UNIVERSITY OF ALBERTA

DISCRIMINANT VALIDITY OF THE STANFORD-BINET INTELLIGENCE SCALE: FOURTH EDITION

BY

CHRISTOPHER WILSON

A thesis submitted to the Faculty of Graduate Studies and Research in partial fulfillment of the requirements for the degree of DOCTOR OF PHILOSOPHY.

IN

SCHOOL PSYCHOLOGY

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

EDMONTON, ALBERTA

SPRING 1995

ISBN 0-612-01778-8

Canada

UNIVERSITY OF ALBERTA

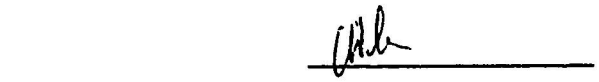
RELEASE FORM

NAME OF AUTHOR: Christopher Wilson

TITLE OF THESIS: Discriminant Validity of the Stanford-Binet Intelligence Scale: Fourth Edition

DEGREE: Doctor of Philosophy

YEAR THIS DEGREE GRANTED: 1995



Christopher Wilson

40 Shaneragh Road
Mandara
Harare
ZIMBABWE

Date: December 6, 1994

UNIVERSITY OF ALBERTA

FACULTY OF GRADUATE STUDIES AND RESEARCH

The undersigned certify that they have read, and recommend to the Faculty of Graduate Studies and Research for acceptance, a thesis entitled DISCRIMINANT VALIDITY OF THE STANFORD-BINET INTELLIGENCE SCALE: FOURTH EDITION submitted by CHRISTOPHER WILSON in partial fulfillment of the requirements for the degree DOCTOR OF PHILOSOPHY in SCHOOL PSYCHOLOGY.

H. L. Janzen (Supervisor)

J. G. Paterson

W. T. Rogers

L. L. Stewin

R. H. Short

K. L. Ward

G. W. Hynd

DATE: December 5, 1994

ABSTRACT

Performance on the Stanford-Binet Intelligence Scale: Fourth Edition (SB:FE) was examined in a Canadian clinic sample (N=1220), age 2 through 23 years, with a range of demographic characteristics and ability levels. Data were analyzed for the samples 2-6-11, 7-11-11, and 12-23-11. SB:FE subtest, Reasoning Area, and Composite Standard Age Scores (SAS's) decreased significantly ($p < .05$) with increasing age. Within each age group, the intercorrelations among subtests, the four Reasoning Area, and the Composite SAS's supported the four cognitive ability areas posited by Thorndike et al. (1986b). Performance of subjects on the SB:FE full battery and SB:FE General Purpose Abbreviated Battery (GPAB) were compared. Significant differences ($p < .05$), attributable to the large sample sizes, were found between means and variances in Reasoning Area and Composite SAS's. Uncorrected correlation coefficients among the two measures were significant ($p < .01$) and close to unity for the Verbal, Quantitative, Short Term Memory, and Composite SAS's. The correlations between Abstract/Visual SAS's, while significant, were somewhat lower. Also, similar and significant ($p < .05$) correlations were observed among the two versions of the SB:FE and the Wide Range Achievement Test-Revised (WRAT-R). Next, internally valid, reliable, and replicable groups displaying differences in profile elevation and/or shape were obtained through application of hierarchical agglomerative and iterative partitioning clustering procedures to SB:FE GPAB data. For the age sample 2-4-11, a two cluster solution, with high average and average groups was optimal. For the samples 5-6-11, 7-11-11, and 12-23-11, a three cluster solution comprising high, average, and low scoring groups was optimal. Mean WRAT-R subtest scores of the groups in all ages samples were significantly different ($p < .01$). However, when cluster solutions were compared with clinically derived a priori learning disability models, clusters were more similar with respect to Composite SAS's or profile elevation, than educational diagnosis. In general, results suggest the SB:FE is most appropriately used as an index of global ability. Caution is needed interpreting Reasoning Area SAS's, although the GPAB may provide a reasonable representation of the full battery.

ACKNOWLEDGEMENTS

It is difficult to adequately express my gratitude to the many people who have contributed, not only this degree, but also to my growth and development over the years since I arrived at the University of Alberta. However, I will try.

First, I would like to express my sincere appreciation and gratitude to my committee: To Dr. Paterson, who, although becoming involved only in the later stages of the dissertation, so sagely contributed to the final examination; To Dr. L. Stewin, who has been involved in my academic progress from the day I arrived in Canada. Thank you for your support, encouragement, and belief in me over the years; To Dr. R. Short, who has also been involved in my academic program since I arrived in Alberta. Many thanks for your assistance over the years; To Dr. K. Ward, thank you for your involvement, insight, and constructive feedback; To Dr. G. W. Hynd many thanks for your invaluable assistance; To Dr. H. Janzen, my supervisor, thank you for your unwavering support and encouragement over the years; and finally sincere thanks to Dr. T. Rogers for your tremendous commitment and support in ensuring this document reached completion.

On a more personal note, thank you to the many people who have supported and encouraged me over the years while I have been in graduate school: In particular, thanks go to Dr. Rod Beattie, whose assistance and support, not only to myself, but to many graduate students over the years, has been immeasurable. To my mother Cicely, and brother David, in Zimbabwe, for their support and encouragement. Thank you. To my Canadian family, Hans, Christine, George, and Chris. Thank you for helping make Canada my home.

Last, and most important of all, I would like to thank Ellen, my wife, for her patience, understanding, support, and love throughout.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE: INTRODUCTION

In this chapter, the Stanford-Binet Fourth Edition (SB:FE) (Thorndike, Hagen, & Sattler, 1986a) is briefly introduced. The roles and purposes of educational assessment are discussed, and the need to establish the validity of the SB:FE in educational assessment is identified. In discussing educational assessment, particular emphasis was placed on investigations of learning disabilities and subtyping or classification research. Multivariate research with the SB:FE poses particular problems because of the way it was designed. Thus, as described in this chapter, a major purpose of the study was to explore the utility of the Stanford-Binet: Fourth Edition General Purpose Abbreviated Battery (SB:FE GPAB) as a means of helping to overcome these difficulties. Lastly, this chapter concludes with a presentation of the major research objectives and an outline of the organization of the dissertation.

The Stanford-Binet Intelligence Scale: Fourth Edition (SB:FE)

The SB:FE is "a major intelligence test" that provides a "continuous scale for appraising cognitive development from age 2 to adulthood" (Thorndike et al., 1986a, p. 8). A three level hierarchical model guided the construction of the SB:FE (Sattler, 1988, 1992). This model postulates a general intelligence factor, g, at the highest level of interpretation, and Crystallized, Fluid, and Short Term Memory factors at the second level. As shown in Figure 1, the three factors included at the third level are "nested" within the factors at the second level as follows: Verbal Reasoning and Quantitative Reasoning reflect Crystallized Abilities; Abstract/Visual Reasoning reflects Fluid Ability; and Short Term Memory which stands independently and does not subsume other factors. Each of Verbal Reasoning, Quantitative Reasoning, Abstract/Visual Reasoning, and Short Term Memory are then measured by specific subtests unique to each. Four subtests comprise the Verbal Reasoning area - - Vocabulary, Comprehension, Absurdities, and Verbal Relations; three subtests the Quantitative area - - Quantitative, Number Series, and Equation Building; four subtests the Abstract/Visual area - - Pattern Analysis, Copying, Matrices, and Paper Folding and Cutting; and four the Short Term Memory area - - Bead Memory, Memory for Sentences, Memory for Digits, and Memory for Objects. Up to 13 subtests of the SB:FE may be

administered, depending on subject age, ability, and examiner choice (Keith, Cool, Novak, White, & Pottebaum, 1988a; Sattler, 1988). Despite differences in the number of subtests administered at the different age and ability levels, the same grouping of the subtests into ability domains is assumed (Molfese, Yaple, Helwig, Harris, & Connell, 1992).

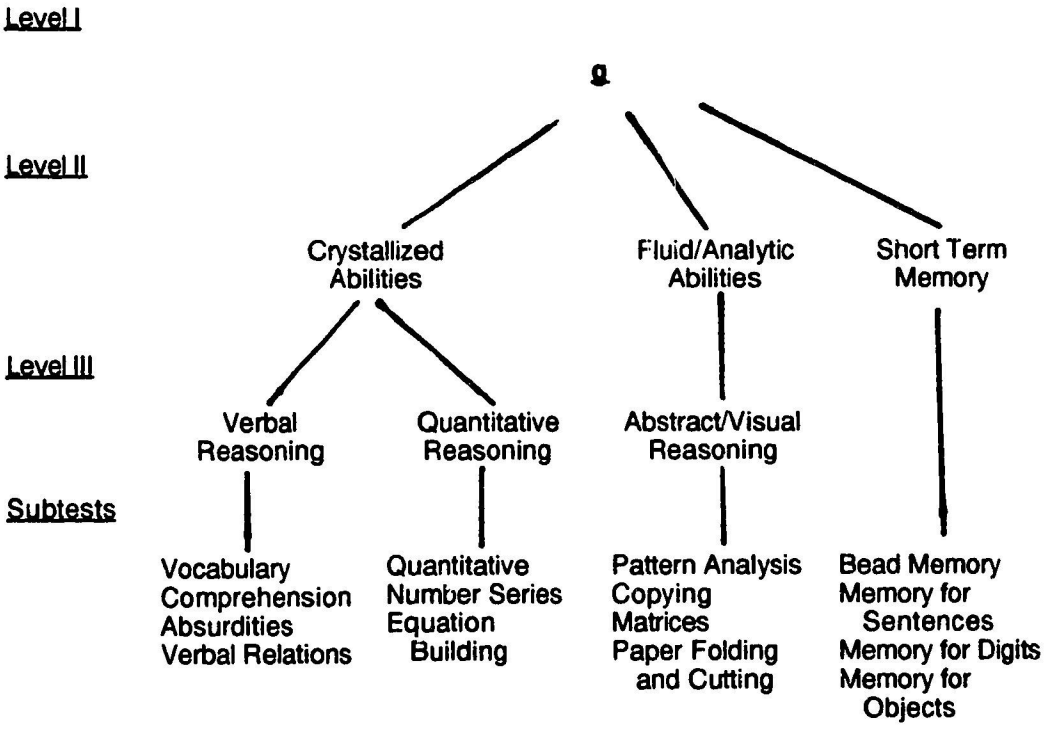


Figure 1. Theoretical model of intelligence underlying the SB:FE.

Thorndike et al. (1986a) noted that the hierarchical model of the SB:FE was adopted largely because of the way clinicians and educators have used the previous editions of the Binet - - "together with other information to make recommendations for educational intervention" (p. 9). These uses have included utilization of the Binet "to identify gifted students, to assess the cognitive abilities of mainstream students who were having difficulty learning, and to identify the mentally retarded" (Thorndike et al., 1986a, p. 9). These potential uses are also applicable to the revised SB:FE, given that the revision was intended "to assess the kinds of cognitive abilities years of research have shown are correlated with school progress" (Thorndike et al., 1986a, p. 9). The 15 subtests comprising the SB:FE, and the organization of these subtests into the four

reasoning areas, make it possible to interpret profile elevations and depressions (Glutting, 1989). Thus, in light of the rationale underlying the construction of the SB:FE and the range of cognitive abilities tapped by the test, the instrument appears to hold promise as a means of providing diagnostic and remedial information in educational settings. Boyle (1989) suggested "the new instrument may well usher in an exciting era for cognitive measurement" with research and applied findings "pertaining to clinical, clinical neuropsychological, vocational and educational domains respectively" (p. 709).

There are abbreviated versions of the SB:FE version available that provide "a reasonably accurate estimate of overall cognitive level and pattern of cognitive abilities" (Thorndike et al., 1986a, p. 35). The four test Quick Screening Battery comprises four subtests administered at all age levels: Vocabulary, Bead Memory, Quantitative, and Pattern Analysis. The six test SB:FE General Purpose Abbreviated Battery (GPAB) includes all six subtests that are administered at all ages: Vocabulary, Bead Memory, Quantitative, Memory For Sentences, Pattern Analysis, and Comprehension. Both abbreviated versions require substantially less testing time than the complete battery and have acceptable internal consistency reliabilities (Thorndike et al., 1986b). For example, the SB:FE GPAB can be administered in about 60 minutes (Carvajal, McVey, Sellers, Weyand, & McKnab, 1987), whereas Sattler (1982) reported that the full battery is much too long to complete in most circumstances. The SB:FE GPAB can be used for placement decisions (Glutting, 1989; Thorndike et al., 1986a).

The Role of Assessment in Educational Diagnosis and Planning

There is an increasing emphasis in education on the use of individual educational plans (IEP's), especially for those students experiencing learning difficulties (Salvia & Ysseldyke, 1988; Sattler, 1988). The use of individualized intelligence tests is integral to the diagnosis of learning difficulties and the formulation of individual educational plans. Indeed, Mueller, Dennis, and Short (1986) suggest that the popularity of the Wechsler Intelligence Scale for Children-Revised (WISC-R) (Wechsler, 1974), for example, arises from its apparent attractiveness "for psychoeducational diagnostic purposes" (p. 22). Diagnosis provides one of the first steps in planning educational

programs. Again, with reference to WISC scales, Kavale and Forness (1984) observed "the structure of the WISC leads to the assumption that there ought to be subtest patterns; in addition, established clinical practice operates as if such patterns were fact" (p. 150). Tests that merely yield single IQ scores are not useful for diagnosis and educational planning. Whilst the previous version of the Binet, the Stanford-Binet Form L-M, was one such test, in contrast, the revised structural design of the SB:FE, with its four cognitive ability or reasoning areas, represents a considerable departure from earlier editions (Keith et al., 1988a), and is "one that is better suited to educational planning" (Fritzke, 1988, p. 50). However, the validity of the SB:FE in educational and differential diagnosis needs to be established.

In the Standards for Educational and Psychological Testing (American Psychological Association, 1985) it is stated:

> Validity is the most important consideration in test evaluation. The concept refers to the appropriateness, meaningfulness, and usefulness of the specific inferences made from test scores. Test validation is the process of accumulating evidence to support such inferences. A variety of inferences may be made from scores produced by a given test, and therefore there are many ways of accumulating evidence to support any particular inference. Validity, however, is a unitary concept. Although evidence may be accumulated in many ways, validity always refers to the degree to which the evidence supports the inferences that are made from the scores. The inferences regarding specific uses of a test are validated, not the test itself. (APA, 1985, p. 9)

Standards 1.1. and 1.3 (APA, 1985) are particularly relevant for this dissertation. Standard 1.1 states: "Evidence of validity should be presented for the major types of inferences for which use of the test is recommended" (APA, 1985, p. 13). Standard 1.3 states: "Whenever interpretation of subscores, score differences, or profiles is suggested, the evidence justifying such interpretation should be made explicit" (APA, 1985, p. 13). The need to establish the validity of the SB:FE within these parameters in the context of educational diagnosis is paramount.

The Guidelines for Educational and Psychological Testing (Canadian Psychological Association) (CPA, 1987) were formulated to be generally consistent with the APA standards. However, in constructing the guidelines, allowances were made for differing legal and social facets. The guidelines are grounded within the Canadian context (CPA, 1987) and provide additional support for the need to establish evidence of the validity of the SB:FE in educational

diagnosis and decision making, particularly within the context of Canadian samples. In the Principles for Fair Student Assessment Practices for Education in Canada (1993) further support for research endeavours of this nature can be found. The second part of this code applies to standardized assessment measures used in student admissions, placement, certification, and educational diagnosis. "Users should select methods that are appropriate for the intended purposes and suitable for the students to be assessed" (Principles for Fair Student Assessment Practices for Education in Canada, 1993, p. 15). One of the purposes of this dissertation is to explore the suitability and appropriateness of the SB:FE in educational assessment and diagnosis. These are the first steps in planning remediation interventions.

Perhaps one of the most important and contentious areas within education today is that of learning disabilities (Hooper & Willis, 1989). Individuals with learning disabilities experience extreme difficulties in making academic progress, despite at least average intelligence and conventional interventions (Childs & Finucci, 1983; Hooper & Willis, 1989; Hynd, 1988; Rourke, 1991; Wilson, 1985; Winzer, 1993). The consequences of learning disabilities are immeasurable, impacting negatively on the educational, emotional, and behavioural well-being of the student (Rourke, 1991). Moreover, the difficulties experienced by individuals with learning disabilities endure (Rourke & Fuerst, 1991; Spreen, 1988) and the deficits associated with learning disabilities typically persist into adulthood (Kaste, 1971; Mendelson, Johnson, & Stewart, 1971; Silver & Hagin, 1964; Spreen, 1982). Learning disabilities are the "largest single focus of special education in many school districts" (Winzer, 1993, p. 243). The study of learning disabilities and associated academic problems is burgeoning into an intensive area of investigation with growth that has been described "as little short of phenomenal" (Winzer, 1993, p. 243). Further, based on both prevalence and costs, the learning disabled population presents practical educational problems of major importance (Keogh, 1990).

Gaddes (1981) and Pirozzolo (1979) estimated that between 10 to 15% of children show seriously deficient academic attainment, although prevalence figures depend on the definition and procedures used to identify learning disabled subjects. In Alberta, it was estimated that, in

1991, there were just over 21,000 students identified as learning disabled, and of over 50,000 students identified as "exceptional", 40.6% were identified as learning disabled (Alberta Education, 1992). In Canada, students with learning disabilities make up the largest single group of children with disabilities, and generally, the best Canadian estimates place the number of students with learning disabilities at from 2 to 4% of the school-age population (Winzer, 1993). Duane (1979) estimated that with increased survival of high risk infants the population of children with learning disabilities will exceed the combined population of children with seizure disorders, cerebral palsies, and severe mental retardation. The consensus of scholarly opinion strongly suggests that cerebral dysfunctions underlie this disorder (Gaddes, 1985; Hooper & Willis, 1989; Hynd & Obrzut, 1981; Obrzut & Hynd, 1986; Pirozzolo, 1979; Rourke, 1991; Winzer, 1993) and the use of intelligence tests for classifying learning disabled children is "entrenched in every form of work with these children" (Francis, Espy, Rourke, & Fletcher, 1991, p. 15).

Winzer (1993, p. 243) maintained too, that the field of learning disabilities brought "changes and innovations to the entire field of special education" in the areas of instruction, assessment, and conceptualization of mild handicaps. Thus, it is important that efforts continue to be directed toward establishing reliable and valid means of identifying the presence of learning disabilities in order to facilitate remedial programming (Hooper & Willis, 1989). This is a complex task made all the more difficult by the fact that research suggests that there are numerous different subtypes of learning disability (Hynd, 1988), as well as controversy over conceptualization and operational definitions of learning disabilities (Keogh, 1990, Winzer, 1993). "The differential diagnosis of learning disability subtypes is a critical first step in developing theoretically sound programs of psychoeducational intervention" (Hynd, 1989, p. vii). Winzer (1993) stated "there is almost universal agreement on the need for efficient diagnosis of students with learning disabilities" (p. 253). Research techniques need to be directed toward the development of valid differential diagnostic procedures, based on theoretical clinically relevant classification schema (Adelman & Taylor, 1985, 1986). Evidence suggests such subtypes exist (Hynd, 1989) and currently much effort is being spent on determining subtypes according to patterns of disorder, particularly

"different areas of underlying cognitive or psychologic dysfunction" (Forness, 1990, p. 195). Early attempts at subtyping were typically characterized by clinical inferential approaches (Forness, 1990). However, empirical multivariate cluster analytic procedures hold promise in developing classification schema in this area (Adelman & Taylor, 1985; Hooper & Willis, 1989; Kavale, 1990; Rourke, 1991) and research utilizing empirical approaches is growing significantly (Kavale, 1990). The role of the SB:FE in the diagnosis of learning disabilities, and its potential to contribute to learning disability subtyping research, in particular, needs to be investigated in order to improve current diagnosis and facilitate better remedial programming. Moreover, Lyon and Risucci (1988) emphasized that the scope of classification transcends mere categorization, but is also concerned with enhancing the theoretical understanding of learning disabilities. Similar lines of reasoning underscore the need to accurately identify and further understand mild handicaps such as Mental Retardation (Shepard, 1989) and the use of standardized intelligence tests is integral to work with such populations (Winzer, 1993). The potential of the SB:FE to contribute to these areas of understanding is presently unclear.

However, the SB:FE poses particular problems for multivariate research. The SB:FE scales are characterized by their adaptive testing and age scale formats (Thorndike et al., 1986a). It is possible for each of the four Reasoning Area scores, and therefore Composite Standard Age Scores (SAS's) to be composed of various numbers and different subtests. This tendency becomes more marked at different age levels. These difficulties are further compounded by the possibility that several tasks are thought to involve different abilities which depend on developmental levels (Keith, 1987; Keith et al., 1988a). It is worth noting that these methodological concerns, in general, may be applicable to virtually all measures of intelligence that span various age ranges. These concerns are exacerbated by the SB:FE's adaptive format and wide age coverage.

In light of the difficulties inherent in multivariate research with the SB:FE, a major purpose of this proposed research was to investigate the relationship between the SB:FE GPAB and the SB:FE, and to investigate the utility of the abbreviated battery in educational diagnosis and

classification. Ideally, use of the abbreviated version may provide a means of ameliorating the difficulties that arise from the idiosyncratic nature of the SB:FE adaptive testing format. As noted, the complete battery for the SB:FE consists from eight to thirteen tests, depending on the age and ability level of the subject. The GPAB comprises six subtests that should be administered to any subject assessed with the SB:FE, regardless of age or ability level (Thorndike, Hagen, & Sattler, 1986b). Thorndike et al. (1986b) indicated that for all age ranges internal consistency (KR-20) (Kuder & Richardson, 1937) reliabilities of the abbreviated version are satisfactory (around .95), and Composite SAS's derived from the abbreviated version correlate very highly with Composite SAS's derived from the complete battery (r=.94-.98). These claims need to be investigated, particularly in terms of the relationship of the abbreviated battery to academic achievement within a Canadian population.

Research Objectives

The research reported here was designed to meet four main objectives:

1. To provide comprehensive descriptive data about the SB:FE used with a Canadian clinic sample;
2. To explore the relationship between the SB:FE full battery and the SB:FE GPAB within this sample;
3. Given close agreement between the two, to then investigate the applicability of multivariate cluster analytic procedures to SB:FE GPAB data in order to derive reliable and replicable (internally valid) groups of individuals with distinct cognitive profiles; and
4. To explore the external validity of groups derived through application of multivariate clustering procedures to SB:FE GPAB data through investigation of subgroup differences on the basis of external achievement criteria and to explore the agreement between empirically derived subtypes and clinical inferential models.

Throughout this study, short hand notation has been used to designate the age groupings. For example, 2-23-11 means 2 years through 23 years, 11 months; 2-6-11 means 2 years through

6 years, 11 months; 7-11-11 means 7 years through 11 years, 11 months; and 12-23-11 means 12 years through 23 years, 11 months.

Delimitations

The primary restrictions of this research centered around the possible idiosyncratic nature of the sample. Issues relating to the sample are more fully addressed in Chapter Three and the final chapter. The study was restricted to a post hoc or retrospective analysis of data. With respect to exploration of learning disabilities, only academic deficits which occurred in the three major areas of difficulty (reading, arithmetic, and spelling) (Winzer, 1993) were considered. The intent of the study was to provide evidence of the SB:FE's validity with Canadian subjects. Messick (1989) pointed out that "validity is an inductive summary of both the existing evidence for and the potential consequences of score interpretation and use" and then cautioned that "validity is a matter of degree, not all or none" maintaining that it is "an evolving property" and "a continuing process" (p. 13). This research was designed to be a part of this ongoing process in generating empirical evidence about the validity of the SB:FE in educational settings, within the parameters identified.

Outline of the Dissertation

This remaining portion of this dissertation is organized in seven chapters. In Chapter Two, a review of the development of the Binet intelligence scales culminating in the 1986 SB:FE is presented. An overview of learning disabilities is also provided, and empirical and clinical clustering or subgrouping research procedures are reviewed. The methodology used to address the objectives of the study (see p. 8) is presented in Chapter Three. The results are presented in four chapters corresponding to the research objectives. In Chapter Four the descriptive data about the SB:FE used with a Canadian clinic sample is presented. In Chapter Five the relationship between the SB:FE full battery and abbreviated version is presented. The results of the multivariate clustering procedures applied to SB:FE GPAB data are presented in Chapter Six. Chapter Seven contains the results of external validation procedures applied to the clusters derived through application of empirical clustering procedures to SB:FE data. A summary of the

study, together with a discussion of the conclusions and the implications for school psychology practice and future research appear in the final chapter.

CHAPTER TWO: REVIEW OF RELEVANT LITERATURE

Overview

The present chapter begins with a description of the development of the SB:FE and the SB:FE GPAB. Next, their psychometric properties are examined. This is followed by a review of trends in educational assessment and planning. Emphasis is placed on a discussion of learning disabilities and definitional and historical issues related to learning disabilities. The conceptualization of learning disabilities as a heterogeneous multi-factor construct is examined. Finally, the notion of subtypes in learning disabilities is introduced and the validity of subtyping or subgrouping in research is reviewed.

Development of the Stanford-Binet Intelligence Scales

Binet and Simon (1905, 1908) initially developed the Binet-Simon scale to provide a screening instrument that would enable the French Minister of Public Instruction to identify mentally retarded children (Fancher, 1985). The scale comprised a series of 30 tasks of increasing levels of difficulty, standardized on groups of about 50 normal children of varying ages and 45 subnormals of varying degrees (Fancher, 1985). The 1905 test was atheoretical and empirically derived. In 1908, Binet and Simon published an extensive revision consisting of 58 items, again arranged in order of increasing difficulty, located at specific age levels between three and 13 years. The concept of mental "level" was also introduced (Freeman, 1955), although Binet cautiously did not use the term "mental age" (Fancher, 1985). In 1911, the third revision of the Binet-Simon scale was published by Binet alone, who extended the scale to include 15 year olds and a limited adult category (Fancher, 1985). Several items were relocated to higher age levels and several omitted, so that there were five items for each age level (Anastasi, 1982).

In 1909, Goddard translated the Binet-Simon Scale from French to English and introduced the scale to the United States with a number of revisions (Sattler, 1988, 1992). In 1911, Goddard tested 2,000 children for standardization purposes (Thorndike & Lohman, 1990) and became one of the world's leading proponents of Binet's testing methods (Fancher, 1985). Other researchers (Kuhlmann, 1912) published English translations of the Binet scales in the United

States (Thorndike & Lohman, 1990). However, in 1916, Terman of Stanford University, completed and published the most successful revision of the 1908 scale. This test was called The Stanford Revision of the Binet-Simon Scale (Terman, 1916). The revised test covered the age range from three to 16 years and was standardized on 1000 Californian children. In addition, groups of items were included for average and superior adult levels. A total of 90 items were included in the 1916 scale, of which 54 had been adopted from the 1911 Binet scale. Although the test remained basically an age-scale yielding a mental age, Stern's Intelligence Quotient (IQ) ratio (Stern, 1914) was adopted in order to report responses on the age scale in a condensed form. This ratio was calculated by the formula:

$$IQ = \frac{\text{Mental Age (MA)}}{\text{Chronological Age (CA)}} \times 100$$

Standardized administration procedures were also instituted, although these remained somewhat subjective and problematic (Freeman, 1955). The test was primarily designed as a measure of global intelligence and no attempt was made to measure separate mental faculties, although the distributions of IQ's was basically normal at each age group (Freeman, 1955). The test subsequently became known as the Stanford-Binet Intelligence Scale, and despite its limitations, became the standard against which all subsequent American intelligence tests would be measured (Fancher, 1985).

Terman and Merrill (1937) revised the scale again, extending the age range of the instrument from age 2 to 18 years and attempting to improve the standardization (based on 3,184 white American born subjects). The emphasis on measurement of general intellectual ability, rather than specific abilities was maintained. Age scale formats and the ratio IQ were retained. Two forms, L and M, were developed, each comprising 129 items. Sattler (1988) commented that the 1937 revision was "recognized as a milestone in the progress of the individual testing of intelligence" and noted that it had "excellent reliability and validity" (p. 246). Factor analytic studies indicated that most tests loaded heavily on a common factor and the tests "served as important tools in clinical and educational settings" (Sattler, 1988, p. 248).

The instrument was revised for a third time in 1960 (Terman & Merrill, 1960). No new content was introduced in this revision, but the best items from the two forms were selected and merged into the Form L-M. Deviation IQ's or standard scores derived from Yerkes (1917) and Wechsler (1939), although with a mean of 100 and a standard deviation of 16, were adopted for the first time in order to present test results for the sample age 2 through 18 years. The concept of mental age, however, was not abandoned. Norms were based on the 1937 sample and a sample of 4,498 subjects who had taken the scale between 1950 and 1954 was used to explore changes in item difficulty and to determine placement of items on the new form. The test remained a measure of general ability (Sattler, 1982).

R. L. Thorndike restandardized the 1960 revision, providing new interpretive norms in 1972, based on a more representative sample of 2,100 nonwhite and white children (Sattler, 1988). The revised norms were published in 1973 (Terman & Merrill, 1973). However, the test items in the scale, and directions for scoring and administration remained the same, and the test yielded a single score or measure of general intelligence. Salvia and Ysseldyke (1985) noted weaknesses in the norming, reliability, validity, and standardization of the 1972 Binet scale. Sattler (1982) provided a comprehensive review of the test.

It is worth noting that the successive revisions of the Stanford-Binet were intended as scales with a unitary focus. All purported to measure general intelligence. However, factor analytic studies (Burt & John, 1942a, 1942b; Hallahan, Ball, & Payne, 1973; Jones, 1949; McNemar, 1942; Ramsey & Vane, 1970; Thompson, 1984; Wright, 1939) suggested the presence of group factors (e.g., memory, verbal, visual-spatial, or numerical) in all editions of the test. In addition, earlier versions of the Stanford-Binet were consistently criticized for over-emphasizing verbal abilities (Krohn & Lamp, 1989). Moreover, the need for a revision of the Binet was paramount as during the 1970's and 1980's use of the Binet scales declined drastically (Lubin, Larsen, & Matarazzo, 1984; Lubin, Larsen, Matarazzo, & Seever, 1985). Thus, in light of current educational trends emphasizing differential abilities and specific areas of educational need (Sattler 1988), previous factor analytic findings, and criticisms of earlier versions of the Binet scales,