ANALYSIS OF ITEM CHARACTERISTICS OF THE SLOSSON INTELLIGENCE TEST FOR BRITISH COLUMBIA SCHOOL CHILDREN

By

BARBARA KATHLEEN GARD

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF
THE REQUIREMENTS FOR THE DEGREE OF
MASTERS OF ARTS

in

THE FACULTY OF EDUCATION
(Department of Educational Psychology and Special Education)

We accept this thesis as conforming
to the required standard

THE UNIVERSITY OF BRITISH COLUMBIA
March 1986

78

In presenting this thesis in partial fulfilment of the requirements for an advanced degree at the University of British Columbia, I agree that the Library shall make it freely available for reference and study. I further agree that permission for extensive copying of this thesis for scholarly purposes may be granted by the head of my department or by his or her representatives. It is understood that copying or publication of this thesis for financial gain shall not be allowed without my written permission.

Department of Educational Psychology & Special Education

The University of British Columbia
1956 Main Mall
Vancouver, Canada
V6T 1Y3

Date March 4, 1986

DE-6 (3/81)

ABSTRACT

This study investigated item characteristics which may affect the validity of the Slosson Intelligence Test (SIT) when used with school children in British Columbia. The SIT was developed as a quick, easily administered individual measure of intelligence to correlate highly with the Stanford-Binet Intelligence Scale as an anchor test. Use of the SIT has become widespread, but little technical information is available to support this.

To examine the internal psychometric properties of the SIT for British Columbia schoolchildren, SIT responses were collected from 319 children (163 males, 156 females) in three age groups (7 1/2, 9 1/2, and 11 1/2 years). These data were subjected to a variety of item analysis procedures. Indices were produced for: item difficulty, item discrimination (item-total test score correlations), rank correlation between empirically determined item difficulties and item order given in the test, test homgeneity, and item-pair homogeneity.

Results of the item analyses suggest that the SIT does not function appropriately when used with British Columbia school children. Two-thirds of the item difficulty indices were found to be outside the desired range: one-third of the items did not discriminate effectively; and many items are not in correct order of difficulty in administration of the SIT. The thesis discusses effects of these findings on the test's internal consistency, criterion validity, and technical utilization. Factors which may underlie the shift in item difficulties are also discussed.

Robert Conry, Ph.D.
Research Supervisor

TABLE OF CONTENTS

TABLE OF CONTENTS CONT'D

LIST OF TABLES

ACKNOWLEDGEMENTS

I wish to express my appreciation to Dr. Barbara Holmes for her kindness in permitting me to use for the purpose of my analyses, the SIT data she collected as part of her large province-wide anchor-norming study.

I also wish to thank my committee chair, Dr. Robert Conry, for his many readings of drafts and invaluable advice; my committee members, Dr. Julie Conry and Dr. Buff Oldridge, for their suggestions and support; and my advisor, Dr. Emily Goetz, for her friendly encouragement to bring this project to a close.

Chapter I

Introduction

The Slosson Intelligence Test (Slosson, 1977, 1983) was developed to meet a need for an easily administered, brief test of intelligence. Since its introduction over twenty years ago, the Slosson Intelligence Test (SIT) has proven popular (Brown & McGuire, 1976). The appeal of the SIT derives from the test's easy administrative and scoring procedures, brevity, low cost, and availability to a broad range of professionals without specific training in intelligence testing. However, review of the technical information available regarding the development of the test suggests that a paucity of research underlies the widespread acceptance of the SIT. It is the purpose of this study to examine the internal psychometric properties of the SIT in regard to the use of the test with British Columbia schoolchildren.

The benefits of standardized tests such as the SIT derive from careful technical construction and selection of test items, and large representative norming procedures. Examination of SIT manuals (Slosson, 1977, 1983) indicates a lack of technical information pertaining to the development of the test. For example, norm tables are based on data collected from a small, sketchily described, regional sample, only one reliability measure is reported, and validity indices cited are based on studies with limited samples. These technical weaknesses suggest that the psychometric properties of the SIT need to be further examined in relation to those populations to whom the test is administered.

Evaluation of the psychometric properties of a test provides information as to how effectively the construct tested is being measured (Jensen, 1980). The functioning of a test of general mental ability, such as the SIT, can be judged through examining the properties of test items relative to the test as a whole (Jensen, 1980). To determine item characteristics, technical procedures known as item analyses are conducted. Item analyses provide information regarding the shape and distribution of test scores, discrimination among test takers, test score variance and the internal consistency reliability of the test (Jensen, 1982; Nunnally, 1978). To evaluate the effectiveness of the SIT in terms of the internal psychometric properties of test items for a British Columbia sample of schoolchildren, SIT responses, collected as part of a province-wide norming study (Holmes, 1981), were subjected to a variety of item analytic procedures. These included analysis of item difficulty, analysis of item-total test score correlations, correlation of the rank order of item difficulties between the British Columbia and norm sample, test homogeneity of adjacent item pairs (item-test homogeneity). For valid and reliable test measures, these item indices should remain relatively stable across the populations to whom the test is applied.

The Slosson Intelligence Test

The SIT is an individually administered test composed of a series of primarily verbal questions arranged in order of increasing difficulty. Each question is credited as a pass or fail depending on whether or not a correct response is given. The starting point varies for each individual based on age and ability, and testing stops when consecutive failures occur.

Since the SIT was designed for administration across a broad-age range, test items of varying difficulty were selected and rank ordered from easiest to most difficult. Rank ordering of items permits the use of basal and ceiling points to shorten test administration without losing test reliability achieved from test length. The basal point is the test item below which it can be assumed that all earlier (less difficult) items will be answered correctly if given, and the ceiling point is the test item above which failure can be assumed on all higher placed (more difficult) items if given. For the SIT, the basal and ceiling points are set, respectively, at ten consecutive correct and incorrect, and individuals are tested only on the subset of items which fall between his or her basal and ceiling points.

The development of the SIT, both in structure and content, was based largely upon the 1960 Stanford-Binet, and high correlations obtained between the two tests were interpreted as a strong indication that the SIT provided a valid substitute for the Stanford-Binet. In 1972 the Stanford-Binet was renormed and it was found that IQ scores dropped by approximately six points, reflecting an increased sophistication among the general population (Terman and Merrill, 1973). The change in the Stanford-Binet IQ tables resulting from the 1972 renorming lowered the correlation between the SIT and the Stanford-Binet and the SIT norms required revision in order to re-establish high correlation validity with the Stanford-Binet.

The equi-percentile method of equating test scores was selected to rescale SIT IQ's to the 1972 revision of the Stanford-Binet. This method involves using the scores of one test (the 1972 Stanford-Binet) as an anchor against which the second test's (the SIT) scores are distributed and matched according to percentile ranks. The SIT norm tables were developed from ratio IQ's and in order to obtain deviation IQ's which correlated highly with the 1972 Stanford-Binet deviation IQ's, the SIT simply matched IQ's along chronological and mental age. As test content on the Stanford-Binet was unchanged, raw scores (mental age) had remained constant and could function as the equating variable (Armstrong & Jensen, 1982, 1984). The 1981 norms were published in the second edition of the SIT (Slosson, 1983) and a correlation of .95 with the 1972 Stanford-Binet was reported (Armstrong & Jensen, 1982, 1984).

Importance of the Study

The study reported here addresses several important issues relevant to the administration and interpretation of the SIT to British Columbia schoolchildren. First, the SIT is generally accepted as a test of general intellectual ability (Brown & McGuire, 1976). However, the widespread use of the SIT is not validated by empirical research available supporting the test's claims. Second, the technical information available on the SIT is meager and outdated: the test questions and item order are based on data collected on a limited sample over twenty years ago. The validity and reliability of any test application and interpretation rests upon the technical strengths of the test itself. Third, the 1981 renorming of the SIT failed to improve the technical weaknesses of the earlier version. No item analyses are reported in the 1983 manual, a representative norm sample was not collected for the purpose of rescaling the IQ scores, and the renorming consisted only of matching IQ's through use of the equi-percentile method to equate them to the 1972 Stanford-Binet norm tables. Fourth, the SIT's psychometric properties have not been examined relative to the British Columbia population so as to support local use and interpretation of test results.

Chapter II

Literature Review

This chapter reviews literature relevant to the use of the Slosson Intelligence Test for use with British Columbia schoolchildren. The development of the SIT and the technical data presented in the test manual are reviewed, and its standardization, validity, and reliability are discussed.

Initial Development of the Slosson Intelligence Test

Richard L. Slosson designed the Slosson Intelligence Test (SIT) as a short, easily administered measure of intelligence. The major intelligence tests available prior to the development of the SIT were the Stanford-Binet and the Wechsler Scales of Intelligence, which require extensive training in administrative, scoring, and interpretive procedures. The SIT was constructed to be attractive to professionals who need an indication of an individual client's intellectual ability, but who either are not specially trained in intelligence test administration or want a less time-consuming assessment instrument. For example, the SIT manual states that "the test has been made for the use of school teachers, principals, psychometrists, psychologists, guidance counselors, social workers, school nurses and other responsible persons who, in their professional work, often need to evaluate an individual's mental ability" (Slosson, 1977, p.iii) and that the SIT "yields sufficiently valid IQ's, for children four years of age into adulthood, as to furnish a useful screening instrument in the hands of responsible, professional persons" (p.viii). Although Slosson uses the term "screening instrument," the interpretative potential of IQ scores from the SIT is equated with that of the Stanford-Binet. For example, the manual states that "IQ is a numerical score...(which) gives an indication of a person's ability to learn, solve and understand problems. It is a 'rough' measure of an individual's capacity to reason, judge and retain knowledge" (1977, p.24); also, that "it is generally proposed and accepted that the results of IQ tests should be or can be used as achievement predictors" (1983, p.41), and that "the Slosson (SIT) is a valid, reliable, individual IQ test that achieves its stated purpose" (Slosson, 1983, p.49). Introduced in 1961, the SIT quickly gained popularity. For example, a survey of test use in clinics across the United States found the SIT to be one of the ten most frequently administered tests across all age levels (Brown & McGuire, 1976). Frequent use of the SIT has also been noted in British Columbia schools, the arena for this study (Holmes, 1981).

The design of the SIT is founded on the assumption that individuals gain knowledge over time. Like the Stanford-Binet, the SIT is composed of a series of test questions ordered by ascending difficulty. SIT items are assigned a chronological age equivalent corresponding to the age at which a child of average ability is expected to pass or fail the item. A number of months of mental age credit is assigned to each question, with the total obtainable per chronological year level equal to twelve. To obtain a full year's mental age credit, all questions within that age level need to be answered correctly.

An individual's Intelligence Quotient (IQ) is determined by comparing the number of months mental age credit received from correct test responses to the individual's chronological age. For example, an individual of average intelligence would be expected to answer test questions correctly up to a difficulty level equal to his chronological age, while an individual with above-average intelligence would be able to answer test items rank ordered in difficulty above his chronological age. SIT test questions range from an infant level (0-0.5 months) to an adult level (27-0 years).

SIT item design was based on two well-established measures of cognitive functioning (Slosson, 1977). At the infant and early childhood age levels, test questions follow the format of the Gesell Infant Scale of Development and are comprised primarily of performance-type items which involve fine motor skills, while school-age child to adult level questions are based on the Stanford-Binet. From items 5-4 up, all questions are administered verbally and require a verbal response.

IQ determination is based on the sum of mental age credits given for all items answered correctly plus credit for all pre-basal items. No credit is given for items after the ceiling point. The SIT originally used the ratio IQ formula to determine IQ where IQ = MA/CA x 100 (MA = number of months mental age credit obtained; CA = chronological age). The 1981 revision of the SIT involved equating or rescaling SIT IQ's to match the 1972 Stanford-Binet deviation IQ norm tables.

SIT Test Construction

The SIT test manual (1977, 1983) provides scanty information regarding the construction of the SIT. The manual states that item design was based on the Stanford-Binet test questions and that "the most favorable" (p.iv) items were selected over several years of testing (Slosson, 1977). Items which teachers reported to be difficult to administer or score were eliminated. No other selection criteria and no item statistics are provided.

Norm Population

No account of the norming procedures is given in the 1977 SIT test manual. Sample size and stratification information such as age, sex or socio-economic status of the sample is not detailed. The information given regarding the sample used for concurrent validity studies of the SIT with the Stanford-Binet is:

> The children and adults used in obtaining comparative results, came from both urban and rural populations in New York State. The referrals came from cooperative nursery schools, public, parochial and private schools, from junior and senior high schools. They came from gifted as well as retarded classes -- White, Black and some American Indian. Some came from a city Youth Bureau, some from a Home for Boys. The very young children resided in an infant home. The adults came from the general population, from various professional groups, from a university graduate school, from a state school for the retarded and from a county jail (Slosson, 1977, p.iv).

No further details regarding the sample composition are provided.

Development of the 1981 norm tables was based on a sample of 1109 subjects, aged 2 years 3 months to 18 years. Data was collected between 1968 and 1977 and included some of the original sample data (S. Slosson, personal communication, February 13, 1986). The sample was drawn only from the New England area. Sample distribution was analyzed within four age groups: below 6-6, 6-7 to 10-7, 10-7 to 13-6, 13-7 and above; and within three IQ ability levels: below 84, 84-116, above 116 (Armstrong & Jensen, 1982; Slosson, 1983). No other sample characteristics or selection procedures are described.

SIT Reliability

Only one reliability measure is reported in the SIT test manuals (Slosson, 1977, 1983). A test-retest reliability of .97 was obtained over a two-month interval for a sample of 139 subjects aged four to fifty. No new reliability information is given,in the 1983 manual but the SIT technical manual reports an additional test-retest correlational finding of .93 for a sample of 350 individuals over a ten-week interval (Armstrong & Jensen, 1982).

SIT Validity

The SIT (1977) reports "sufficient" test validity, based on a total of nine concurrent validity studies (p.viii). Correlation coefficients are given for only four of the studies while IQ scores alone are reported in the other studies. Six studies report the relationship of the SIT with the Stanford-Binet, two with the WAIS or WISC as well as the Stanford-Binet, and one with the Cattell Infant Intelligence Scale. Six of the studies interpret findings based on data from less than 25 individuals. Concurrent validity coefficients, reported by age levels (four and up), for the SIT and the Stanford-Binet fall in the mid-90's (r =.23 to .71). On the basis of the high correlations obtained between Stanford-Binet and SIT IQ scores, the test author concludes that the SIT is a valid assessment instrument for individuals four years of age and up (Slosson, 1977, 1983). The 1983 edition of the SIT reviews validity data from various correlational studies of the SIT with the Stanford-Binet, Wechsler scales and other achievement measures. A median correlation of .90 (range .96 to .60) is reported for 18 studies correlating Stanford-Binet and SIT IQ scores carried out between 1963 and 1974; a median correlation of .75

with Wechsler full-scale IQ's (range .96 to .52), .82 with Wechsler verbal IQ's (range .96 to .44) and .62 with Wechsler performance scale scores (range .84 to .10) was based on 18 correlational studies conducted between 1968 and 1974. The median of eighteen correlational studies between the SIT and various achievement tests was found to be .55 with a range of .83 to .24.

Concurrent Validity of the Stanford-Binet and the SIT

A ten-year review (1963-1974) of research involving the concurrent validity of the SIT was carried out by Stewart and Jones (1976). Ten concurrent validity studies of the SIT and the Stanford-Binet were reviewed (Armstrong & Jensen, 1972; Armstrong & Mooney, 1971; Carlisle, Shinedling & Weaver, 1970; DeLapa, 1973; Johnson & Johnson, 1971; Jongeward, 1968; Lamp, Traxler & Gustafson, 1973; Ritter, Duffey & Fischman, 1973; Stewart, Wood & Gallman, 1971; Stewart & Myers, 1974). Validity correlation coefficients ranged from .60 to .94 with a median of .90. Although some studies did not obtain validity coefficients as high as those reported in the SIT manual (DeLapa, 1973; Johnson & Johnson, 1971; Jongeward, 1968; Lamp et al., 1973; Stewart & Myers, 1974), the median correlation of .90 generally supports Slosson's finding that the SIT measures a construct similar to that of the Stanford-Binet. Mean IQ scores obtained on the two measures differed by four points or less in each of the 10 studies. Stewart and Jones concluded that the ranked ordering of children on the SIT and the Stanford-Binet was nearly equivalent. However, they cautioned against substitution of the SIT for the Stanford-Binet because large enough discrepancies occurred between IQ scores on the two tests to have resulted in misclassification of a significant proportion of individuals.

Rotatori, Sedlak, and Freagon (1979) administered both tests to 40 severely or profoundly retarded children, aged 11 to 19. A correlation of .90 was obtained between IQ scores on the two measures, which concurs with the earlier findings. High agreement was found between the rank ordering of individuals on the two tests. However, it was noted that scores on the SIT were more than seven points higher than those on the Stanford-Binet in 75% of the cases. Rogers (1982) cautioned against use of the SIT after finding that IQ scores on the SIT were 9 to 40 points ($\overline{x}$ = 20) higher than Stanford-Binet IQ scores for nine 3 to 6 year olds. The larger difference in IQ scores obtained between these two studies and the earlier studies may be an artifact of the renorming of the Stanford-Binet in 1972.

Concurrent Validity of the SIT and the Wechsler Scales

Stewart and Jones (1976) also summarize the findings of 15 studies reporting on the correlation of the SIT with WISC or WAIS IQ's (Houston & Otto, 1968; Jerrolds, Calloway & Gwaltney, 1972; Jongeward, 1968; Kaufman & Ivanoff, 1969; Lamp et al., 1973; Lessler & Galinksy, 1971; Martin & Rudolph, 1972; Maxwell, 1971; Stewart et al., 1971; Stewart & Myers, 1974; Swanson & Jacobson, 1970). Overall, it was noted that the SIT correlated highest with the Wechsler Verbal Scale, slightly lower with the Wechsler Full-Scale, and considerably lower with the Performance Scale. SIT correlations ranged from .52 to .96 with a median of .83 with the Verbal Scale; from .44 to .94 with a median of .74 with the Full-Scale; and from .10 to .84 with a median of .65 with the Performance Scale. Stewart and Jones (1976) conclude that it is not "justifiable to treat the SIT IQ as a direct substitute for the Wechsler IQ" (p.375) because the two tests differ in the skills they measure, especially the Wechsler Performance Scale. They also note that use of the SIT usually results in substantially higher IQ scores than the Wechsler scales, which could lead to misclassification of intellectual ability.

A number of studies reporting Wechsler-SIT correlations have been published in the ten years since the Stewart and Jones review (Baum & Kelly, 1979; Covin, 1977a, 1977b; Crofoot & Bennett, 1980; Dirks, Wessels, Quarforth & Quervon, 1980; Lowrance & Anderson, 1979; Mize, Calloway & Smith, 1979; Rotatori et al., 1979; Rust & Lose, 1980; Smith, 1981; Vance, Lewis & DeBell, 1979). In general, their findings support the conclusions drawn from the earlier studies. SIT IQ scores correlate highest with the Wechsler Verbal Scale (range .41 to .92; median .61) and lowest with the Performance Scale (range .003 to .70; median .51). A number of studies found the SIT to yield significantly different IQ scores than the Wechsler. The SIT tended to overestimate IQ scores at the higher end of the IQ range and to underestimate IQ at the lower range of intelligence (Baum & Kelly, 1979; Covin, 1977a, 1977b; Crofoot & Bennett, 1980; Dirks et al., 1980; Lowrance & Anderson, 1979; Mize et al., 1979). These researchers caution against substitution of the SIT for the Wechsler scales because of the potential misclassification of individuals.

For the British Columbia sample used in this study, Holmes (1981) reports a correlation of .75 between the SIT and the WISC-R Verbal Scale, .48 with the Performance Scale, and .71 with the Full-Scale.

Concurrent Validity of the SIT with Achievement Tests

The correlation of various achievement tests with the SIT has also been reported in a number of studies. Stewart and Jones (1976) summarize fourteen studies carried out between 1967 and 1974 with a variety of tests including the Wide Range Achievement Test, the Peabody Picture Vocabulary Test, and the California Achievement Test. Correlations ranged from .24 to .83 with a median correlation of .55.

A review of the research shows that more recent studies have found similar correlations between the SIT and achievement tests (Baum & Abelson, 1981; Cianflone & Zullo, 1980; Colarusso, McLesky & Gill, 1977; Coleman, Brown & Ganong, 1980; Covin, 1977a, 1977b; Crofoot & Bennett, 1980; Grossman & Johnson, 1983; Hale, Douglas, Cummins, Rittgarn, Breeds & Dabbert, 1978; Klein, 1978; Martin, Blair & Vickers, 1979; Rust & Lose, 1980; Smith, 1981; Vance et al., 1979). Correlational findings of the SIT with achievement tests, including the Peabody Picture Vocabulary Test, the Wide Range Achievement Test, Stanford Achievement Tests, the McCarthy, and the Shipley Institute of Living Scale, ranged from .31 to .94 with a median of .56. Similarly, a correlation of .62 was found between the SIT and the PPVT for a British Columbia norming sample (Holmes, 1981).

Content Analysis of SIT Items

Several researchers have examined the content of SIT items in terms of the intellectual functions they measure relative to the Stanford-Binet and the WISC. Nicholson (1970) applied Sattler's Stanford-Binet classification scheme to SIT items and Stone (1975) adapted Valett's classification scheme to determine the degree of similarity of item content between the two tests. Both reports note a high, but not exact, correspondence between the proportion and type of mental functions evaluated. Boyd (1974) and Fudala (1979) analyzed the item content of the SIT relative to the WISC and conclude that SIT item content corresponds to the WISC Verbal Scale. Comparison of the four categorization schemas shows no major discrepancies between classification of individual items if allowance is made for the different terms used by Nicholson (i.e. language for vocabulary, social intelligence for information, memory for digit span). SIT items were categorized as vocabulary, information, arithmetic, similarities, digit span, and visual-motor.

Summary

In summary, the above review of the development of the SIT and studies which have examined the test's concurrent validity indicate technical weaknesses in both original and revised editions of the SIT. Test norms are limited, reliability information is lacking, item statistics are not given, and concurrent test validity is based on small samples. These areas of weakness suggest a need for further evaluation of the SIT for the populations to whom the test is given.

Purpose of the Study

The present study is designed to examine the internal psychometric properties of the SIT in relation to use of the test with British Columbia schoolchildren. It should be noted here that Holmes' (1981) data used for the present item analyses was collected on the 1977 edition of the SIT. The findings of this study, however, are equally applicable to the administration of the revised edition of the SIT (1983) because no changes were made in the test items themselves, or in the order of their presentation. As test items and item order are identical in the original and the revised editions of the SIT, administration procedures and questions asked remain the same for both editions. IQ scores which do differ between editions, are not involved in item analysis conducted in this study. Therefore, the findings reported in this paper are applicable to the interpretation of results arising from use of either edition of the SIT.

Five research questions are addressed in this study and all of them relate to the adequacy of use of SIT with British Columbia schoolchildren:

1. How adequate or effective are the range and distribution of SIT item difficulty indices?
2. How adequate or effective are the range and distribution of SIT item total test score correlations (item discrimination)?
3. How adequate or effective is the correlation between the rank order of SIT item difficulties for the British Columbia sample and the rank order given in the test?
4. How adequate is the SIT's test homogeneity?
5. How adequate is the range and distribution of adjacent item pair homogeneity?

Chapter III

Methodology

This chapter presents methods used to collect and analyze the data. Subject characteristics, testing procedures, and data analysis methodology are outlined.

Sample Characteristics

In this investigation, analysis of SIT item reliability for British Columbia children used Holmes' (1981) data collected to norm several psycho-educational measures frequently administered to British Columbia schoolchildren (Wechsler Intelligence Scale for Children-Revised, Raven's Standard Progressive Matrices, the Peabody Picture Vocabulary Test, the Mill Hill Vocabulary Test, and the Slosson Intelligence Test).

Holmes selected children in three age groups as a representative sample of the British Columbia population of schoolchildren. The stratification variables identified were based on those used in the standardization of the WISC-R (Wechsler, 1974) and included: age, sex, geographic region, community size, and size of school. A breakdown of sample characteristics is given in Appendix A.

SIT tests were given to a total of 319 children (163 males and 156 females) in three age groups: 7 1/2 year olds ($\underline{n}$ = 108), 9 1/2 year olds ($\underline{n}$ = 111), and 11 1/2 year olds ($\underline{n}$ = 100). At the time of testing children were within 3 months of the midyear, i.e. 7 years 3 months to 7 years 9 months. The three age groups correspond to grades 2, 4, and 6 and were chosen to be representative of elementary schoolchildren.

Data Collection

The SIT (1977) was individually administered to each child in the study. Testing was conducted during school hours in a quiet location within the child's school. Test administration was counter-balanced. All tests were administered by trained personnel familiar with SIT test procedures.

Data Analysis

As basal and ceiling points were not constant for all children, the set of children and the number of responses varied among SIT items. In order to carry out item analyses, a set of items administered to a majority of the children in the sample was identified. These items are referred to as the common item range, and for the purpose of item analysis was established as those items given to fifty percent or more of the children tested for each of the three age levels. This criterion was used in an item analysis of the Peabody Picture Vocabulary Test (Berry, 1977) which also had variable data per individual. It was adopted in the present study for comparability of research method, and because it provided more than 50 responses per item for analysis. The items analyzed at each age level are identified in Table 1.

Table 1

Common item range composition by age

Age	Common Item Range	Number of Items
7 1/2	5-4 to 12-4	43
9 1/2	6-10 to 15-4	52
11 1/2	7-0 to 18-6	65
All Groups	5-4 to 18-6	75

To investigate the psychometric properties of the SIT for the British Columbia sample, the following five types of item analysis were conducted.

Item Difficulty

Item difficulty (p) values indicate the proportion of individuals who answer a dichotomously scored item correctly and reflect the extent to which items discriminate between individuals. Item difficulty values range from 0 to 1. Items which approach 0 increase in difficulty (fewer pass the item) while items which approach 1 decrease in difficulty (more pass the item). As item difficulties approach .5, the distribution of test scores becomes more normal, and standard deviation increases (Nunnally, 1978). Items in the middle range of difficulty (.25 to .75) are preferred for their potential to disperse test scores and enhance individual differences (Stanley & Hopkins, 1981).

Difficulty values influence the discriminating power of items which in turn influences test variance and internal consistency reliability (Stanley & Hopkins, 1981). As item difficulty values approach .5, item intercorrelations and internal consistency reliability is maximized. Items which fall at either end of the difficulty continuum fail to discriminate between individuals and add no reliable variance (Nunnally, 1978). Out of order items, in terms of increasing difficulty level, and item difficulties which do not maximize discrimination, are two common test problems (Bornstein, McLeod, McLurg & Hutchinson, 1983).

To compute the index of difficulty, correct responses were assigned a value of 1 and incorrect responses 0. Total test score was equal to the number of correct items in the common item range for the relevant age level.

Item Discrimination (Item-Test Correlation)

Item discrimination, sometimes referred to as item validity, is the correlation of an item with a criterion and is a form of the Pearson product-moment correlation coefficient. Point-biserial r (r_{pbis}) is the preferred product-moment correlation coefficient used to estimate the relationship between a dichotomously scored test item and a continuous variable, here total test score (Nunnally, 1978). Item discrimination values reflect how well an item discriminates between high and low scores on the overall test and range from -1.0 to +1.0.

Items that correlate highly with total test score (approach +1.0) increase the reliability of individual differences and the test's standard deviation (Jensen, 1980; Nunnally, 1978). Test reliability is greatest when item-total test score correlations fall above .30 (Nunnally, 1978).

Discrimination values are related to item difficulty and are maximized when item difficulty levels approach .5.

For the purpose of this item analysis, point biserial correlations were determined for each item in the common item range. Total test score for each age level was defined as the sum of the number of items answered correctly with credit for pre-basal items.

Comparison of Rank Order Item Difficulties

By arranging test items in order of increasing difficulty and using basal and ceiling rules as entry and exit points, tests such as the SIT can be administered to a wide age and ability range of individuals and yet be kept brief. The use of basal and ceiling rules involves the prediction that all pre-basal items would be passed and all post-ceiling items would be failed and requires the rank order by difficulty of test items to remain relatively constant across samples (Nunnally, 1978).

Items which differ in rank order between groups may be suspected of bias and may reflect different learning opportunities or changes in the cultural knowledge base over time (Jensen, 1982; Terman and Merrill, 1973). If internal criteria suggest the presence of bias, then the test's predictive validity may be biased for different cultural groups, and the presence of bias lowers the validity of the test as a whole. Items which maintain their same rank order placement across groups may be considered to be measuring the same ability (Jensen, 1982).

Two comparisons of the rank order of item difficulties were made between the British Columbia sample responses and the test presentation order. First, Spearman's rank order correlation coefficient between the two groups was obtained (Jensen, 1982). Spearman's r ranges from 0 (no rank association) to 1 (perfect correspondence in ranks). Correlations above .95 are desired and represent a very high degree of similarity in the order of item difficulty (Jensen, 1982). Second, rank order can also be evaluated in terms of change in item position between groups. For the purpose of this study, items were identified as misranked if rank order placement changed by more than ten positions between the two samples. A movement of more than ten rank positions was selected to take into account the SIT's basal and ceiling criterion of ten correct and incorrect responses in a row, respectively. Rank position of item difficulty was defined as location of placement in the test for the SIT items and by order of item difficulty values for the sample group.

Loevinger's Coefficient of Test Homogeneity

Loevinger's index of test homogeneity measures internal consistency or the degree to which items are ordered according to increasing difficulty (Loevinger, 1947; Cliff, 1979). The coefficient of test homogeneity is defined in terms of "the probabilities of passing successive items and probabilities of passing the easier of two items granted that the harder of the two is passed, for all pairs of items" (Loevinger, 1947, p.31). For a perfectly homogeneous test, in theory, every individual would pass all items up to a certain point and fail all subsequent items. A test departs from homogeneity when an individual passes an item(s) after a failure has occurred. The coefficient of homogeneity equals 1.0 for a perfectly homogeneous test and 0 for a perfectly heterogeneous test. Three implications may be drawn when a test of ability is perfectly homogeneous: (1) that all easier or earlier items have been passed when it is known that a harder or later item has been passed; (2) that all individuals with correct responses to an item have higher total test scores than those individuals who fail the item; and (3) that an individual who obtains a higher score on the test than another individual has more of the ability the test is measuring.

For the purpose of data analysis, the coefficient of SIT test homogeneity was determined for each age level for items responded to by 30 or more individuals within the age group. Item difficulty was based on each item's rank order placement in the test. Loevinger's coefficient of test homogeneity is capable of handling incomplete or tailored dichotomous response data, such as on the SIT. The computational formula for the coefficient of test homogeneity is given in Appendix B.

Loevinger's Coefficient of Item Pair Homogeneity

Loevinger's (1947) coefficient of item pair homogeneity identifies discrepancies in difficulty values between pairs of items. A discrepancy occurs when an individual passes the harder of two items but fails the easier when items are presumed ordered by increasing difficulty. The coefficient of homogeneity of an item pair is equal to 1 when the item pair is perfectly homogeneous and equal to 0 when the items are unrelated.

For the purpose of this analysis of the SIT, homogeneity coefficients were computed for all pairs of items answered by at least 100 out of the total 319 subjects administered the SIT. The coefficient takes into account chance expectancy. The formula for the coefficient of item pair homogeneity used in the present analysis is given in Appendix B.

In summary, this study examined the validity of the SIT, for use with schoolchildren in British Columbia, by means of assessing its internal psychometric properties according to the following five criteria:

1. item difficulty levels approach the desired value of .5 and fall within the range of .25 and .75;

2. item-total test score correlations (item discrimination indices) fall above .30, and approach the desired value of 1.0;

3. rank order correlation between the order of items by difficulty for the British Columbia sample of schoolchildren and the SIT test order of items (by difficulty level) falls at or above .95; items given to the British Columbia sample of schoolchildren do not differ by more than ten rank positions from SIT test item order;

4. test homogeneity of the SIT will approach perfect homogeneity of 1.0; and

5. item-pair homogeneities will be significantly positive and approach the "perfect homogeneity" value of one.

Chapter IV

Results

This chapter reports the results of five item analyses of the SIT for British Columbia schoolchildren: (1) analysis of item difficulty; (2) analysis of item discrimination; (3) comparison of rank order of item difficulty for the British Columbia sample and the norm sample; (4) test homogeneity; and (5) homogeneity of adjacent item pairs.

Analysis of Item Difficulty Indices

Item difficulty values (p) were determined for items falling within the common item range for each of the three age-groups. The obtained values are given for each age-group in Table 2. Examination of the values indicates that item difficulty indices ranged from .04 to 1.0 where actual item difficulty decreases as a value of 1 is approached. Table 2 suggests that, for the British Columbia sample, many items are not functioning effectively, as items with a difficulty value of .25 to .75 are desired and test variance and reliability is maximized when item difficulty approaches .5.

Table 2 also suggests that items do not consistently increase in difficulty over age, as desired for tests of ability employing basal and ceiling points and given to a broad age range of individuals. The interpretation may be drawn, therefore, that for this sample, some SIT items may be misplaced relative to their rank order of item difficulty.

Table 2

SIT Items: Frequency of Response, Index of Difficulty, Index of Discrimination, and Rank Order Misplacement by Difficulty for British Columbia Schoolchildren on the SIT

	Response Frequency (f)			Item Difficulty (p)			Item Discrimination (r_{pbis})			Rank Order Misplacement		
Age	7	9	11	7	9	11	7	9	11	7	9	11
N	108	111	100	108	111	100	108	111	100	108	111	100
Item Number												
5-4	61			1.0						0		
5-6	63			.98			.08			.5		
5-8	63			.98			.01			.5		
5-10	75			.93			.32			1.5		
6-0	78			.91			.37			3		
6-2	87			.95			.39			2		
6-4	91			.88			.44			2.5		
6-6	97			.88			.36			1.5		
6-8	97			.94			.26			4		
6-10	102	57		.76	.91		.38	.30		5	7	
7-0	106	90	53	.60	.90	.94	.39	.14	.24	9	7.5	6.5
7-2	106	90	53	.82	.98	.96	.49	.19	.37	.5	1	2
7-4	106	91	53	.93	.99	1.0	.17	.14		6.5	3	2
7-6	107	92	53	.60	.97	.92	.59	.21	.40	6	2	7
7-8	107	98	54	.82	.96	.96	.35	.16	.32	2.5	1.5	1
7-10	108	100	54	.43	.88	.94	.60	.47	.34	7	4.5	1.5
8-0	108	100	54	.81	.92	.93	.25	.20	.19	3	1	2.5
8-2	108	103	58	.33	.70	.79	.51	.52	.51	8	9	11
8-4	108	106	67	.59	.81	.88	.43	.21	.28	3	5	5
8-6	108	106	69	.60	.83	.8	.27	.32	.43	0	3	5
8-8	108	106	70	.13	.20	.26	.46	.35	.50	13	25	35.5
8-10	108	107	71	.83	.96	.99	.17	.19	.36	11	8.5	10
9-0	108	107	72	.75	.93	.96	.28	.08	.40	7	8	9
9-2	108	107	75	.62	.88	.93	.44	.15	.34	6	3.5	4.5
9-4	108	108	78	.18	.51	.82	.38	.53	.49	13	8.5	1
9-6	108	110	89	.24	.71	.91	.35	.42	.39	3	0	4
9-8	108	111	92	.73	.90	.95	.41	.33	.39	10	8.5	11
9-10	107	111	94	.06	.24	.61	.20	.39	.46	9.5	16.5	10
10-0	107	111	95	.03	.32	.61	.23	.51	.40	11.5	11.5	9
10-2	107	111	95	.27	.64	.67	.38	.41	.40	3	1	4
10-4	105	111	95	.39	.60	.81	.36	.32	.45	7	.5	4
10-6	105	111	95	.19	.55	.78	.45	.54	.51	1	0	.5
10-8	104	111	94	.22	.85	.89	.41	.33	.34	3	11	10
10-10	98	111	95	.34	.74	.78	.46	.27	.24	9	9	2.5
11-0	92	111	95	.15	.32	.59	.45	.35	.42	2	5.5	4.5
11-2	92	111	95	0	.08	.17		.20	.46	6.5	19	28
11-4	91	111	95	.25	.60	.64	.49	.28	.14	9	6.5	1

Table 2 Cont'd

	Response Frequency (f)			Item Difficulty (p)			Item Discrimination (r_{pbis})			Rank Order Misplacement		
Age	7	9	11	7	9	11	7	9	11	7	9	11
N	108	111	100	108	111	100	108	111	100	108	111	100
Item Number												
11-6	77	110	99	0	.08	.35		.27	.47	4.5	17	15.5
11-8	77	110	97	.03	.25	.52	.19	.37	.46	.5	4	4
11-10	77	110	97	.06	.36	.74	.09	.41	.42	2.5	2	7
12-0	74	110	97	.08	.34	.45	.30	.39	.34	6	2	6.5
12-2	60	108	97	.07	.50	.66	.23	.31	.38	6	7	7
12-4	55	106	98	.05	.38	.79	.27	.66	.58	4	6	14
12-6		106	96		.04	.16		.25	.43		14	11.5
12-8		102	96		.15	.35		.32	.35		35	8.5
12-10		101	96		.13	.23		.25	.23		5	13
13-0		99	96		.06	.20		.30	.51		10	15
13-2		91	95		0	.03			.18		13	24
13-4		91	95		.14	.13		.32	.27		1	14
13-6		91	95		.24	.43		.54	.54		5.5	.5
13-8		84	95		.65	.79		.61	.53		23	22
13-10		78	94		.18	.40		.31	.32		5	0
14-0		71	92		.15	.49		.48	.53		4.5	9
14-2		64	89		.09	.30		.41	.43		1	1
14-4		64	89		.28	.61		.49	.59		13	17
14-6		63	87		.02	.16		.16	.25		3.5	9.5
14-8		60	86		.12	.43		.45	.55		5	6.5
14-10		59	86		.08	.23		.27	.49		3	1
15-0		59	85		.41	.48		.55	.42		13	14
15-2		59	85		.02	.02		.09	.26		1	14
15-4		59	85		.51	.59		.54	.41		27.5	20.5
15-6			80			.45			.51			14.5
15-8			79			.47			.43			17
15-10			78			.54			.53			22
16-0			78			.44			.51			16
16-3			77			.23			.40			7
16-6			76			.01			.00			8
16-9			75			.21			.21			7
17-0			75			.09			.13			.5
17-3			73			.10			.32			3
17-6			73			.03			.18			1
17-9			69			.09			.26			4.5
18-0			66			.03			.30			1
18-3			62			.26			.37			17.5
18-6			54			.04			.15			5

The percentage of items which fall below, within or above the desired difficulty range of .25 to .75 is given in Table 3. Table 3 indicates that approximately one-third of the items are too easy and one-third too difficult. Therefore, for this sample of British Columbia schoolchildren, two-thirds of the items do not work effectively to maximize test reliability.

Comparison of item difficulties across the three age levels shows that, as expected, test items decrease in difficulty over age. This finding suggests that the items are functioning as desired between age levels; however, the two-year age differences between the groups tested weakens the significance of this finding.

Table 3

Percentage of items falling below, within and above the preferred range of difficulty for British Columbia children at three age levels

	Item Difficulty (p) Range			
Age	(p)<.25	(p)=.25 to .75	(p)>.75	Number of Items
7 1/2	35	30	35	43
9 1/2	34.5	36.5	29	52
11 1/2	28	38	34	65

Analysis of Item Discrimination Indices

Item-test correlation coefficients were computed for items falling in the common range for each of the three age levels. The obtained point-biserial correlation coefficients (r_{pbis}) are given in Table 2.

Correlations range from .00 to .66 where correlations of .30 and above are accepted as contributing to test reliability. These findings suggest that of items in the common item range, 35%, 39% and 20% (n = 43, 52, 65) are not discriminating well at the 7 1/2, 9 1/2, and 11 1/2 year old age levels, respectively, for this sample.

Comparison of Rank Order of Item Difficulty

The rank order of item difficulty for the British Columbia sample of schoolchildren was compared to item presentation order. Spearman's rank order correlation coefficients were computed for each of the three age levels over the common item range. Correlation values of .88, .79, and .81 were obtained for the 7 1/2, 9 1/2 and 11 1/2 year olds, respectively, indicating a degree of similarity in the rank ordering of items by difficulty for the two samples.

Degree of difference in the rank order placement of items for the two samples was determined for each age level and presented in Table 2. Rank order changes in difficulty ranged from 0 to 35. Items were found both to be easier and harder for the British Columbia sample relative to their placement in the test. The items which changed more than 10 positions in terms of rank order of difficulty are listed by age level in Table 4.

Table 4

Items changing rank order position by more than ten places from the order of presentation for the British Columbia sample

	Age Level	
7 1/2	9 1/2	11 1/2
8-8**	8-8**	8-8**
8-10	9-10	9-8
9-4	10-0*	11-2*
10-0*	10-8	11-6*
	11-2*	12-4
	11-6*	12-6*
	12-6*	12-10
	13-2*	13-0
	13-8*	13-2*
	14-4*	13-4
	15-0*	13-8*
	15-4*	14-4*
		15-0*
		15-2
		15-4*
		15-6
		15-8
		15-10
		16-0
		18-3

* change in rank position by more than ten positions at two age levels

** change in rank position by more than ten positions at three age levels

Twenty-five items were found to have shifted more than ten positions. Table 5 presents these items by content and direction of rank position movement. Items were classified as information, similarities, short-term memory, arithmetic-reasoning, arithmetic-information, vocabulary or visual-motor. Categories were based on previously developed schemes with the exception of numerical reasoning which was broken down into the categories of arithmetic-reasoning and arithmetic-information on the basis of item content, and digit span and/or sentence memory which were combined into the category of short-term memory (Boyd, 1974; Fudala, 1979; Nicholson, 1970; Stone, 1975). Classification across schemes was generally consistent. The majority of discrepancies which did occur was over categorization of items as information versus arithmetic. This was corrected for in the present study by the inclusion of both an arithmetic-reasoning and arithmetic-information category. The number and percentage of items by category are given in Appendix C for each of four previously developed item classification schemas. The classification of items falling in the common item range and the number and percentage of items per category which changed relative difficulty are also presented in Appendix C. Test item content is not consistent over age.

Of the questions which were easier for the British Columbia sample, 8 out of 12 were classified as vocabulary items and 7 out of 13 of the harder items were arithmetic-information. Over all SIT items analyzed, 46% of the vocabulary items and 100% of the arithmetic-information items changed rank position by more than 10 places.

Table 5

Categorization of items changing rank position of difficulty by more than ten positions

Items Less Difficult than Test Placement

Item	Age	Question	Category
8-10	7	What does destroy mean?	Vocabulary
9-8	11	What was a dungeon used for?	Vocabulary
12-4	11	What does scarce mean?	Vocabulary
13-8	9, 11	What does tremendous mean?	Vocabulary
14-4	9, 11	What is the principal kind of work done by a pharmacist?	Vocabulary
15-0	9, 11	What is the principal kind of work done by an architect?	Vocabulary
15-4	9, 11	What does fragrant mean?	Vocabulary
15-6	11	What is the area or how many square feet are there in a room 9' wide and 12' long?	Arithmetic Reasoning
15-8	11	In what ways are an octopus and an octave alike?	Similarity
15-10	11	What does environment mean?	Vocabulary
16-0	11	A boy who had $5.00 took his girl to the movies. If the tickets cost $.75 each, and they both had $.30 milkshakes, how much did he have left?	Arithmetic Reasoning
18-3	11	Say these numbers backwards: '8 3 2 9 4 7'	Short-term Memory

Table 5 Cont'd

Items More Difficult than Test Placement			
Item	Age	Question	Category
8-8	7, 9, 11	Listen carefully and say exactly what I say "The train goes fast on the tracks carrying people and bags of mail"	Short-term Memory
9-4	7	What does vacant mean?	Vocabulary
9-10	9	How many inches in two feet?	Arith-Info
10-0	7	How many minutes in 3/4 of an hour?	Arith-Info
10-8	9	If a boy had 45 cents, how many nickel or 5 cent candy bars could he buy?	Arithmetic Reasoning
11-2	9, 11	What does it mean to be thrifty?	Vocabulary
11-6	9, 11	What would a man do if he took an inventory of his store?	Vocabulary
12-6	9, 11	How many inches in two yards?	Arith-Info
12-10	11	What should be a healthy person's temperature?	Information
13-0	11	How many feet in thirteen yards?	Arith-Info
13-2	9, 11	How many pints in a gallon?	Arith-Info
13-4	11	How many pounds in a ton?	Arith-Info
15-2	11	How many feet in a mile?	Arith-Info

Test Homogeneity

Loevinger's coefficient of test homogeneity was computed for the SIT items. At the 7 1/2, 9 1/2 and 11 1/2 age group, the obtained coefficients of test homogeneity were equal to .003, .004, and .006, respectively.

These low coefficients suggest that items on the SIT are not arranged according to order of difficulty for the sample of children tested.

Homogeneity of Pairs of Test Items

Loevinger's coefficient of item-test homogeneity was determined for item pairs responded to by one hundred or more of the children tested, grouped over age. Coefficients ranged from -.22 to .88, where items which are perfectly homogenous have a coefficient of 1.0 and items which are unrelated approach 0, and are listed in Appendix D. The coefficient median fell at .21 and 75% of the responses fell in the range of .06 to .44. The large number of low coefficients reflect discrepancies in item pairs where the harder item is passed and the easier failed, suggesting that items on the SIT are not in order of increasing difficulty for the British Columbia sample.

Summary

The results of the analyses suggest that SIT items are not working as desired for the British Columbia sample of schoolchildren tested. Two-thirds of the items in the common item range were not functioning effectively in terms of item difficulty, one-third of the items were not discriminating well, and many items were misordered according to increasing level of difficulty for the sample tested. The item weaknesses identified are recognized to lower a test's overall internal consistency reliability and, consequently, its criterion validity.

Chapter V

Summary and Conclusions

This chapter includes a summary of the findings of the item analyses of the SIT and a discussion of them in relation to the effectiveness of the SIT for use with British Columbia schoolchildren.

Purpose of the Study

The purpose of the present study was to examine the effectiveness of the SIT as a measure of general intelligence for British Columbia schoolchildren through analysis of the internal psychometric properties of the test items. Test item's psychometric properties affect the distribution of test scores, internal consistency reliability and criterion validity. Five item characteristics which influence test interpretation were examined: item difficulty, item discrimination, rank order of item difficulty correlations between the British Columbia sample and the standardization group, test homogeneity, and homogeneity of item pairs. For a test to discriminate most effectively, item difficulty values should be in the range of .25 to .75, item discrimination values should fall above .30, and items should be arranged by increasing level of difficulty.

Summary of Test Findings

The item difficulty values obtained for the British Columbia sample indicate that less than one-third of the items, at each of the three age levels tested, achieved the desired range of difficulty of .25 to .75. This suggests that, for this sample, approximately two-thirds of the items are failing to discriminate between individuals on the trait measured by the test. Items falling outside the optimal difficulty range do not raise test variance nor lower internal consistency reliability. For tests of general intelligence, such as the SIT, a wide dispersion of test scores is desirable in order to maximize discrimination between individuals.

For the British Columbia sample, approximately one-third of the item discrimination values were found to be too low to discriminate effectively between high and low scorers on the test as a whole for each of the age levels assessed. Items which are good discriminators are passed by individuals with higher test scores than those who fail the item. Items which are poor discriminators diminish the reliability of individual differences and the test's internal consistency reliability.

The SIT's incorporation of basal and ceiling entry and exit points assumes that items are presented in order of increasing difficulty. Analysis of the item difficulty indices showed that items falling in the common item range do not consistently increase in difficulty within each age-group for the British Columbia sample. A consistent decrease in difficulty over age was noted. However, the two year gap between age-groups measured reduces the significance of the finding.

Spearman's rank order correlation coefficient was used to compare the relative rank order of item difficulty for the British Columbia sample and the order of item presentation. If a test is measuring the same ability across groups, the relative rank order of the items should not vary. The obtained rank order correlations of .79 to .88 between the norm and the British Columbia sample for the three age-groups is respectable but falls short of the desired value of .95.

To further examine which items varied by difficulty between groups, items which changed rank order placement by more than ten positions were identified. This criterion was chosen to take into account the SIT's basal and ceiling rules. Rank order position changes ranged from 0 to 35, and twenty-five items were found to have shifted by more than ten positions. Analysis of items by content suggested that different types of items were easier or harder for the British Columbia sample than the standardized group. Of the easier items, eight were classified under vocabulary. These items called for the meaning of destroy, dungeon, scarce, tremendous, pharmacist, architect, fragrant and environment. Of the 13 harder items, seven involved non-metric arithmetic knowledge. These items included converting feet to inches, yards to inches, gallons to pints, tons to pounds and miles to feet. Three vocabulary words, thrifty, vacant and inventory, were also more difficult for the British Columbia sample.

Shifts in difficulty between groups may reflect different learning opportunities or cultural experiences, or may be an artifact of changes in the common knowledge base which occur over time (Jensen, 1982). Examination of the content of items which significantly altered in difficulty can be attributed either to cultural differences or changes which occur over time (time-factor). The finding that 46% of the vocabulary items in the common item range significantly changed in difficulty, becoming either easier or harder supports the time-factor interpretation as vocabulary use is recognized to alter over time. For example, the words "tremendous" and "environment" are more commonly used today than twenty years ago. A cultural or educational difference interpretation of changes in vocabulary difficulty is less supported since

British Columbia children have exposure to a similar language base (through TV and print media) as their American counterparts. The role of cultural factors in the variation of the psychometric properties between the two groups is suggested by the increase in difficulty of the non-metric arithmetic problems for the British Columbia sample, although this interpretation is not clear-cut because the age of the test is a confounding factor. A cultural interpretation may be drawn because the metric system was adopted in Canada in the early 1970's and children are less familiar with non-metric arithmetic values than at the time the SIT was developed. Therefore, the increase in difficulty of the arithmetic problems may be related to the change in the math system taught to British Columbia schoolchildren. However, as twenty years have elapsed since the standardization of the SIT, the age of the test confounds the interpretation of this finding. An analysis of item difficulty for a comparable present-day American sample is needed to determine whether the change in item difficulty may be attributable to cultural differences or to the test's age.

Loevinger's coefficient of test homogeneity was also used to evaluate the degree to which items on the SIT were ordered by increasing difficulty for the British Columbia sample. When a test is perfectly homogeneous, an individual's total test score reflects his or her ability relative to the trait measured by the test, and higher test scores can be interpreted in terms of greater ability. The obtained coefficients of test homogeneity approached zero, indicating that for this sample, test items are not ordered in terms of increasing difficulty.

Lack of homogeneity is also reflected in the low adjacent item pair homogeneity coefficients. For this sample of British Columbia schoolchildren, the median fell at .21 and 75% of the responses fell in the range of .06 to .44, where a coefficient of 0 suggests that the pair of items are unrelated in regard to the ability that they measure.

Limitations of the Study

The findings of the present study must be interpreted relative to the following limitations on its generalizability. First, all subjects in this study resided within the province of British Columbia and may not, therefore, be representative of children in other provinces of Canada or the United States. Second, the sample selected was limited to three age groups, 7 1/2, 9 1/2, and 11 1/2 (corresponding to grades 2, 4, and 6), and is therefore not representative of the total age-range of individuals to whom the SIT is administered. Third, all subjects were drawn from regular class placements and findings may not, therefore, be generalizable to children in special education programs. The sample was, however, carefully selected to be representative of the British Columbia population for the age groups tested and the findings, therefore, may be generalizable to those age groups within the British Columbia population and may possibly be extended to the British Columbia population of schoolchildren at large.

It must be noted that lack of an American comparison group limits the interpretation of this study's findings. Data from an American comparison group might provide insight as to whether the shift in difficulty values noted reflect the limited nature and age of the norm data or cultural differences. For example, if the results of an American comparison group matched the findings of the norm data or they matched on all but the arithmetic-information items, then a cultural difference interpretation would be favoured. If the variations are a factor of the age of the test, then similar results would be expected for an American comparison sample as those found for the British Columbia sample.

The degree to which this sample of British Columbia children is or is not representative of Canadian and American children's performance on the Slosson Intelligence Test limits the generalizability of the study's findings. Although cultural influences cannot be separated from the technical weakness and age of the SIT item analysis data without obtaining comparative information for a present-day sample of American schoolchildren, the information provided by these item analyses should be considered useful for interpretation of the SIT with British Columbia schoolchildren.

Conclusion

Technical weaknesses evident in the SIT manuals (1977, 1983) suggested a need to examine the psychometric properties of the test when administered to British Columbia schoolchildren. Analyses indicated that a significant proportion of test items are not working to discriminate between individuals on the trait measured by the test. Additionally, items were not consistently ordered by their increasing level of difficulty. Misranking of items can have two results. One, children of younger ages, who would gain credit for knowing the item if it were administered, receive no credit for it if the item comes after the child's ceiling point: it would not be administered. Two, the limit (ceiling) is extended upward for children who pass a misplaced, easy item. In the most extreme case, a child may correctly answer the easy (misplaced) item after nine consecutive failures; he must then be given additional questions until ten consecutive items are failed. This could amount to 19 wrong responses in 20 items. This pattern was noted several times during data collection. Repetitive failure is undesirable during test administration, as it can have deleterious effects on the child and invalidate further testing.

Two possible factors have been presented in explanation of the noted differences in item difficulty and discrimination values for the British Columbia sample and the standardization group: (1) cultural or educational bias resulting from different learning experiences encountered by the two groups, or (2) changes in the content knowledge base of the general population.

An alternative explanation of the differences in obtained item difficulty values is related to the limited nature of the original norm sample. The standardization data was collected solely from New England residents, and therefore may not be representative of the United States at large. Discrepant difficulty values would perhaps not have been found if the norm sample had been more representative. As there is no way of assessing what item difficulty indices would have been in a well-stratified sample collected at the time of the limited norm data, this alternative cannot be evaluated empirically. However, analysis of a well-stratified present-day American sample would shed light on this issue. Evidence which indicates that item difficulties do change over time weakens support for this interpretation (Terman & Merrill, 1973; Wechsler, 1974; Dunn & Dunn, 1981). The possibility that testing a present-day American sample would support a cultural difference interpretation of the shift in difficulty on the arithmetic items also cannot be dismissed without getting empirical evidence.

In conclusion, it is hypothesized that the differences in item difficulty between the British Columbia sample and the norm sample is a function of the age of SIT norms. Nearly a quarter of a century has passed since the Slosson Intelligence Test was constructed and the test's items have not been re-evaluated or updated, despite the appearance of a 1981 revision of norms. To support this hypothesis, further research is needed to gather comparative item analysis data for a present-day American sample of children. Until such data is collected, it is not possible to determine the factors contributing to the differences in item difficulty. This information is not, however, necessary for caution to be drawn against use of the SIT with British Columbia schoolchildren.

References

Armstrong, R.J. & Jensen, J.A. (1972). The validity of an abbreviated form of the Stanford-Binet Intelligence Scale, Form L-M. Educational and Psychological Measurement, 32, 463-467.

Armstrong, R.J. & Jensen, J.A. (1982). Slosson Intelligence Test (SIT) for Children and Adults. Technical Manual. East Aurora, NY: Slosson Educational Publications.

Armstrong, R.J. & Jensen, J.A. (1984). Slosson Intelligence Test (SIT) for Children and Adults. Norms tables: Application and Development. East Aurora, NY: Slosson Educational Publications.

Armstrong, R.J. & Mooney, R.F. (1971). The Slosson Intelligence Test: Implications for reading specialists. Reading Teacher, 24, 336-340.

Baum, D.D. & Abelson, G. (1981). Comparison of Slosson and Peabody IQs among white kindergarten children. Psychological Reports, 48, 754.

Baum, D.D. & Kelly, T.J. (1979). The validity of the Slosson Intelligence Test with learning disabled Kindergartners. Journal of Learning Disabilities, 12, 268-270.

Berry, G.M. Jr. (1977). An investigation of the item ordering of the Peabody Picture Vocabulary Test by sex and race. (Doctoral dissertation, University of Connecticut, 1977). Dissertation Abstracts International, 38, 6642A.

Borstein, R.A., McLeod, J., McClurg, E. & Hutchinson, B. (1983). Item difficulty and content bias on the WAIS-R information subtest. Canadian Journal of Behavioral Science, 15, 27-34.

Boyd, J.E. (1974). Use of the Slosson Intelligence Test in Reading Diagnosis. Academic Therapy, 9, 441-444.

Brown, W.R. & McGuire, J.M. (1976). Current psychological assessment practices. Professional Psychology, 7, 475-484.

Carlisle, A., Shinedling, M. & Weaver, R. (1970). Note on the use of the Slosson Intelligence Test with mentally retarded residents. Psychological Reports, 26, 865-866.

Cianflone, R. & Zullo, T.O. (1980). The relationship of an early measure of intelligence to the ability to learn sight vocabulary words and to later achievement. Educational and Psychological Measurement, 40, 1197-1200.

Cliff, N. (1979). Test theory without true scores. Psychometrika, 44, 373-391.

Colarusso, R., McLesky, J., Gill, S.H. (1977). Use of the Peabody Picture Vocabulary Test and the Slosson Intelligence Test with urban black kindergarten children. Journal of Special Education, 41(1), 19-23.

Coleman, M., Brown, G. & Genong, L. (1980). A comparison of PPVT and SIT scores of young children. Psychology in the Schools, 17, 178-180.

Covin, T.M. (1977a). Comparison of SIT and WISC-R IQ's among special education candidates. Psychology in the Schools, 14, 19-23.

Covin, T.M. (1977b). Relationship of the SIT and PPVT to the WISC-R. Journal of School Psychology, 15, 259-260.

Crofoot, M.J. & Bennett, T.S. (1980). A comparison of the screening tests and the WISC-R in special education evaluations. Psychology in the Schools, 17, 474-478.

DeLapa, G. (1973). Correlates of Slosson Intelligence Test, Stanford-Binet Form L-M, and achievement indices. (Doctoral dissertation, West Virginia University, 1967). Dissertations Abstracts International, 3498-A. (University Microfilms no. 68-2678).

Dirkes, J., Wessels, K., Quarforth, J. & Quervon, B. (1980). Can short-form WISC-R IQ tests identify children with high full scale IQ? Psychology in the Schools, 17, 40-41.

Dunn, L.M. and Dunn, L.M. (1981). Peabody Picture Vocabulary Test-Revised. Circle Pines, MN:American Guidance Service.

Fudala, J.B. (1979). Differential evaluation of students with the SIT. Academic Therapy, 15, 61-64.

Grossman, F.M. & Johnson, K.M. (1983). Validity of the Slosson and Otis-Lennon in predicting achievement of gifted students. Educational and Psychological Measurement, 43, 617-622.

Hale, R.L., Douglas, B., Cummins, A., Rittgarn, G., Breeds, B. & Dabbert, D. (1978). The Slosson as a predictor of Wide Range Achievement Test performance. Psychology in the Schools, 15, 507-509.

Holmes, B.J. (1981). Individually-administered intelligence tests: An application of anchor test norming and equating procedures in British Columbia. (Doctoral dissertation, University of British Columbia, 1981). Dissertation Abstracts International, 42, 2626A.

Houston, C. & Otto, W. (1968). Poor readers' functioning on the WISC, Slosson Intelligence Test and Quick Test. Journal of Educational Research, 62, 157-159.

Jensen, A.R. (1980). Bias in Mental Testing. New York: The Free Press.

Jensen, A.R. (1982). Straight Talk about Mental Tests. New York: The Free Press.

Jerrolds, B.W., Callaway, B., & Gwaltney, W.K. (1972). Comparison of the Slosson Intelligence Test and WISC scores of subjects referred to a reading clinic. Psychology in the Schools, 9, 409-410.

Johnson, D.L. & Johnson, C.A. (1971). Comparison of four intelligence tests used with culturally disadvantaged children. Psychological Reports, 28, 209-210.

Jongeward, P.A. (1968). A validity study of the Slosson Intelligence Test for use with educable mentally retarded students. Journal of School Psychology, 7, 59-63.

Kaufman, H. & Ivanoff, J. (1969). The Slosson Intelligence Test as a screening instrument with a rehabilitation population. Exceptional Children, 35, 745.

Klein, A.E. (1978). The reliability and predictive validity of the Slosson Intelligence Test for pre-kindergarten pupils. Educational and Psychological Measurement, 38, 1211-1217.

Lamp, R.E., Traxler, A.J. & Gustafson, P.P. (1973). Predicting academic achievement of disadvantaged fourth grade children using the Slosson Intelligence Test. Journal of Community Psychology, 1, 339-341.

Lessler, K. & Galinksy, M.D. (1971). Relationship between the Slosson Intelligence Test and Wechsler Intelligence Scale for Children scores in special education candidates. Psychology in the Schools, 8, 341-344.

Loevinger, J. (1947). A systematic approach to the construction and evaluation of tests of ability. Psychological Monographs, 61, No.4.

Lowrance, D. & Anderson, H.N. (1979). A comparison of the Slosson Intelligence Test and the WISC-R with elementary school children. Psychology in the Schools, 16, 361-364.

Martin, J.D., Blair, G.E. & Vickers, D.M. (1979). Correlation of the Slosson Intelligence Test with the California Short-Form Test of Mental Maturity and the Shipley-Institute of Living Scale. Educational and Psychological Measurement, 39, 193-196.

Martin, J.D. & Rudolph, L. (1972). Correlates of the Wechsler Adult Intelligence Scale, the Slosson Intelligence Test, ACT scores and grade point averages. Educational and Psychological Measurement, 32, 459-462.

Maxwell, M.T. (1971). The relationship between the Wechsler Intelligence Scale for Children and the Slosson Intelligence Test. Child Study Journal, 1, 164-171.

Mize, J.M., Calloway, B. & Smith, J.W. (1979). Comparison of reading disabled children's scores on the WISC-R, Peabody Picture Vocabulary Test and Slosson Intelligence Test. Psychology in the Schools, 16, 356-358.

Nicholson, C.I. (1970). Analysis of functions of the Slosson Intelligence Test. Perceptual and Motor Skills, 31, 627-631.

Nunnally, J.C. (1978). Psychometric Theory. New York: McGraw Hill.

Ritter, D., Duffey, J. & Fischman, R. (1973). Comparability of Slosson and Stanford-Binet estimates of intelligence. Journal of School Psychology, 11, 224-227.

Rogers, S.J. (1982). Problems with the Slosson Intelligence Test for preschool children. Journal of School Psychology, 20(1), 65-68.

Rotatori, A.F., Sedlak, B. & Freagon, S. (1979). Usefulness of the Slosson Intelligence Test with severely and profoundly retarded children. Perceptual and Motor Skills, 48, 334.

Rust, J.O. & Lose, B.D. (1980). Screening for giftedness with the Slosson and the scale of rating behavioral characteristics of superior students. Psychology in the Schools, 17, 446-451.

Slosson, R.L. (1977). Slosson Intelligence Test (SIT) for Children and Adults. East Aurora, New York: Slosson Educational Publishers.

Slosson, R.L. (1983). Slosson Intelligence Test (SIT) for children and adults (2nd ed.). East Aurora, New York: Slosson Educational Publishers.

Smith, S.A. (1981). Slosson and Peabody IQ's of mentally retarded adults. Psychological Reports, 48, 786.

Stanley, J.C. & Hopkins, K.D. (1981). Educational and psychological measurement and evaluation (6th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Stewart, K.D. & Jones, E.C. (1976). Validity of the SIT. A ten-year review. Psychology in the Schools, 13, 372-380.

Stewart, K.D. & Myers, D.G. (1974). Long term validity of the Slosson Intelligence Test. Journal of Clinical Psychology, 30, 180-181.

Stewart, K.D. & Wood, D.Z. & Gallman, W.A. (1971). Concurrent validity of Slosson Intelligence Test. Journal of Clinical Psychology, 27, 218-220.

Stone, M. (1975). An interpretive profile for the Slosson Intelligence Test. Psychology in the Schools, 12, 330-333.

Swanson, M.S. & Jacobson, A. (1970). Evaluation of the Slosson Intelligence Test for screening children with learning disabilities. Journal of Learning Disabilities, 3, 318-320.

Terman, L.M. & Merrill, M.A. (1973). Stanford-Binet Intelligence Scale (3rd ed.). New York: Houghton-Mifflin.

Vance, H.B., Lewis, R. & DeBell, S. (1979). Correlations of the Wechsler Intelligence Scale for Children - Revised, Peabody Picture Vocabulary Test, and Slosson Intelligence Test for a group of learning disabled students. Psychological Reports, 44, 735-738.

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

Appendix A

Breakdown of Sample by Stratification Variables

Appendix A

Breakdown of Sample by Stratification Variables

Variable	Stratification	N
Sex	Male	163
	Female	156
Age	7 years 3 months to 7 years 9 months	108
	9 years 3 months to 9 years 9 months	111
	11 years 3 months to 11 years 9 months	100
Community Size	Under 1000	45
	1000 to 50,000	80
	Over 50,000	192
School Size	Under 150	47
	151 to 300	80
	Over 300	192
Zone	Okanagan	43
	Metropolitan Vancouver	133
	Fraser Valley	26
	Vancouver Island	48
	Kootenay	26
	Northern British Columbia	43

Appendix B

Computational Formulas for Determining Test Homogeneity and Item Pair Homogeneity

Appendix B

Computational Formula for Determining Test Homogeneity (Loevinger, 1947)

$$\text{Est } H_t = \frac{N(X_k{}^2 - X_k) + N_1{}^2 - (X_k)^2}{2N(E_i N_i - X_k) + N_i{}^2 - (X_k)^2}$$

Where:

H_t = coefficient of test homogeneity

X = raw score

N_i = number passing the *i*th item, when items are ordered according to decreasing number passing

k = summation for all *N* individuals

i = summation for all *m* items

Computational Formula for Determining the Coefficient of Homogeneity of Item Pairs (Loevinger, 1947)

$$H_{ii} = 1 - \frac{NK}{P_h O_e}$$

Where:

H_{ii} = the coefficient of homogeneity of two items

N = number of cases

P_h = number passing the harder item

O_e = number failing the easier item

K = number passing the harder and failing the easier item

Appendix C

Item Categorization Schemas of the SIT

Item Categorization Schemas of the SIT

	Nicholson N = 149 (2:0-27:0)		Boyd N = 65 (4:8-15:10)		Stone N = 122 (2:0-20)		Fudala N = 159 (2:1-27:0)	
Category	N	%	N	%	N	%	N	%
Information	26	17	15	23	35	29	26	16
Similarities	26	17	12	18	17	14	20	13
Short-Term Memory	13	9	6	9	12	10	11	7
Arithmetic	31	21	15	23	26	22	36	23
Vocabulary	49	33	17	26	28	23	61	38
Visual-Motor	4	3			4	3	3	2

SIT Items by Category in the Common Item Range

Category	N	Too Easy	Too Hard	Sum	% of Total
Information	12	1		1	8
Similarities	11		1	1	9
Short-Term Memory	7	1	1	2	29
Arithmetic-Reasoning	13	1	2	3	23
Arithmetic-Information	7	7		7	100
Vocabulary	24	3	8	11	46
Sum	75	13	12	25	33

Appendix D

Coefficients of Homogeneity of Item Pairs

Appendix D

Coefficients of Homogeneity of Item Pairs (N>100)

Passed	Failed	H_{ii}	N	Passed	Failed	H_{ii}	N
5-10	5-8			11-6	11-4	-.11	282
6-0	5-10	.10	100	11-8	11-6	.21	284
6-2	6-0	.26	107	11-10	11-8	.50	284
6-4	6-2	.44	121	12-0	11-10	.25	281
6-6	6-4	.35	134	12-2	12-0	.15	265
6-8	6-6	.11	160	12-4	12-2	.23	258
6-10	6-8	.60	171	12-6	12-4	.88	254
7-0	6-10	.12	185	12-8	12-6	.18	234
7-2	7-0	.14	249	12-10	12-8	.24	228
7-4	7-2	.06	249	13-0	12-10	.24	226
7-6	7-4	.21	250	13-2	13-0	.62	201
7-8	7-6	.19	252	13-4	13-2	.002	201
7-10	7-8	.43	259	13-6	13-4	.05	199
8-0	7-10	.07	262	13-8	13-6	.15	190
8-2	8-0	.22	262	13-10	13-8	.30	179
8-4	8-2	.14	270	14-0	13-10	.09	168
8-6	8-4	.17	282	14-2	14-0	.25	158
8-8	8-6	.47	284	14-4	14-2	.21	158
8-10	8-8	.01	282	14-6	14-4	.38	155
9-0	8-10	.38	287	14-8	14-6	.07	151
9-2	9-0	.16	286	14-10	14-8	.43	150
9-4	9-2	.56	290	15-0	14-10	.06	149
9-6	9-4	.30	294	15-2	15-0	-.18	149
9-8	9-6	.12	307	15-4	15-2	.004	149
9-10	9-8	.52	310	15-6	15-4	.22	131
10-0	9-10	.39	312	15-8	15-6	.35	126
10-2	10-0	.13	313	15-10	15-8	.25	125
10-4	10-2	.14	311	16-0	15-10	.36	124
10-6	10-4	.44	311	16-3	16-0	.26	122
10-8	10-6	.25	309	16-6	16-3	-.22	121
10-10	10-8	.44	303	16-9	16-6	.05	120
11-0	10-10	.23	298	17-0	16-9	-.16	117
11-2	11-0	.32	298	17-3	17-0	-.08	110
11-4	11-2	.03	297	17-6	17-3	.00	109