do not exist (i.e., the value of D^2 that might correspond to acceptable levels of stability).

Cattell's coefficient of profile similarity. Cattell's coefficient of profile similarity (r_p) is similar to D, but the metric has undergone a transformation (Cronbach & Gleser, 1953). As such, r_p is easy to understand because values can be interpreted in the same manner as the common correlation coefficient (Livingston et al., 2003). An r_p value of 1 indicates perfect similarity between two profiles, 0 represents no agreement, and -1 means an exact inverse relation. In addition, r_p is invariant across number of elements and across metric. Both r_p and Euclidean distance measures rank profiles in the same order of dissimilarity (Cronbach & Gleser), and like D and D^2, r_p is sensitive to all profile dimensions (Livingston et al.).

One difficulty associated with the use of r_p is that it assumes independence of element scores, which is not always the case. For example, WISC-III subtest scores, which are often the subject of profile analysis, are correlated (Wechsler, 1991). This underlying assumption is not restrictive when measuring profile similarity, though, given its relevance to significance testing rather than to examination of the absolute value (Livingston et al., 2003). Another problem related to r_p is the difficulty involved with its interpretation. Cronbach and Gleser (1953) asserted that having a measure of similarity

that can be interpreted like the correlation coefficient is not advantageous. For example, assigning a limit to the degree of dissimilarity between two profiles (i.e., -1) is not sensible as there is no boundary to how far apart two profiles can be (Cronbach & Gleser). Similarly, the idea that two profiles can be totally dissimilar holds no real meaning.

Q correlation. The *Q* correlation (Burt, 1937) was once a popular method of measuring profile similarity (Cronbach & Gleser, 1953). This measure represents the correlation coefficient between the elements of a pair of profiles and, as such, is sensitive only to shape (Livingston et al., 2003). The *Q* correlation can be generalized across studies and does not have restrictive underlying assumptions.

Cronbach and Gleser (1953) argued against removing level information from profiles in the calculation of profile similarity. For example, relying on a measure of profile similarity that does not take level into account may result in two profiles being judged similar even though they differ widely on the variable of interest, such as intelligence. On the other hand, Cronbach and Gleser noted that there may be occasions when elevation information is not of concern. In general, if the information conveyed by the level data is meaningful, it should not be disregarded in profile similarity analyses. Further, removal of scatter from profiles is not recommended as shape scores are unreliable if scatter is not large compared to error (Cronbach & Gleser). That is, if some profiles are flat, analysis involving only shape will be swayed by error.

Similarity measure employed in the current study. Overall, then, similarity measures that disregard elevation and scatter information are not generally preferred (Cronbach & Gleser, 1953). Further, level information should not be ignored when examining stability

of cognitive profiles, as in this study, given that elevation information (i.e., intelligence) is the variable of interest. Finally, a similarity measure representing all profile dimensions was desirable in order to be consistent with the nonlinear multivariate nature of core profiles belonging to the WISC-III subtest (Konold et al., 1999) and factor (Donders, 1996) profile taxonomies.

Given that Euclidean distance and r_p rank profiles in the same order of dissimilarity (Cronbach & Gleser, 1953), both these similarity measures would have been adequate for the present study. That is, decisions regarding profile-type membership would have been the same no matter which measure of similarity was employed. D^2 was chosen as the similarity measure because of the difficulties associated with the interpretation of r_p (Cronbach & Gleser).

Core Profile Membership or Designation as Unusual

In order to calculate the long-term classification stability of membership to WISC-III subtest and factor profile taxonomies, classification decisions were made for participants at both Time 1 and Time 2 based on the D^2 similarity measure. The D^2 value was calculated between each individual's WISC-III profile and every core profile in Konold et al.'s (1999) subtest taxonomy based on 10 subtests (Sample 1) or, for participants belonging to Sample 2, between WISC-III profiles and core profiles in Donders' (1996) core factor taxonomy. Classification to a profile type was based on the lowest D^2 value, as this indicated the greatest similarity.

Participants whose profiles were unlike all core profiles were classified as *unusual*. Consistent with core profile membership, the designation of unusual was also determined by D^2. That is, using the method outlined by Konold et al. (1999), a value above the

critical D^2 identified a participant as belonging to a subgroup representing the 5% of children having profiles most discrepant from core profiles. As discussed by Konold et al., it is likely that most psychologists would consider a rate of 5% to represent few enough children such that they should be considered unusual. Further, re-classification of normative sample children to the WISC-III taxonomy based on 10 mandatory subtests (Konold et al.) according to D^2 was found to be most accurate when those in the most discrepant 5.4% were considered unusual as compared to lower percentages (e.g., rarest 4%; Konold et al.). For the taxonomy that they derived, Konold et al. (1999) calculated the critical D^2 value to be 98. That is, in the event that every D^2 value calculated between a given Sample 1 participant's profile and each profile in the taxonomy was $\geq$ 98, the profile was determined to be unusual compared to the general population.

In order to determine the critical D^2 value for the taxonomy of core factor profiles (Donders, 1996) an identical procedure was used (Konold et al., 1999). The D^2 values between the profiles of all 2,200 students in the WISC-III standardization sample and each of the core factor profiles were calculated and the most discrepant 5.4% (n = 118) were identified (i.e., those whose profiles were least similar to any core factor profile type). The critical D^2 value for the WISC-III factor taxonomy was determined to be 820.14. When every D^2 value calculated between a Sample 2 participant's profile and each profile in the taxonomy was $\geq$ 820.14, the profile was designated unusual.

Donders (1996) suggested another method of determining whether a given factor profile should be considered unusual with reference to the WISC-III factor profile taxonomy. However, unlike the D^2 technique, this method focuses on differences in profile elevation rather than considering all profile dimensions simultaneously.

Specifically, the guidelines provided by Donders are based on the number of core profile scores that fall outside the 90% confidence intervals of the observed scores.

Once 90% confidence intervals are obtained for a given child's four factor scores, these intervals are compared to the factor scores of the core profile defined by level that has an average FSIQ most similar to the child's (Donders, 1996). In addition, the observed 90% confidence intervals are compared to the factor scores belonging to both shape-defined core profiles in the taxonomy. Donders advocated that a profile be considered unusual when at least three fourths of core profile scores from level-defined profiles fall outside the 90% confidence intervals, and when at least half of core profile scores from both shape-defined profiles fall outside 90% confidence limits. With respect to the three-fourths application, more conservative criteria were not suggested as alpha is inflated each time a student's 90% confidence interval for a given score is compared to the corresponding score in the core profile. Also, less stringent guidelines were given for profiles defined by shape as "one would typically be interested in the potential clinical relevance of relative levels of elevation of individual (i.e., not all) specific factor scores" (Donders, p. 316). Table 6 outlines the steps taken to determine whether a profile is unusual according to this *standard error method*. In order to take Donders' suggestion into consideration, both the D^2 procedure as well as the standard error technique were used to identify unusual factor profiles (Sample 2), despite the limitations of the latter procedure. Once unusual profiles as defined by Donders were removed, remaining participants' profiles were classified as members of core profile types based on the lowest D^2.

Table 6

Steps to Determine Whether a Factor Profile is Unusual According to the Standard Error Method

1.	Determine the 90% confidence intervals for each of the child's four factor scores
2.	Based on the child's FSIQ score, determine the level-defined core profile that has an FSIQ most similar in value
3.	Determine the number of factor scores of the level-defined core profile selected in Step 2 that fall outside of the child's 90% confidence intervals
4.	Repeat Step 3 for both core profiles defined by shape
5.	If at least three of the four core profile scores from level-defined profile fall outside the 90% confidence intervals *and* if at least two of the four core profile scores from *both* shape-defined profiles fall outside 90% confidence limits, then a profile can be considered unusual

Note. Donders, 1996.

Determination of Profile Membership Stability

Classification of participants to core profiles or designation as unusual was repeated at both Time 1 and Time 2 for both Sample 1 and Sample 2; in addition, categorization as unusual was determined in two separate ways for members of Sample 2. Classification stability was calculated across time using Fleiss' k_m. A k_m coefficient yields the percent agreement of profile classification between Time 1 and Time 2 corrected for agreement due to chance. An extension of Cohen's k, k_m can be applied to cases with more than two raters and can be used in instances where the raters may vary with each observation, although the number of raters must stay constant (Fleiss). The present study reported k_m

coefficients in order to be consistent with studies that have examined the short-term classification stability of cognitive profiles (Holland & McDermott, 1996; Glutting & McDermott, 1990a, 1990b; Glutting et al., 1992; McDermott, Glutting, Jones, Watkins, & Kush, 1989).

MacKappa (Watkins, 1998) was used to calculate general and partial k_m coefficients across Time 1 and Time 2. That is, for classification choices made for Sample 1, an overall or general k_m coefficient was calculated in addition to 9 partial k_m coefficients, representing each core profile type as well the group of unusual profiles. Similarly, for Sample 2 when the critical D^2 technique of designating profiles as unusual was utilized, an overall k_m coefficient was calculated and 6 partial k_m coefficients were calculated, one for each core profile type and one for unusual profiles. Finally, general and partial k_m coefficients were again calculated for Sample 2 when the standard error method was used to identify unusual profiles. The level of statistical significance was determined for each k_m coefficient calculated.

Although not theoretically or empirically supported, different guidelines have appeared in the literature regarding the interpretation of kappa statistics in terms of clinical significance (e.g., Fleiss, Levin, & Paik, 2003; Landis & Koch, 1977). Cicchetti (1994) summarized suggested interpretations: A kappa coefficient less than .40 indicates poor clinical significance; a kappa value between .40 and .59 is considered fair; good clinical significance is defined as a kappa coefficient ranging form .60 to .74; and a kappa coefficient of .75 and above is excellent.

Based on these guidelines, it was decided a priori, with respect to the current study, that statistically significant general and partial k_m coefficients of $\geq$ .40 would indicate

that future research is warranted in order to determine whether nonlinear multivariate profile type membership information is useful when making educational decisions. Although much higher values must be obtained before making decisions about individual students in order to avoid deleterious practice (e.g., a correlation coefficient of at least .90 was recommended by Salvia and Ysseldyke [2001]), there would be merit in further studying core profile taxonomies or individual core profiles that display at least a fair degree of clinical significance with respect to classification stability. By choosing this conservative cutoff point, helpful practices in educational decision-making based on nonlinear multivariate profile analysis would not be overlooked. On the other hand, for k_m coefficients found to be < .40 there would be no support for conducting further research to determine whether there is evidence of validity for the interpretation and use of core profile membership information. Reliability is a prerequisite for validity (AERA, APA, & NCME, 1999) and, therefore, poor classification stability of core profile membership over time necessarily excludes the possibility that these membership decisions are valid.

Two final analyses were conducted to determine whether the number of unusual cases, one form of outlier, or the amount of instability across time varied with region of the country, state, or reporting psychologist. Regional practices or personal styles may have interfered with final test results, increasing the number of unusual cases or the degree of instability unproportionally. That is, findings of widely discrepant instances of unusual WISC-III profiles or amount of instability across region or psychologist may be an indication that designation as unusual or determination of stability are functions of WISC-III administration and/or scoring practices. Should this be the case, classification stability results would be greatly distorted.

The multivariate outliers considered in this study were profiles determined to be unusual with respect to the profiles in the WISC-III subtest or factor taxonomies. The extent to which unusual profile classification corresponded to region of the country was evaluated by first dividing the country into the four regions identified in the WISC-III manual: West, South, North Central, and Northeast. The percentage of unusual cases across region as determined by the D^2 method was then calculated at Time 1 and Time 2 for both Sample 1 and Sample 2. In addition, a separate analysis was conducted for Sample 2 using the standard error method of identifying unusual cases. Finally, visual inspection was used to determine whether the percentage of students with unusual profiles differed considerably across the four regions. In order to determine whether an apparent correspondence existed between the number of unusual cases reported and state or psychologist, visual inspection of profile type membership was conducted where a sufficient number of cases were reported.

General k_m coefficients were visually compared across the four geographic regions in order to determine whether instability had an obvious relation to area of the country. In addition, where an adequate number of cases were reported, the degree of instability was informally investigated for correspondence with particular psychologists or various states.

Results

Results for Sample 1

WISC-III Data

Data for all 10 WISC-III mandatory subtests were available for Sample 1 participants. Table 7 displays the IQ, index, and subtest scores for this sample at both Time 1 and Time 2.

Table 7

Means and Standard Deviations of WISC-III IQ, Index, and Subtest Scores for Sample 1 at Both Time 1 and Time 2

	Time 1		Time 2	
	M	*SD*	*M*	*SD*
FSIQ	88.68	16.12	88.41	17.00
VIQ	88.71	15.84	88.18	15.79
PIQ	90.79	16.75	90.83	17.91
VC	90.32	15.77	89.77	15.71
PO	91.81	17.11	92.66	18.60
FD	85.80[a]	14.46[a]	85.43[b]	14.13[b]
PS	92.06[c]	15.70[c]	90.33[d]	15.63[d]
PC	8.70	3.35	9.05	3.38
IN	7.75	3.13	7.94	3.16
CD	8.30	3.42	7.68	3.25

(table continues)

Table 7 (continued)

	Time 1		Time 2	
	M	*SD*	*M*	*SD*
SM	8.20	3.40	8.39	3.23
PA	8.47	3.58	8.67	3.92
AR	7.25	3.08	7.18	2.95
BD	8.41	3.72	8.30	4.03
VO	8.02	3.22	7.49	3.13
OA	8.42	3.38	8.52	3.62
CM	8.66	3.72	8.40	3.53
SS	8.43[e]	3.65[e]	8.60[d]	3.55[d]
DS	7.34[f]	2.70[f]	7.39[g]	2.77[g]

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 503. [b]Due to missing data, results are based on a sample size of 484. [c]Due to missing data, results are based on a sample size of 250. [d]Due to missing data, results are based on a sample size of 277. [e]Due to missing data, results are based on a sample size of 247. [f]Due to missing data, results are based on a sample size of 497. [g]Due to missing data, results are based on a sample size of 483.

Descriptive Information for Participants Belonging to the Various Profile Types

Participants in Sample 1 were divided into 9 different profile types: Profiles 1 through 8 as defined by Konold et al. (1999), and a profile type reserved for those determined to have unusual profiles in accordance with the critical D^2 method. The

number of children in each profile type is displayed in Table 8 along with information about student age for both Time 1 and Time 2. Descriptive information for Sample 1 participants across the different profiles can be found in Table 9 for Time 1 and Table 10 for Time 2. Finally, Tables 11 and 12 display mean WISC-III IQ, index, and subtest scores for Sample 1 at Time 1 and Time 2 across profile types.

Table 8

Number of Children and Mean, Standard Deviation, and Range of Ages for Children from Sample 1 Across Profile Types at Both Time 1 and Time 2

	Time 1				Time 2			
		Age (years)				Age (years)		
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range
Profile 1	15	8.7	2.2	6.0 - 13.3	12	11.4	2.1	8.8 – 16.2
Profile 2	25	9.0	1.9	6.0 – 12.9	40	11.7	2.0	8.9 – 16.0
Profile 3	15	8.3	.9	7.0 – 9.6	11	10.7	1.4	9.2 – 14.0
Profile 4	58	8.6	2.0	6.0 – 13.9	50	11.6	1.6	9.0 – 15.5
Profile 5	74	9.2	1.7	6.0 – 13.2	81	12.1	1.8	8.0 – 16.8
Profile 6	127	9.3	2.0	6.0 – 13.9	127	12.0	2.2	8.3 – 16.9
Profile 7	92	8.8	1.8	6.0 – 12.7	75	11.8	2.2	7.9 – 16.6
Profile 8	116	9.7	2.3	6.0 – 14.6	123	12.1	2.2	7.5 – 16.6
Unusual profile	63	9.2	2.3	6.0 – 14.0	66	12.4	2.3	7.6 – 16.8

Table 9

Percent of Sample 1 Participants at Time 1 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
Gender									
Boys	60	84	60	71	78	68	65	58	68
Girls	40	16	40	29	22	32	35	42	32
Race/Ethnicity									
White	100	92	100	79	81	80	70	67	70
Black	0	0	0	7	11	11	21	24	21
Hispanic	0	0	0	9	3	5	9	7	6
Native American	0	0	0	2	0	2	0	0	2
Asian/Pacific	0	0	0	0	0	0	1	0	0
Other	0	4	0	0	1	1	0	1	0
Missing	0	4	0	3	4	2	0	1	2
Disability									
Not disabled	7	4	0	0	0	3	3	7	2
LD	67	64	73	91	82	77	63	30	41
MR	0	0	0	0	0	2	1	29	32
ED	13	16	7	2	8	5	5	8	13
SLD	0	0	0	2	1	2	3	6	2

(table continues)

Table 9 (continued)

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
Other disabilities	7	4	13	0	3	2	8	16	5
Unspecified	7	12	7	5	5	9	16	3	6
Geographic Region									
West	0	20	27	34	23	22	26	14	19
South	33	28	40	40	36	33	33	41	35
North Central	40	28	27	24	30	36	29	33	32
Northeast	27	24	7	2	11	9	12	13	14

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 10

Percent of Sample 1 Participants at Time 2 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
Gender									
Boys	83	88	73	62	81	73	60	50	68
Girls	17	13	27	38	19	27	40	50	32
Race/Ethnicity									
White	100	88	82	82	77	77	76	63	85
Black	0	5	9	12	16	11	17	27	6
Hispanic	0	3	9	0	4	9	7	7	5
Native American	0	0	0	2	0	1	0	1	2
Asian/Pacific	0	0	0	0	0	0	0	1	0
Other	0	3	0	0	0	1	0	1	2
Missing	0	3	0	4	4	2	0	1	2
Disability									
Not disabled	0	8	0	6	5	5	11	8	2
LD	75	70	73	72	70	76	57	36	41
MR	0	0	0	0	0	2	0	20	38
ED	8	13	0	4	11	4	11	6	6
SLD	0	0	0	0	0	2	3	7	0

(table continues)

Table 10 (continued)

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
Other disabilities	8	0	0	4	7	2	9	15	6
Unspecified	8	10	27	14	6	9	9	9	8
Geographic Region									
West	0	10	27	24	25	31	16	14	29
South	33	35	45	36	35	31	36	44	30
North Central	50	33	18	32	33	28	31	34	29
Northeast	17	23	9	8	7	10	17	8	12

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 11

Mean WISC-III IQ, Index, and Subtest Scores for Sample 1 at Time 1 Across Profile Types

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
FSIQ	126	112	108	102	97	88	86	71	82
VIQ	124	110	111	96	100	84	91	72	84
PIQ	123	114	104	109	95	96	83	73	83
VC	124	111	111	97	102	85	93	74	87
PO	123	117	100	107	98	98	83	74	84
FD	114[a]	97[b]	102	93[c]	91[d]	85[e]	86[f]	74[g]	79[h]
PS	119[i]	106[j]	114[k]	108[l]	90[m]	93[n]	88[o]	82[c]	83[p]
PC	13	13	10	11	10	10	8	6	7
IN	14	11	12	8	10	7	8	5	7
CD	12	9	13	12	7	8	8	7	7
SM	14	12	11	9	11	7	9	5	8
PA	14	13	11	12	9	9	8	5	7
AR	13	10	11	9	9	7	7	5	5
BD	15	13	10	11	10	10	6	5	7
VO	15	12	12	9	10	7	9	5	8
OA	13	12	10	11	9	10	7	6	7
CM	14	13	13	11	11	7	10	6	7

(table continues)

Table 11 (continued)

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
SS	15[i]	12[j]	12[k]	11[l]	9[m]	9[q]	7[r]	6[s]	7[p]
DS	12[a]	9[b]	9	9[h]	8[t]	7[u]	8[f]	6[g]	6[v]

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 13. [b]Due to missing data, results are based on a sample size of 19. [c]Due to missing data, results are based on a sample size of 53. [d]Due to missing data, results are based on a sample size of 68. [e]Due to missing data, results are based on a sample size of 103. [f]Due to missing data, results are based on a sample size of 81. [g]Due to missing data, results are based on a sample size of 100. [h]Due to missing data, results are based on a sample size of 51. [i]Due to missing data, results are based on a sample size of 6. [j]Due to missing data, results are based on a sample size of 11. [k]Due to missing data, results are based on a sample size of 7. [l]Due to missing data, results are based on a sample size of 28. [m]Due to missing data, results are based on a sample size of 36. [n]Due to missing data, results are based on a sample size of 47. [o]Due to missing data, results are based on a sample size of 39. [p]Due to missing data, results are based on a sample size of 23. [q]Due to missing data, results are based on a sample size of 46. [r]Due to missing data, results are based on a sample size of 38. [s]Due to missing data, results are based on a sample size of 52. [t]Due to missing data, results are based on a sample size of 67. [u]Due to missing data, results are based on a sample size of 102. [v]Due to missing data, results are based on a sample size of 49.

Table 12

Mean WISC-III IQ, Index, and Subtest Scores for Sample 1 at Time 2 Across Profile Types

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
FSIQ	126	112	106	104	98	87	85	72	77
VIQ	124	110	109	97	100	84	90	74	77
PIQ	124	113	103	110	97	94	82	74	82
VC	124	111	111	98	102	85	91	76	80
PO	124	118	99	109	101	97	83	74	84
FD	108[a]	101[b]	100[c]	95[d]	92[e]	83[f]	87[g]	76[h]	75[i]
PS	116[j]	100[k]	119[l]	111[m]	91[n]	92[o]	88[p]	83[q]	79[b]
PC	13	13	9	11	11	10	8	6	8
IN	14	12	12	10	10	7	8	5	6
CD	12	9	13	12	7	8	8	6	6
SM	14	12	12	10	11	8	9	6	7
PA	14	13	10	13	9	9	8	6	7
AR	13	10	10	9	9	7	8	5	5
BD	16	13	11	11	10	9	6	5	7
VO	14	12	11	9	9	6	8	5	6
OA	13	13	8	11	10	10	6	6	7
CM	14	12	13	10	10	8	9	6	6

(table continues)

Table 12 (continued)

	Profile type								
	1	2	3	4	5	6	7	8	Unusual
SS	14[j]	12[k]	15[l]	13[m]	9[n]	9[o]	8[p]	7[q]	7[b]
DS	9[a]	10[b]	10[c]	9[d]	8[e]	7[f]	8[g]	6[r]	6[i]

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 11. [b]Due to missing data, results are based on a sample size of 34. [c]Due to missing data, results are based on a sample size of 10. [d]Due to missing data, results are based on a sample size of 41. [e]Due to missing data, results are based on a sample size of 66. [f]Due to missing data, results are based on a sample size of 98. [g]Due to missing data, results are based on a sample size of 64. [h]Due to missing data, results are based on a sample size of 104. [i]Due to missing data, results are based on a sample size of 56. [j]Due to missing data, results are based on a sample size of 4. [k]Due to missing data, results are based on a sample size of 27. [l]Due to missing data, results are based on a sample size of 2. [m]Due to missing data, results are based on a sample size of 17. [n]Due to missing data, results are based on a sample size of 35. [o]Due to missing data, results are based on a sample size of 63. [p]Due to missing data, results are based on a sample size of 40. [q]Due to missing data, results are based on a sample size of 55. [r]Due to missing data, results are based on a sample size of 103.

Profile Membership Agreement Across Time

A general k_m coefficient that represented all profile types was computed as were partial k_m coefficients for each individual category. Table 13 displays general and partial k_m coefficients for Konold et al.'s (1999) WISC-III subtest taxonomy for the 10 mandatory subtest scores using Sample 1.

Table 13

General and Partial k_m Coefficients for the WISC-III Subtest Taxonomy for the 10 Mandatory Subtest Scores (Konold et al., 1999) Using Sample 1

	k_m coefficient
General k_m	.39*
Partial k_m	
Profile 1	.43
Profile 2	.40
Profile 3	.37
Profile 4	.35
Profile 5	.32
Profile 6	.43*
Profile 7	.36*
Profile 8	.51*
Unusual	.26

* $p < .002$ (with Bonferroni correction adjusting for 24 comparisons; experimentwise error rate = .05).

Results for Sample 2 (Unusual Cases Defined by the Critical D^2 Method)

WISC-III Data

Data for all four WISC-III index scores were available for Sample 2 participants. Table 14 displays the IQ, index, and subtest scores for this sample at both Time 1 and Time 2.

Table 14

Means and Standard Deviations of WISC-III IQ, Index, and Subtest Scores for Sample 2 at Both Time 1 and Time 2

	Time 1		Time 2	
	M	*SD*	*M*	*SD*
FSIQ	89.37	14.93	88.21	15.98
VIQ	89.78	15.55	88.30	15.51
PIQ	91.03	15.60	90.33	16.31
VC	91.56	15.52	89.92	15.74
PO	92.11	16.00	92.55	17.66
FD	85.51	15.15	85.60	13.42
PS	92.85	16.18	90.85	14.83
PC	8.66[a]	3.26[a]	9.01	3.27
IN	7.99[a]	3.14[a]	7.95	3.16
CD	8.36[a]	3.31[a]	7.37	2.82
SM	8.40[a]	3.13[a]	8.38	3.08
PA	8.59[a]	3.43[a]	8.67	3.66
AR	7.28[a]	3.16[a]	7.16	2.70
BD	8.41[a]	3.41[a]	8.21	3.66
VO	8.31[a]	3.36[a]	7.42	3.11
OA	8.59[a]	3.24[a]	8.57	3.54

(table continues)

Table 14 (continued)

	Time 1		Time 2	
	M	*SD*	*M*	*SD*
CM	8.83[b]	3.68[b]	8.54	3.57
SS	8.50[b]	3.78[b]	8.70	3.42
DS	7.26[c]	2.94[c]	7.42	2.69

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 176. [b]Due to missing data, results are based on a sample size of 175. [c]Due to missing data, results are based on a sample size of 174.

Descriptive Information for Participants Belonging to the Various Profile Types

Participants in Sample 2 were divided into 6 different profile types: Profiles 1 through 5 as defined by Donders (1996), and a profile type reserved for those determined to have unusual profiles in accordance with the critical D^2 method. The number of children in each profile type is displayed in Table 15 along with information about student age for both Time 1 and Time 2. Descriptive information for Sample 2 participants across the different profiles can be found in Table 16 for Time 1 and Table 17 for Time 2. Finally, Tables 18 and 19 display mean WISC-III IQ, index, and subtest scores for Sample 2 at Time 1 and Time 2 across profile types.

Table 15

Number of Children and Mean, Standard Deviation, and Range of Ages for Children from Sample 2 (Unusual Cases Defined by the Critical D^2 Method) Across Profile Types at Both Time 1 and Time 2

	Time 1				Time 2			
		Age (years)				Age (years)		
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range
Profile 1	24	8.2	1.5	6.3 – 12.0	14	11.4	1.6	9.3 – 14.9
Profile 2	56	9.1	1.9	6.0 – 13.1	66	11.9	1.8	9.1 – 16.0
Profile 3	5	7.5	1.4	6.0 – 9.4	5	10.6	1.4	9.0 – 12.8
Profile 4	50	9.1	1.3	6.0 – 11.4	44	11.7	1.5	7.9 – 15.9
Profile 5	10	8.9	1.4	7.0 – 11.7	20	11.7	2.1	9.2 – 16.0
Unusual profile	32	8.8	2.1	6.0 – 13.0	28	11.7	2.0	7.5 – 15.0

Table 16

Percent of Sample 2 Participants (Unusual Cases Defined by the Critical D^2 Method) at Time 1 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type					
	1	2	3	4	5	Unusual
Gender						
Boys	67	55	60	82	100	63
Girls	33	45	40	18	0	38
Race/Ethnicity						
White	75	82	100	80	100	84
Black	8	13	0	10	0	6
Hispanic	13	5	0	8	0	6
Native American	4	0	0	0	0	3
Asian/Pacific	0	0	0	0	0	0
Other	0	0	0	2	0	0
Missing	0	0	0	0	0	0
Disability						
Not disabled	0	4	0	2	10	0
LD	83	48	100	80	80	41
MR	0	5	0	0	0	41
ED	0	9	0	4	0	3

(table continues)

Table 16 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SLD	4	5	0	2	0	6
Other disabilities	4	16	0	2	10	6
Unspecified	8	13	0	10	0	3
Geographic Region						
West	33	16	0	34	20	16
South	33	21	40	16	20	44
North Central	17	32	40	36	30	31
Northeast	17	30	20	14	30	9

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 17

Percent of Sample 2 Participants (Unusual Cases Defined by the Critical D^2 Method) at Time 2 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type					
	1	2	3	4	5	Unusual
Gender						
Boys	93	56	80	80	80	57
Girls	7	44	20	20	20	43
Race/Ethnicity						
White	86	83	100	75	100	75
Black	0	8	0	16	0	14
Hispanic	7	9	0	7	0	7
Native American	7	0	0	0	0	4
Asian/Pacific	0	0	0	0	0	0
Other	0	0	0	2	0	0
Missing	0	0	0	0	0	0
Disability						
Not disabled	0	3	0	0	15	0
LD	79	55	80	70	65	46
MR	0	8	0	0	0	39
ED	0	5	0	11	5	0

(table continues)

Table 17 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SLD	0	6	0	0	0	4
Other disabilities	0	14	0	9	0	11
Unspecified	21	11	20	9	15	0
Geographic Region						
West	21	21	0	27	20	29
South	29	24	40	25	30	25
North Central	14	35	60	32	30	25
Northeast	36	20	0	16	20	21

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 18

Mean WISC-III IQ, Index, and Subtest Scores for Sample 2 (Unusual Cases Defined by the Critical D^2 Method) at Time 1 Across Profile Types

	Profile type					
	1	2	3	4	5	Unusual
FSIQ	101	80	127	94	111	77
VIQ	99	80	122	97	113	77
PIQ	104	82	127	93	107	82

(table continues)

Table 18 (continued)

	Profile type					
	1	2	3	4	5	Unusual
VC	100	82	123	99	113	80
PO	102	83	124	96	111	83
FD	95	80	118	91	105	69
PS	114	87	125	90	100	85
PC	10	7[a]	12	10	11	7
IN	10	6[a]	13	9	12	6
CD	12	8[a]	14	7	8	7
SM	10	7[a]	13	10	12	6
PA	10	7[a]	17	10	11	6
AR	9	6[a]	13	8	11	4
BD	10	7[a]	13	9	12	7
VO	10	6[a]	15	10	12	6
OA	10	7[a]	14	9	12	7
CM	10[b]	7[a]	14	11	14	6
SS	13[b]	7[a]	15	8	11	7
DS	8[b]	6[a]	13	8	10	4[c]

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 55. [b]Due to missing data, results are based on a sample size of 23. [c]Due to missing data, results are based on a sample size of 31.

Table 19

Mean WISC-III IQ, Index, and Subtest Scores for Sample 2 (Unusual Cases Defined by the Critical D^2 Method) at Time 2 Across Profile Types

	Profile type					
	1	2	3	4	5	Unusual
FSIQ	97	79	122	93	110	76
VIQ	94	81	123	93	109	75
PIQ	101	81	116	95	109	82
VC	96	82	123	94	111	78
PO	99	83	116	99	113	84
FD	93	81	106	91	101	71
PS	110	86	121	90	102	82
PC	9	8	13	10	12	8
IN	9	6	14	9	12	6
CD	11	6	12	7	8	6
SM	9	7	13	9	12	6
PA	10	7	14	10	12	6
AR	8	6	13	8	10	5
BD	11	6	11	9	12	6
VO	9	6	14	8	11	5
OA	10	7	12	10	12	8
CM	10	7	16	9	13	6

(table continues)

Table 19 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SS	12	8	16	8	12	7
DS	9	7	9	8	10	5

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

Profile Membership Agreement Across Time

A general k_m coefficient that represented all profile types was computed as were partial k_m coefficients for each individual category. Table 20 displays general and partial k_m coefficients for Donders' (1996) WISC-III factor taxonomy using Sample 2, where unusual profiles were identified using the critical D^2 method.

Table 20

General and Partial k_m Coefficients for the WISC-III Factor Taxonomy (Donders, 1996) Using Sample 2 (Unusual Cases Defined by the Critical D^2 Method)

	k_m coefficient
General k_m	.37*
Partial k_m	
Profile 1	.17
Profile 2	.45
Profile 3	.59
Profile 4	.36
Profile 5	.34
Unusual	.36

* $p < .002$ (with Bonferroni correction adjusting for 24 comparisons; experimentwise error rate = .05).

Results for Sample 2 (Unusual Cases Defined by the Standard Error Method)

WISC-III Data

WISC-III data for Sample 2 participants were already presented and can be found in Table 14.

Descriptive Information for Participants Belonging to the Various Profile Types

Participants in Sample 2 were also divided into 6 different profile types as follows: Profiles 1 through 5 as defined by Donders (1996) and a profile type reserved for those determined to have unusual profiles as defined by the standard error method. The number of children in each profile type is displayed in Table 21 along with information about student age at both Time 1 and Time 2. Descriptive information for Sample 2 participants

across the different profiles can be found in Table 22 for Time 1 and Table 23 for Time 2. Finally, Tables 24 and 25 display mean WISC-III IQ, index, and subtest scores for Sample 2 at Time 1 and Time 2 across profile types.

Table 21

Number of Children and Mean, Standard Deviation, and Range of Ages for Children from Sample 2 (Unusual Cases Defined by the Standard Error Method) Across Profile Types at Both Time 1 and Time 2

	Time 1				Time 2			
		Age (years)				Age (years)		
	n	*M*	*SD*	Range	*n*	*M*	*SD*	Range
Profile 1	16	8.4	1.5	6.6 – 12.0	10	11.3	1.3	9.4 – 13.3
Profile 2	51	9.1	1.9	6.0 – 13.1	58	11.7	1.9	7.5 – 15.6
Profile 3	4	7.4	1.6	6.0 – 9.4	4	10.7	1.6	9 – 12.8
Profile 4	49	9.0	1.4	6.0 – 11.4	34	11.8	1.7	7.9 – 14.3
Profile 5	9	8.9	1.6	7.0 – 11.7	17	11.9	2.0	9.9 – 16
Unusual profile	48	8.7	1.9	6.0 – 13.0	54	11.7	1.8	8.0 – 16.0

Table 22

Percent of Sample 2 Participants (Unusual Cases Defined by the Standard Error Method) at Time 1 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type					
	1	2	3	4	5	Unusual
Gender						
Boys	69	49	50	84	100	69
Girls	31	51	50	16	0	31
Race/Ethnicity						
White	88	78	100	80	100	83
Black	13	14	0	10	0	8
Hispanic	0	8	0	8	0	4
Native American	0	0	0	0	0	4
Asian/Pacific	0	0	0	0	0	0
Other	0	0	0	2	0	0
Missing	0	0	0	0	0	0
Disability						
Not disabled	0	2	0	0	11	4
LD	81	39	100	84	78	58
MR	0	12	0	0	0	21
ED	0	10	0	4	0	2

(table continues)

Table 22 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SLD	6	6	0	2	0	2
Other disabilities	6	20	0	2	11	4
Unspecified	6	12	0	8	0	8
Geographic Region						
West	31	14	0	35	11	23
South	38	24	25	20	11	33
North Central	19	31	50	31	44	31
Northeast	13	31	25	14	33	13

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 23

Percent of Sample 2 Participants (Unusual Cases Defined by the Standard Error Method) at Time 2 Distributed Across Gender, Race/Ethnicity, Disability, and Geographic Region for Each Profile Type

	Profile type					
	1	2	3	4	5	Unusual
Gender						
Boys	100	55	75	85	82	61
Girls	0	45	25	15	18	39
Race/Ethnicity						
White	90	83	100	76	100	78
Black	0	9	0	15	0	11
Hispanic	10	9	0	6	0	7
Native American	0	0	0	0	0	0
Asian/Pacific	0	0	0	0	0	4
Other	0	0	0	3	0	0
Missing	0	0	0	0	0	0
Disability						
Not disabled	0	3	0	0	18	9
LD	80	57	75	71	59	56
MR	0	7	0	0	0	22
ED	0	3	0	9	6	6

(table continues)

Table 23 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SLD	0	7	0	0	0	2
Other disabilities	0	12	0	12	0	0
Unspecified	20	10	25	9	18	6
Geographic Region						
West	20	21	0	29	24	24
South	20	24	50	21	24	31
North Central	20	34	50	35	35	24
Northeast	40	21	0	15	18	20

Note. LD = Learning disability; MR = Mental retardation; ED = Emotional disability; SLD = Speech and language disability.

Table 24

Mean WISC-III IQ, Index, and Subtest Scores for Sample 2 (Unusual Cases Defined by the Standard Error Method) at Time 1 Across Profile Types

	Profile type					
	1	2	3	4	5	Unusual
FSIQ	103	78	125	95	112	84
VIQ	100	79	120	96	116	84
PIQ	106	80	126	95	105	88

(table continues)

Table 24 (continued)

	Profile type					
	1	2	3	4	5	Unusual
VC	100	80	120	98	117	87
PO	103	81	123	98	109	88
FD	98	78	114	90	108	79
PS	114	86	125	91	100	91
PC	11	7[a]	13	10	11	8
IN	10	6[a]	13	9	13	7
CD	12	8[a]	14	8	8	8
SM	9	6[a]	13	10	12	8
PA	10	7[a]	17	10	11	8
AR	11	6[a]	12	8	12	6
BD	10	6[a]	12	9	12	8
VO	10	6[a]	14	9	13	7
OA	10	7[a]	14	9	11	8
CM	11	6[a]	14	10	15	7[b]
SS	13	7[a]	15	8	12	8[b]
DS	8	6[a]	12	8	11	6[c]

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

[a]Due to missing data, results are based on a sample size of 50. [b]Due to missing data, results are based on a sample size of 47. [c]Due to missing data, results are based on a sample size of 46.

Table 25

Mean WISC-III IQ, Index, and Subtest Scores for Sample 2 (Unusual Cases Defined by the Standard Error Method) at Time 2 Across Profile Types

	Profile type					
	1	2	3	4	5	Unusual
FSIQ	98	79	120	94	110	84
VIQ	95	80	118	94	110	83
PIQ	102	81	118	96	109	87
VC	97	81	119	95	111	85
PO	101	83	117	99	113	89
FD	93	80	100	91	104	80
PS	110	85	124	89	102	89
PC	9	8	13	10	12	8
IN	10	7	13	9	12	7
CD	11	6	13	7	8	7
SM	9	7	13	10	12	7
PA	10	7	14	9	12	8
AR	8	6	12	8	11	6
BD	11	6	11	9	12	7
VO	9	6	13	8	11	7
OA	9	7	12	10	11	8
CM	10	7	15	9	13	8

(table continues)

Table 25 (continued)

	Profile type					
	1	2	3	4	5	Unusual
SS	12	7	16	8	12	8
DS	9	7	8	8	10	6

Note. FSIQ = Full Scale IQ; VIQ = Verbal IQ; PIQ = Performance IQ; VC = Verbal Comprehension; PO = Perceptual Organization; FD = Freedom From Distractability; PS = Processing Speed; PC = Picture Completion; IN = Information; CD = Coding; SM = Similarities; PA = Picture Arrangement; AR = Arithmetic; BD = Block Design; VO = Vocabulary; OA = Object Assembly; CM = Comprehension; SS = Symbol Search; DS = Digit Span.

Profile Membership Agreement Across Time

A general k_m coefficient that represented all profile types was computed as were partial k_m coefficients for each individual category. Table 26 displays general and partial k_m coefficients for Donders' (1996) WISC-III factor taxonomy using Sample 2, where unusual profiles were identified using the standard error method.

Table 26

General and Partial k_m Coefficients for the WISC-III Factor Taxonomy (Donders, 1996) Using Sample 2 (Unusual Cases Defined by the Standard Error Method)

	k_m coefficient
General k_m	.28*
Partial k_m	
Profile 1	.09
Profile 2	.35
Profile 3	.49
Profile 4	.32
Profile 5	.42
Unusual	.15

* $p < .002$ (with Bonferroni correction adjusting for 24 comparisons; experimentwise error rate = .05).

Results of Analyses to Determine Whether Distribution of Unusual Cases and Degree of Instability Varied Across Geographic Regions, States, and Reporting Psychologists

An analysis was conducted to determine whether unusual cases, a type of outlier, were distributed unevenly across geographic region, states, or individual psychologists. Upon examining the proportion of unusual cases in each region of the country, it became apparent that designation as unusual was not likely a function of differing regional practices. The percentage of students with unusual profiles remained fairly constant across the four regions within both Sample 1 and Sample 2 (both when unusual cases were defined by the critical D^2 method and when they were defined by the standard error

method) at each testing interval. In addition, inspection of profile type membership for individual states and psychologists where a sufficient number of cases were reported did not reveal an obvious correspondence between the number of unusual cases and the source of data.

In order to determine whether instability was represented unproportionally across geographic region, general k_m coefficients for each region were visually compared for Sample 1, Sample 2 where unusual profiles were defined by the critical D^2 method, and Sample 2 where unusual profiles were determined by employing the standard error technique (Table 27).

Table 27

General k_m Coefficients Across Geographic Regions

	West	South	North Central	Northeast
Sample 1	.29	.41	.47	.27
Sample 2[a]	.18	.42	.51	.26
Sample 2[b]	.14	.24	.35	.30

[a]Unusual cases defined by the critical D^2 method; [b]Unusual cases defined by the standard error method.

Consistent with k_m coefficients where all regions were considered together, lower agreement was found when the standard error method of defining profiles as unusual was employed. In this instance agreement was poor (i.e., below .40) across all regions. When the critical D^2 method was employed poor agreement was found in the West and Northeast regions, while fair agreement characterized the South and North Central

regions. Thus, the current study cannot discount membership instability, where unusual cases were identified using the D^2 method, as being reflective of regional practices.

Where an adequate number of cases were available, the degree of instability was informally investigated for correspondence with reporting psychologists or particular states. With the exception of one psychologist whose nine cases all showed membership agreement across time with respect to Konold et al.'s (1999) subtest taxonomy, no obvious connection was noted between degree of instability and reporting psychologist or state.

Discussion

The purpose of the current study was to examine the long-term stability of profile-type membership based on taxonomies of WISC-III subtest (Konold et al., 1999) and factor (Donders, 1996) profiles. Practitioners tend to rely on profile analysis of Wechsler cognitive scores to help make educational decisions (Alfonso et al., 2000; Pfeiffer et al., 2000). Given that reliability is a prerequisite to the valid interpretation and use of test results (AERA, APA, & NCME, 1999), it is imperative that the stability of results based on profile analysis be investigated. That is, instability of results of profile analysis would indicate that their application when making educational choices is both unreliable and invalid. Given that clinically based methods of profile analysis have many limitations (e.g., McDermott et al., 1992) and little empirical support (e.g., Kavale & Forness, 1984; Kline et al., 1993; McDermott et al., 1992; Watkins & Glutting, 2000; Watkins & Worrell, 2000), the long-term stability of profile-type membership based on empirically derived WISC-III taxonomies of both subtest and factor profiles (i.e., nonlinear multivariate profile analysis) was examined.

Results of this study indicated that cluster membership based on nonlinear multivariate profile analysis over a 3-year period was generally not stable: Agreement coefficients across profile types were all poor according to Cicchetti's (1994) guidelines. Specifically, overall k_m coefficients were .39 and .37 for Sample 1 and Sample 2, respectively, using the D^2 approach for classification. The k_m coefficient for Sample 2 when unusual profiles were designated using the standard error technique was .28. This low coefficient of agreement was not unexpected given that the standard error approach described by Donders (1996) does not take all profile dimensions into account and,

instead, focuses only on elevation. Consideration of only one profile dimension is not consistent with a nonlinear multivariate approach to profile analysis.

Interestingly, general k_m values for the WISC-III factor taxonomy were not higher than the k_m value for the WISC-III subtest taxonomy, despite the higher reliability of factor versus subtest scores (Canivez & Watkins, 1998; Wechsler, 1991). Although all three general k_m coefficients were statistically significant, none reached the a priori value of .40, the minimum k_m value that would warrant further research into the potential utility of nonlinear multivariate profile type membership information when making educational decisions. That is, poor classification stability of core profile membership over time indicated that interpretation or use of these membership decisions cannot be valid (AERA, APA, & NCME, 1999).

Of the agreement coefficients that were calculated for each profile type across both samples, only three reached statistical significance. Profiles 6, 7, and 8 of Konold et al.'s (1999) WISC-III subtest taxonomy for the 10 mandatory subtest scores had partial k_m values that likely represented true stability over time. Using Cicchetti's (1994) standards, the k_m coefficient for Profile 7 showed poor clinical significance (k_m = .36), while k_m values for Profile 6 (k_m = .43) and Profile 8 (k_m = .51) were fair in terms of clinical significance. Thus, according to the a priori k_m value set at .40, future validity research for Profile 6 and Profile 8 membership should be conducted.

Profile 6 and Profile 8: Demographics and Patterns of Cognitive Scores

In terms of demographics, participants in Profile 6 and 8 (both at Time 1 and Time 2) were different from Sample 1. Although average age was similar for Sample 1 and Profile 6, through visual inspection it became apparent that Profile 6 contained more

students diagnosed as LD and fewer diagnosed with mental retardation and *other* disabilities compared to Sample 1. Having a smaller number of participants with mental retardation is not a surprising finding given that the average FSIQ for this profile ranged from 87 (Time 2) to 88 (Time 1), while the generally accepted cutoff for diagnosis of mental retardation is 70 (Spruill, 1998). Gender and ethnicity were similar in proportion for Sample 1 and Profile 6, but more children than expected in Profile 6 came from the western part of the country.

Students in Sample 1 and Profile 8 had similar average ages. Compared to Sample 1, Profile 8 contained more students with mental retardation, speech and language disabilities, and *other* disabilities and fewer students with LDs. Having a higher proportion of students with mental retardation is not unexpected given that the average FSIQ of this profile approaches 70. In addition, Profile 8 included more females and not as many males compared to Sample 1 to the extent that, at Time 2, 50% of Profile 8 members were male and 50% were female. This finding is somewhat unexpected given that mild mental retardation (overall IQ = 50 – 70) has been found to be more prevalent among boys (McLaren & Bryson, 1987). On the other hand, the average FSIQ for students in Profile 8 was slightly higher than 70 meaning that Profile 8 does not represent a group of children with mild mental retardation. More Black students and fewer White students were members of Profile 8. This finding is not unanticipated given that the same trend was seen among children aged 6 to 21 from the 50 states, the District of Columbia, and Puerto Rico who received special education services under IDEA-97 during the 2000-2001 school year: A higher proportion of Black children (33.83%) were classified as having mental retardation compared to what was expected given the proportion of

Black children (19.87%) when all disabilities were considered (USDOE, 2001). The opposite pattern was found for White children, where 62.46% of those with all types of disabilities were White, while only 52.19% of those diagnosed with mental retardation were White (USDOE). Finally, Profile 8 contained more students than expected from the South and fewer from the West.

Average WISC-III IQ and subtest scores of students in Profile 6 and 8 (Time 1 and Time 2) resembled those described by Konold et al. (1999) for these profiles. In the present study, Profile 6 students at Time 1 displayed a 12-point split between their average PIQ and VIQ scores (PIQ > VIQ), a significant ($p < .05$), but not rare difference (35.8% of the normative sample have a split this large or larger). The difference was 10 points at Time 2 (PIQ > VIQ): This difference is significant ($p < .15$), but not rare (44.5%). Similarly, Konold et al. found Profile 6 students to have a mean PIQ-VIQ difference of 10.7 points (PIQ > VIQ) and to display a higher proportion of PIQ > VIQ profiles than expected and a lower number of VIQ > PIQ profiles. Profile 8 students in the current study at Time 1 and 2 displayed a relatively flat profile with no significant difference between PIQ and VIQ. This was consistent with Konold et al.'s finding that the number of PIQ > VIQ discrepancies and VIQ > PIQ discrepancies were not different than what would be expected in the WISC-III standardization sample.

Although significant PIQ > VIQ discrepancies were found when comparing the average scores for students in Profile 6 and none were found for Profile 8, in order to be consistent with Konold et al. (1999), profile trends were identified in a slightly different manner. Trends were established when more than 3% of students in the profile of interest displayed a strength or weakness of magnitude found in only 3% of the WISC-III

standardization sample. This standard has been applied when describing the characteristics of core profile taxonomies of various tests of intelligence (e.g., Konold et al.; McDermott, Glutting, Jones, Watkins, & Kush, 1989).

Although larger sample sizes are needed to determine trends with certainty, high PO scores in comparison to both VC and FD scores, and, more specifically, relative subtest strengths in Picture Completion and Object Assembly and relative subtest weaknesses in Information, Arithmetic, and Vocabulary were more than twice and up to almost five times as frequent among students in Profile 6, at both Time 1 and Time 2, as that expected in the general population. Unlike what might be expected from the comparison of Profile 6 to Konold et al.'s (1999) Profile 6, no disproportional number of PIQ > VIQ discrepancies was found in the current study's Profile 6.

In contrast to Profile 6, no shape trends were found among the IQ, index, or subtest scores for students in Profile 8 at either Time 1 or Time 2. This last finding was not surprising given that people with mental retardation generally do not exhibit much variability across subtest scores (Kaufman & Lichtenberger, 2000), and members of Profile 8 had, on average, an FSIQ score approaching what is generally considered the cutoff score for diagnosis of mental retardation. On the other hand, inconsistent findings suggested that children with very depressed IQ scores may display a PIQ > VIQ trend (Spruill, 1998).

Instances where no student in Profile 6 or 8 demonstrated given score patterns may have been indicative of lower than expected occurrences of these trends. Further, for both students in Profile 6 and for those in Profile 8, no unexpected frequency of cases with a large degree of scatter was found within the 10 mandatory WISC-III subtests, nor within

the Verbal, Performance, VC, or PO subtests. Finally, missing data for the Symbol Search subtest precluded the comparison of certain WISC-III scores for students in Profiles 6 and 8.

Directions for Future Research

Classification into Profile 6 or 8 is only meaningful if the trends identified for those in each profile are useful in making predictions for the students or if they can be used to generate effective interventions (Glutting, McDermott, Prifitera, & McGrath, 1994). For example, Reschly (1997) stated that "a context in which intellectual assessment is not related to interventions rarely is in the best interests of clients" (p. 438). According to Kaufman and Lichtenberger (2000) a number of interpretations are possible for students in Profile 6 given the relative strengths and weaknesses commonly found among subtest scores, as well as frequent trends among index scores in conjunction with no abnormal degree of scatter. For example, relative weaknesses in Information, Arithmetic, and Vocabulary may be indicative of deficits in acquired knowledge, long-term memory, or school learning (Kaufman & Lichtenberger). In fact, numerous interpretive possibilities exist, limited only by the examiner's imagination; however, analyses must be derived on a case- by-case basis, taking into account results of additional testing, observations, and other information (Kaufman & Lichtenberger). Through case studies Kaufman and Lichtenberger illustrated how hypotheses generated via WISC-III interpretive guidelines that employ profile analysis are translated into educational recommendations.

Despite the many available suggestions for interpretation, recommendations are clinically based and are not supported empirically. Kaufman and Lichtenberger (2000) wrote that "it is important to note that there is little empirical validation for the diverse

clinical hypotheses suggested by many clinicians." (p. 91). Similarly, Reschly (1997) stated that, with respect to ideas for interpretations based on trends seen among index, factor, and subtest scores, "empirical tests of these interpretations are virtually nonexistent" (p. 444). Further, the scant amount of research that has investigated the utility of IQ, index, and subtest score shape and scatter in diagnosis and hypothesis generation did not yield optimistic findings (Watkins, 2003; Watkins et al., in press), an unsurprising discovery given the many difficulties associated with clinically based profile analysis (e.g., McDermott et al., 1992).

It is difficult to know whether the trends uncovered for the higher than expected number of children in Profile 6 are of any practical value. Although Watkins (2003) outlined research dismissing the potential utility of the FD scores in detecting the presence of attention deficit hyperactivity disorder, searches on both ERIC and PsychINFO revealed very little additional useful empirical findings related to the patterns detected among Profile 6 students. Most commonly, the search terms used were *WISC-III* and the name of the subtest or index of interest (e.g., Picture Completion). Thus, there may be merit in directing future research toward the uncovering of any predictive meaning or treatment utility of Profile 6 trends. In addition, future research might focus on the meaning of PO > VC differences, PO > FD discrepancies, or the various subtest strengths and weaknesses detected among members of Profile 6, but for children with below average ability, similar to that of students belonging to Profile 6.

On the other hand, the premise of the current paper's investigation of an empirical approach to profile analysis (i.e., nonlinear multivariate profile analysis) was due to the many limitations of clinically based profile analysis methods in the interpretation of

WISC-III scores, such as those outlined by Kaufman and Lichtenberger (2000). For example, clinically based profile analysis relies on ipsative scores, which have been found to be problematic (McDermott et al., 1992) and disregards the literature on practitioner judgment (e.g., Faust, 1986) and on ATI (Gresham & Witt, 1997; Reschly, 1997). Future research, then, will be better directed by exploration of the meaning of membership in Profile 6 or 8, rather than the meaning of isolated components of those profiles. That is, membership in Profile 6 and 8 can be investigated for correspondence with certain outcome variables, such as an aspect of classroom behavior or response to a specific mode of instruction. Further, outcomes of interest can begin to be considered from a multivariate perspective, perhaps in conjunction with a multivariate view of intelligence.

Noting that the commonly employed, univariate model of detecting IQ-achievement discrepancies ignores the multifaceted nature of intelligence as well as prevalence distinctions between those with a discrepancy in a given achievement area, a discrepancy in any achievement domain, or multiple discrepancies, Glutting et al. (1994) proposed a multivariate approach that can account for both linear and nonlinear aspects of IQ and achievement profiles. Using the MEG clustering procedures, a taxonomy was derived based on the scores of 824 students from the WISC-III and Wechsler Individual Achievement Test (WIAT; Wechsler, 1992) linking sample. Scores subjected to analysis included the four WISC-III index scores and four composite WIAT scores: Reading, Mathematics, Language, and Writing. The resulting taxonomy had 6 core profiles primarily distinguishable by level. However, only one profile type was flat and findings revealed that univariate IQ-achievement discrepancies were fairly common. Identification

of unusual profiles constitutes a multivariate approach to IQ-achievement discrepancy detection.

Thus, future validity research can be conducted to determine whether students in Profile 6 or 8 demonstrate multivariate IQ-achievement discrepancies with disproportional frequency using the taxonomy derived and guidelines outlined by Glutting et al. (1994). Should this be the case, research must then focus on whether the existence of a multivariate IQ-achievement discrepancy is predictive of a given outcome or indicative of an effective intervention (Glutting et al., 1994). Preliminary work in this area has been conducted: Ward, Ward, Glutting, and Hatt (1999) identified two subcategories of children who displayed multivariate IQ-achievement discrepancies; characteristics of each group were described. However, participants included only children diagnosed with a LD, presumably based on a univariate discrepancy model, while almost half of those found by Glutting et al. (1994) to have a multivariate discrepancy (IQ > achievement) did not exhibit a univariate discrepancy. Further, Ward et al. noted that their results do not constitute research on the potential predictive ability or treatment utility of multivariate discrepancies.

When conducting the suggested future research, it will be important to design experiments that are methodologically sound. That is, the fundamental errors that plague much of the research on clinical profile analysis (Glutting et al., 1998; Watkins et al., in press) must be avoided. For example, to abstain from circular reasoning, heterogeneous samples of children should be selected and when problems later surface, researchers should then identify possible predictive factors, such as membership to Profile 6 or 8 or multivariate IQ-achievement discrepancies (Glutting et al., 1994).

Finally, it is important to keep in mind that in order to be meaningful, profile membership information must combine with overall intelligence scores to produce results that exceed what is currently predictable through knowledge of global IQ scores alone. That is, knowledge of profile membership must add incremental validity relative to global intelligence scores (Lubinski, 2004). Research has established that the overall score on Wechsler scales of intelligence can be used as a fairly good predictor (r = .4 to .7) of both academic and occupational accomplishment (Reschly, 1997). Also, global IQ is related to a number of other factors, such as likelihood of being out of compliance with the law (Reschly). So, for example, future research findings that Profile 8 membership predicts poor classroom performance would not be meaningful because, as Reschly stated in his discussion on Wechsler scales as measures of general intelligence, "the Wechsler scales are valid as measures of diagnostic constructs involving learning ability and likely performance in an academic setting" (p. 444-445). Thus, students belonging to Profile 8 will likely have difficulty in the regular educational environment by virtue of their depressed overall intelligence scores, rather than as a function of overall intelligence combined with the level, scatter, and shape trends that simultaneously characterize membership to this profile.

Limitations and Additional Directions for Future Research

Interpretation of the results of this study must be made within the context of its limitations. Participants in this study represented a subset of the sample obtained by Canivez and Watkins (1998). Although Canivez and Watkins (1998) sent their survey to 2,000 randomly selected school psychologists from among members of the National Association of School Psychologists, there were only 114 respondents. This 5.70%

response rate is low and removed the randomness with which participants were originally selected. That is, participation was based on the voluntary decision of school psychologists to respond. Further, respondents selected which students' scores were reported, and it was ultimately up to practitioners whether a student was administered all 12 WISC-III subtests necessary for factor score calculation. Given that the sample in the current study represented a subset of the sample obtained by Canivez and Watkins (1998), participants of the present study were also not randomly selected. Non-randomness of a sample reduces generalizability of results. On the other hand, there was no reason to suspect a selection bias among over 100 respondents from 33 states.

A few other factors also limit generalizability of the current findings. Only scores of students who were re-evaluated could be selected as participants. As such, results should not be extended to students who were only evaluated once, such as those no longer requiring special education services. In addition, results are most representative of students in Grades 1 through 5, as they represented the majority of both Sample 1 (83.42%) and Sample 2 (92.09%). Although Sample 1 and 2 were representative of the population of students receiving special education (USDOE, 2001), it is difficult to generalize results to children who are not White as well as to students without disabilities and to those having a disability other than an LD. Most of the participants in Sample 1 and Sample 2 had a disability; further, the majority of students in both samples had an LD. In addition, the sample sizes for many disability categories other than LD were small. Some disability groups were not included at all in one or both of the present analyses; for example, neither Sample 1 nor Sample 2 included children with visual impairments. Certain subtest scores for these students were not available due to

administration difficulties resulting from a mismatch between children's capabilities and demands of the WISC-III. Additional research would be needed to make conclusions regarding the profile-type membership stability of children without disabilities or of other groups not adequately represented in the current study.

Cognitive profiles analyzed in the current study contained subtest and factor scores from the WISC-III. Therefore, further research is needed before results can be generalized to students' profiles on other measures of cognitive functioning. For example, the WISC-III has been replaced by the WISC-IV and it is important that the current study be replicated using this updated measure and that suggestions for future research based on the results of this study be modified for similar studies involving the WISC-IV.

Another difficulty of this study was related to examiner effects. For example, the accuracy of WISC-III administration to students at both Time 1 and at Time 2 could only be assumed. On the other hand, the relatively even distribution across reporting psychologists of both unusual cases and instances of instability, as well as the case of one psychologist's 100% stability rate across reported cases, were not suggestive of unstandardized WISC-III administration and scoring. The presence of multiple raters is another way in which examiner effects may have influenced results. That is, because the assessors of a given student at Time 1 and Time 2 may have varied, different degrees of standardized administration on the part of the examiners might have influenced WISC-III results and distorted classification stability findings. In addition, examiner familiarity effects may have had an impact on findings (Fuchs & Fuchs, 1986), a concern that can be addressed in future research.

Also, although the proportion of unusual cases did not vary appreciably with geographic region or state, nor did cases of instability correspond to particular states, the potential influence of geographic region on results cannot be disregarded due to higher agreement coefficients in the South and North Central regions compared to the West and Northeast when unusual cases were identified using the D^2 method. That is, low classification stability may have resulted in part from test administration or scoring practices that differed across geographic regions, resulting in disproportional instances of instability.

Interpretation of classification agreement coefficients was rendered difficult given that some clusters were small in size of membership. That is, lack of statistical significance may have resulted from this small sample size. Future research may take this into account by enlisting a larger number of participants across profile types.

A final consideration relates to the possibility that poor classification stability may be related to practice effects or to real change in intelligence over the 3-year period. For instance, Wechsler (1991) found a practice effect with a WISC-III test-retest interval ranging from 12 to 63 days: An increase of about 7 or 8 FSIQ points were noted over this short retest interval. However, there is reason to believe that in the current study participants' intelligence scores remained stable over the retest interval. Using WISC-III data from almost all of the children in Sample 1 (n = 579), Watkins and Canivez (2004) found that no IQ or factor score was statistically significantly different ($p < .05$) over the 2.8 year period, with the largest change being 2.4 points (PS index score). Only three subtest scores were found to differ significantly ($p < .05$) from test to retest, but the biggest change in magnitude (.7 points) was not considered to be of practical importance.

Conclusion

Although it appears that profile-type membership possesses some degree of stability in the short-term across a number of cognitive measures (Glutting & McDermott, 1990a, 1990b; Glutting et al., 1992; McDermott, Glutting, Jones, Watkins, & Kush, 1989), results of the current study revealed that profile-type membership of empirically derived subtest and factor WISC-III profiles did not remain stable in the long-term (i.e., 3 years). That is, having a particular WISC-III subtest or factor profile at Time 1 did not predict designation of the same profile at Time 2. As such, empirically-based WISC-III subtest and factor profile-type membership cannot be relied upon to make educational decisions for students. Even though a nonlinear multivariate approach to profile analysis has advantages over clinically based techniques, to date neither approach has been supported in its contribution to diagnosis or educational decision-making.

Two exceptions were Profiles 6 and 8 from Konold et al.'s (1999) WISC-III subtest taxonomy for the 10 mandatory subtest scores, which had fair agreement coefficients indicating that future research is warranted. Thus, the suggested future validity research as well as replication research taking stated limitations into account will be important for ultimate interpretation of current findings. Finally, it must be kept in mind that results of the current study do not extend to global scores on Wechsler tests that have been shown to be stable over several years (e.g., Canivez & Watkins, 1998) as well as predictive of certain important characteristics, such as achievement and occupational success (Reschly, 1997).

References

Aldenderfer, M. S., & Blashfield, R. K. (1984). *Cluster analysis*. Beverly Hills, CA: Sage.

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

Alfonso, V. C., & Pratt, S. I. (1997). Issues and suggestions for training professionals in assessing intelligence. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 326-347). New York: The Guilford Press.

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

Bannatyne, A. (1968). Diagnosing learning disabilities and writing remedial prescriptions. *Journal of Learning Disabilities, 1,* 242-249.

Belter, R. W., & Piotrowski, C. (2001). Current status of doctoral-level training in psychological testing. *Journal of Clinical Psychology, 57,* 717-726.

Borgen, F. H., & Barnett, D. C. (1987). Applying cluster analysis in counseling psychology research. *Journal of Counseling Psychology, 34,* 456-468.

Burt, C. L. (1937). Correlations between persons. *British Journal of Psychology, 28,* 59-96.

Canivez, G. L., & Watkins, M. W. (1998). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition. *Psychological Assessment, 10,* 285-

291.

Canivez, G. L., & Watkins, M. W. (1999). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition among demographic subgroups: Gender, race/ethnicity, and age. *Journal of Psychoeducational Assessment, 17,* 300-313.

Canivez, G. L., & Watkins, M. W. (2001). Long-term stability of the Wechsler Intelligence Scale for Children-Third Edition among students with disabilities. *School Psychology Review, 30,* 438-453.

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

Cicchetti, D. V. (1994). Guidelines, criteria, and rules of thumb for evaluating normed and standardized assessment instruments in psychology. *Psychological Assessment, 6,* 284-290.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37-46.

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

Davidow, J., & Levinson, E. M. (1993). Heuristic principles and cognitive bias in decision making: Implications for assessment in school psychology. *Psychology in the Schools, 30,* 351-361.

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

Elliott, C. D. (1990). *Differential Ability Scales: Introductory and technical handbook.*

San Antonio, TX: Psychological Corporation.

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

Flanagan, D. P., Andrews, T. J., & Genshaft, J. L. (1997). The functional utility of intelligence tests with special education populations. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 457-483). New York: The Guilford Press.

Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin, 76,* 378-382.

Fleiss, J. L., Levin, B., & Paik, M. C. (2003). *Statistical methods for rates and proportions* (3rd ed.). Hoboken, NJ: John Wiley & Sons.

Forness, S. R. (2001). Special education and related services: What have we learned from meta-analysis? *Exceptionality, 9*, 185-197.

Friend, M., & Bursuck, W. D. (2002). *Including students with special needs: A practical guide for classroom teachers* (3rd ed.). Boston: Allyn and Bacon.

Fuchs, D., & Fuchs, L.S. (1986). Test procedure bias: A meta-analysis of examiner familiarity effects. *Review of Educational Research, 56,* 243-262.

Gall, M. D., Borg, W. R., & Gall, J. P. (1996). *Educational research: An introduction* (6th ed.). White Plains, NY: Longman.

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

Glutting, J. J., & McDermott, P. A. (1990b). Score structures and applications of core profile types in the McCarthy Scales standardization sample. *Journal of Special*

Education, 24, 212-233.

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

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

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

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

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

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

Gresham, F. M., & Witt, J. C. (1997). Utility of intelligence tests for treatment planning, classification, and placement decisions: Recent empirical findings and future

directions. *School Psychology Quarterly, 12,* 249-267.

Hair, J. F., Jr., Anderson, R. E., Tatham, R. L., & Black, W. C. (1998). *Multivariate data analysis* (5th ed.). Upper Saddle River, NJ: Prentice Hall.

Holland, A. M., & McDermott, P. A. (1996). Discovering core profile types in the school-age standardization sample of the Differential Ability Scales. *Journal of Psychoeducational Assessment, 14,* 131-146.

Individuals with Disabilities Education Act Amendments of 1997, 20 U.S.C. §1400 *et seq.* (EDLAW, 1997).

Jastak, S., & Wilkinson, G. S. (1984). *Wide Range Achievement Test-Revised*. Wilmington, DE: Jastak Associates.

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

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

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

Kaufman, A. S., & Kaufman, N. L. (1983a). *K-ABC administration and scoring manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Kaufman, N. L. (1983b). *K-ABC interpretive manual.* Circle Pines, MN: American Guidance Service.

Kaufman, A. S., & Lichtenberger, E. O. (2000). *Essentials of WISC-III and WPPSI-R assessment.* New York: John 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 Disability Quarterly, 7,* 136-156.

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

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

Landis, J. R., & Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics, 33,* 159-174.

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

Lorr, M. (1983). *Cluster analysis for social scientists: Techniques for analyzing and simplifying complex blocks of data.* San Francisco: Jossey-Bass.

Lubinski, D. (2004). Introduction to the special section on cognitive abilities: 100 years after Spearman's (1904) "'general intelligence,' objectively determined and measured." *Journal of Personality and Social Psychology, 86,* 96-111.

McDermott, P. A. (1998). MEG: Megacluster analytic strategy for multistage hierarchical grouping with relocations and replications. *Educational and Psychological*

Measurement, 58, 677-686.

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

McDermott, P. A., Fantuzzo, J. W., Glutting, J. J., Watkins, M. W., & Baggaley, A. R. (1992). Illusions of meaning in the ipsative assessment of children's ability. *The 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: A Journal of Consulting and Clinical Psychology, 1,* 118-125.

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

McFall, R. M., & Treat, T. A. (1999). Quantifying the information value of clinical assessments with signal detection theory. *Annual Review of Psychology, 50,* 215-241.

McLaren, J., & Bryson, S. E. (1987). Review of recent epidemiological studies of mental retardation: Prevalence, associated disorders, and etiology. *American Journal of Mental Retardation, 92,* 243-254.

McLean, J. E., Reynolds, C. R., & Kaufman, A. S. (1990). WAIS-R subtest scatter using the profile variability index. *Psychological Assessment, 2,* 289-292.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. *Psychometrika, 50,* 159-179.

Milligan, G. W., & Hirtle, S. C. (2003). Clustering and classification methods. In J. A. Schinka & W. F. Velicer (Eds.), *Handbook of psychology: Vol. 2. Research methods in psychology* (pp. 165-186). Hoboken, NJ: Wiley.

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

Mojena, R. (1977). Hierarchical grouping methods and stopping rules: An evaluation. *Computer Journal, 20,* 359-363.

National Institute of Child Health and Human Development. (2000). *Report of the National Reading Panel. Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). Washington, DC: U.S. Government Printing Office.

Nunally, J. (1978). *Psychometric theory* (2nd ed.). New York McGraw-Hill.

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

Otis, A. S., & Lennon, R. T. (1989). *Otis-Lennon School Ability Test, Form I (6th ed.).* San Antonio, TX: The Psychological Corporation.

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

Plake, B. S., Reynolds, C. R., & Gutkin, T. B. (1981). A technique for the comparison of the profile variability between independent groups. *Journal of Clinical Psychology*,

37, 142–146.

Popper, K. R. (1959). *The logic of scientific discovery.* London: Hutchinson.

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

Reynolds, C. R. (1997). Measurement and statistical problems in neuropsychological assessment of children. In C. R. Reynolds & E. Fletcher-Janzen (Eds.), *Handbook of clinical child neuropsychology* (2nd ed., pp. 180-203). New York: Plenum Press.

Ridling, Z. (1994, April). *The effects of three seating arrangements on teachers' use of selective interactive verbal behaviors*. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, LA.

Rieth, H. J., & Semmel, M. I. (1991). Use of computer-assisted instruction in the regular classroom. In G. Stoner, M. R. Shinn, & H. M. Walker (Eds.), *Interventions for achievement and behavior problems* (pp. 215-239). Silver Springs, MD: National Association of School Psychologists.

Salvia, J., & Ysseldyke, J. E. (2001). *Assessment* (8th ed.). Boston: Houghton Mifflin Company.

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

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

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

Speece, D. L. (1994-95). Cluster analysis in perspective. *Exceptionality, 5,* 31-44.

Spengler, P. M., Strohmer, D. C., Dixon, D. N., & Shivy, V. A. (1995). A scientist-practitioner model of psychological assessment: Implications for training, practice, and research. *The Counseling Psychologist, 23,* 506-534.

Spruill, J. (1998). Assessment of mental retardation with the WISC-III. In A. Prifitera & D. Saklofske (Eds.), *WISC-III clinical use and interpretation: Scientist-practitioner perspectives* (pp. 73-90). San Diego, CA: Academic Press.

Tatsuoka, M. M., & Lohnes, P. R. (1988). *Multivariate analysis: Techniques for educational and psychological reserach* (2nd ed.). New York: Macmillan.

U.S. Department of Education. (2001). *To assure the free appropriate public education of all children with disabilities: Twenty-fourth annual report to Congress on the implementation of the Individuals with Disabilities Education Act* (USDOE Contract No. HS97020001 with Westat). Washington, D.C.: Author.

Ward, J. H., Jr. (1963). Hierarchical grouping to optimize an objective function. *American Statistical Association Journal, 58,* 236-244.

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

Watkins, C. E., Jr., Campbell, V. L., Nieberding, R., & Hallmark, R.(1995). Contemporary practice of psychological assessment by clinical psychologists. *Professional Psychology: Research and Practice, 26,* 54-60.

Watkins, M. W. (1998). MacKappa [Computer software]. Pennsylvania State University: Author.

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

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

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

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

Watkins, M. W., Glutting, J. J., & Youngstrom, E. A. (in press). Issues in subtest profile analysis. In D. P. Flanagan, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed.). New York: The Guilford Press.

Watkins, M. W., & Kush, J. C. (1994). Wechsler subtest analysis: The right way, the wrong way, or no way? *School Psychology Review*, *23*, 640-652.

Watkins, M. W., & Worrell, F. C. (2000). Diagnostic utility of the number of WISC-III subtests deviating from mean performance among students with learning disabilities. *Psychology in the Schools, 37,* 303-309.

Wechsler, D. (1944). *The measurement of adult intelligence* (3rd ed.). Baltimore: Williams & Wilkins.

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

Wechsler, D. (1967). *Manual for the Wechsler Preschool and Primary Scale of Intelligence.* New York: The Psychological Corporation.

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

Wechsler, D. (1981). *Manual for Wechsler Adult Intelligence Scale-Revised.* New York: The Psychological Corporation.

Wechsler, D. (1989). *Manual for the Wechsler Preschool and Primary Scale of Intelligence-Revised.* San Antonio, TX: The Psychological Corporation.

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

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

Wechsler, D. (1997). *Manual for the Wechsler Adult Intelligence Scale-Third Edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003a). *Wechsler Intelligence Scale for Children-Fourth Edition administration and scoring manual.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003b). *Wechsler Intelligence Scale for Children-Fourth Edition technical and interpretive manual.* San Antonio, TX: The Psychological Corporation.

Vita

ELLEN BORSUK
61 Windermere, D.D.O., Québec, H9A 2C5, Canada, erb155@psu.edu

Education:

- The Pennsylvania State University (PSU), University Park, PA
  Ph.D. School Psychology, anticipated Summer 2005
  M. S. School Psychology, Spring 2003
  Cumulative GPA: 3.99
- McGill University, Montreal, QC, Canada
  B. S. Physical Therapy, Spring 2000, Cumulative GPA: 3.81

Awards and Affiliations:

- Received a passing score on The Praxis Series' School Psychologist test
- Member of the National Association of School Psychologists and the Association of School Psychologists of Pennsylvania
- Packard Professional Development Endowment for Students, PSU, Spring 2004
- Conrad Frank, Jr. Graduate Fellowship, PSU, 2004-2005
- Susan Beth Robson Scholarship in Education, PSU, 2001-2002
- School Psychology Graduate Award, PSU, 2001-2002
- J. W. McConnell entrance scholarship, McGill University, 1997-2000

Relevant Experience:

- Internship in School Psychology, Ossining Union Free School District, 2004-Present
- CEDAR School Psychology Clinic Student Supervisor, PSU, 2003-2004
- CEDAR School Psychology Clinic Student Clinician, PSU, 2001-2003
- School Psychology Practicum Student, Bellefonte Elementary School, Spring 2002 and Fall 2001
- Research Assistant, Mifflin County School District, Summer 2002; McGill University, Spring-Summer 2000

Other Work Experience:

- Test Librarian, PSU, Summer 2001-Summer 2003
- Teaching Assistant, PSU, 2001-2004 (intermittent; across 6 courses in 3 departments)

Activities:

- Food, Household Goods, and Clothing Bank Volunteer, Le Mercaz, Summer 2003
- Poster Presenter, Association of School Psychologists of Pennsylvania Conference, Spring 2003
- Conference Volunteer, Fifteenth Annual International Precision Teaching Conference, Fall 2002
- Member of Good Schools Pennsylvania, PSU, Spring 2002
- Professional Organization Representative, PSU, Fall 2001-Spring 2003

Languages Spoken:

- English
- French