5-1995

An Analysis of Item Bias in the WISC-R with Kainaiwa Native Canadian Children

Deborah Faith Pace
Utah State University

AN ANALYSIS OF ITEM BIAS IN THE WISC-R WITH KAINAIWA NATIVE CANADIAN CHILDREN

by

Deborah Faith Pace

A thesis submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

Psychology

James P. Shaver, Ed.D.
Dean of Graduate Studies

UTAH STATE UNIVERSITY
Logan, Utah

1995

ABSTRACT

An Analysis of Item Bias in the WISC-R with

Kainaiwa Native Canadian Children

by

Deborah Faith Pace, Master of Science

Utah State University, 1995

Major Professor: Dr. Glendon Casto
Department: Psychology

The present study examined the responses of 332 Kainai students ranging in age from 6 to 16 years to the Information, Arithmetic, and Picture Completion subtests of the Wechsler Intelligence Scale for Children–Revised (WISC-R) in order to determine the validity of these subtests as a measure of their intelligence. Two indices of validity were assessed: (a) subtest unidimensionality, and (b) order of item difficulty. With regard to the assumption of unidimensionality, examination of the data indicated low item-factor loadings on the Information, Arithmetic, and Picture Completion subtests. Examination of difficulty parameters revealed a nonlinear item difficulty order on all three subtests.

These results support the conclusion of previous research that the WISC-R does not adequately assess the intelligence of Native children. Possible bases for

the invalidity of the WISC-R for this population are discussed and recommendations for future research are presented.

(47 pages)

ACKNOWLEDGMENTS

I would like to express my sincere appreciation to my family, especially my daughter, Tanya, my mother, Mary, and my granddaughter, Chloe, and my dad, Wallace, who were all patient and supportive in believing my dream. I am particularly grateful to my parents who have taught me well in a loving environment.

I would like to thank Dr. Glendon Casto, who chaired my thesis committee. His guidance and support enabled me to complete my thesis. I would also like to thank Drs. Bertoch, Barcus, and Fifield for their support and guidance.

I am grateful for all my relations who offered their spiritual support in helping me complete my studies, especially Margaret Hindman.

Finally, I would like to acknowledge two of my closest friends and colleagues, Dr. Roland Chrisjohn and Dr. Shelagh Towson, for believing in me and encouraging me to complete my thesis.

To all my friends, colleagues, and family–thank you.

Deborah Faith Pace

CONTENTS

LIST OF TABLES

CHAPTER I

INTRODUCTION

Problem Statement

Within the educational system, the assessment of intelligence depends to a great extent on standardized intelligence tests. The most widely used of these tests, accepted as both valid and reliable for all North American children regardless of their ethnic background, is the Wechsler Intelligence Scale for Children–Revised (WISC-R) (Wechsler, 1974).

However, a number of researchers in the educational field have claimed bias in standardized assessments and IQs obtained by minority children, in particular the Native American children (Chrisjohn, Towson, Pace, & Peters, 1988; Mishra, 1982; Mueller, Mulcahy, Wilgosh, Watters, & Mancini, 1986; Reynolds & Reschly, 1983). These past studies conducted examined item bias in various subtests of the WISC-R which accounted for depressed scores in both verbal and performance scales. St. John and Kricher (1976) have suggested that the WISC-R is culturally biased and that reliance on the WISC-R results in the misclassification of Native children. They believe that the test is a failure in identifying gifted Native children and misidentifying Native children of average intelligence as either being intellectually deficient or having specific learning disabilities (Wilgosh, Mulcahy, & Watters, 1986).

Children whose pattern of scores on the various WISC-R subtests is atypical are often classified as learning disabled, especially if the Performance IQ

is notably higher than the Verbal IQ (Kaufman, 1979). If there is a 1.5 standard deviation or 23-point difference in the discrepancy between achievement tests and intelligence tests (Lerner, 1981), as a result, a student may receive some kind of special education as a result. The educational objectives of the school system are based on one culture, but the lifestyles, values, and goals of students attending it come from different cultural backgrounds (Common & Frost, 1988).

Other researchers have denied the existence of bias in the WISC-R. Sattler (1988) argued that intelligence tests are not systematically content biased to favor one group or another. On the basis of his review of the literature, Jensen (1980) also concluded there is no evidence of internal bias in standardized tests of mental ability according to his review of existing studies. Gordon and Rudert (1979) made a strong argument that IQ tests are not culturally biased and found that "race-by-item" interactions with the analysis of variance (ANOVA) method is sufficient to detect questionable items when they are present.

However, this may not be the best method for detecting item bias. Analysis of variance (ANOVA) is one of the predominant methods for detecting bias in internal analysis of test instruments. The ANOVA indication of bias is a significant group-by-item interaction. However, Camilli and Sheppard (1987) have suggested that ANOVA is inadequate for detecting internal test bias. For instance, even though the ANOVA generates group-by-group interactions for methods of comparisons, it is not able to detect bias that adds or subtracts from the true score of an individual.

A more promising approach for assessing bias in WISC-R items is based on latent trait theory and item response theory (Hambleton & Swaminathan, 1985). In the present study, statistical techniques based on these theories were used to test two possible sources of item bias on the WISC-R with Native children of the Kainaiwa Reserve. One assumption central to the WISC-R is that each subtest is unidimensional, measuring a single underlying or latent trait. If this assumption is incorrect, then the same item may have different meanings for different students. A second assumption is that the items on the WISC-R subtests are presented in an increasing order of difficulty. To date, there have been insufficient data on the WISC-R performance at the item analysis level to precisely identify item difficulty patterns for Native Americans. Evidence that either or both of these assumptions do not hold true for Native children would provide additional support for the contention that WISC-R is culturally biased. Therefore, the present study was designed to investigate the psychometric properties of the WISC-R to determine if these assumptions hold true.

CHAPTER II

REVIEW OF THE LITERATURE

This section discusses the bias against Native students found in the WISC-R, outlines possible sources of cultural bias, and presents literature relating to item difficulty levels.

A primary source of the argument is that the WISC-R is biased against Native students. In fact, Native students typically perform much better on the performance than the verbal WISC-R subtests. For many Native American Indians, the pattern of scores tends to report a discrepancy as much as 25 to 30 points between Verbal IQ and Performance IQ, with Verbal IQ being lower than the Performance IQ (McShane, 1980; McShane & Plas, 1982). In fact, the discrepancy between performance and verbal scores is large enough that for non-Native students it would be interpreted in itself as a sign of learning difficulty or disability. Some researchers have argued that this discrepancy indicates that the majority of Native students do have specific learning disabilities. However, the data sets reported on Native samples have extremely small Ns. This would deem the study useless in reporting bias if only the verbal and performance discrepancy scores are reported. No information has been undertaken to determine or check the utility of the exact bias reported.

For example, Wilson (1981) and Peters (1963) utilized small samples of 12 and 59 Native Americans, respectively. Sachs (1974) studied 33 elementary and 38 junior high Mescalero Apache students. Thurber (1976) employed only 44

Navajo students. St. John and Krichev (1976) reported in their study of 160 Cree and Ojibwa children, youth, and adults in Northwestern Ontario that the mean Verbal IQ ranged from 69.7 to 91.1, with higher Performance IQ scores overall. In this study, they found that the greatest differences were found among younger children ages 6 to 7, with the magnitude of the differences decreasing with age, attributing this to the Native language spoken at home. The children who spoke only the Native language scored lower. As the child became more aculturated into the predominant white school systems, their scores increased. However, St. John and Krichev also reported that there was a nonexistent relationship between achievement and IQ scores and that a gross misdiagnosis of mental deficiency could be made with the 6- and 7-year-olds. They concluded that the verbal and performance IQ should be interpreted separately; otherwise, inaccurate impressions could be made if the Full Scale IQ is used for decision making.

Another study by Seyfort, Spreen, and Lahmer (1980) also found the typical pattern of average performance score and poor verbal scores amongst Native children in southwestern British Columbia. However, their study showed that the test items of the Information, Vocabulary, and Comprehension were out of sequence in terms of difficulty as compared to the WISC-R normed population. They concluded that WISC-R results ought to be interpreted with caution. (It seems unlikely that Native people would have been able to survive as a group if they suffered this kind of global disability.) Whatever the reasons for this discrepancy, its existence serves to illustrate the problem of evaluating Native

students on the basis of non-Native norms. The counterargument is made that the Native African-American and other minority group members were included in the WISC-R norming sample in proportion to their numbers in the larger population (Wechsler, 1974). However, this approach has merely served to obscure possible subgroup differences in response patterns.

Another argument against the possibility of the WISC-R being culturally biased is that scores on the WISC-R do predict future academic performance for all children (Sattler, 1988). The high correlation between WISC-R and academic performance does not in itself prove that the WISC-R is necessarily tapping some sort of pure underlying intelligence; rather, it may be measuring whatever kind of intelligence is most helpful in performing in North American school systems. In any case, research cited by Common and Frost (1988) and Chrisjohn and Lannigan (1986) suggests that, for Native students, performance on the WISC-R is not a reliable predictor of future school performance.

Researchers have focused on two possible sources of cultural bias–the items themselves and the context in which the test is administered. With regard to the test administration context, it has been pointed out that Native children may be intimidated by non-Native testers asking them strange questions in an unfamiliar room (Sattler, 1988). Although not extensive, some researchers support this argument (Chrisjohn & Lannigan, 1986; Common & Frost, 1988).

With regard to the items themselves, various researchers have pointed to individual items that discriminate against all Canadian children. In fact, Vernon

(1977) developed items to be substituted when Canadian children took the test and compared the "Canadian" and "American" versions. In addition, Beal (1988) has provided evidence suggesting the effect of American versus Canadian of the WISC-R items has been overstated. Common and Frost's studies (1988) have pointed to items that, theoretically, rely on knowledge not available to Native (or indeed non-Native) children living in isolated contexts (e.g., Chisasibi, etc.) or items that reflect non-Native cultural values.

These findings have two important implications for the validity of the WISC-R in assessing Native students' intelligence. First, certain items may have a different meaning for minority group children than that assumed by the WISC-R. The validity of each WISC-R subtest depends in part on the assumption that it is measuring a single underlying dimension or latent trait. The possibility that this is not the case for native children needs to be examined. Unfortunately, however, no published research has addressed this issue directly.

Second, various researchers have argued that the assumption that WISC-R subtests are ordered in terms of increasing item difficulty may not be valid for Native students (Mueller et al., 1986). Therefore, the actual order of the items may serve to depress Native student scores given that testing on each subscale ends when the child has failed to answer a certain number of items in a row correctly. In a study conducted by Reynolds and Reschly (1983), item bias was detected in six subtests of the verbal scale. Mishra (1982) also detected item bias

in 15 of the 79 items on three subtests of the verbal scale on her study with 40 Navajo.

The question of differing item difficulty indices is subject to empirical verification. Unfortunately, relatively few studies have been conducted. Seyfort, Spreen, and Lahmer (1980) administered the WISC-R to a sample of Native children in southwestern British Columbia. They found that the Native students showed typical patterns of low verbal scores with higher Performance IQ scores and that many items on the subtests were out of sequence in terms of increasing difficulty when compared to the normed population. Mueller and his colleagues (1986) conducted the most exhaustive investigation of differing WISC-R item difficulty levels to date and the research most pertinent to the proposed study. Mueller et al. conducted a psychometric test norming project in the Northwest Territories using a sample which included Inuit, Caucasian, and Dene children. They analyzed the WISC-R item responses of the Canadian Inuit children who had been included in this larger study. Based on the results of the six verbal and three performance subtests for which items could be coded dichotomously as correct or incorrect, the researchers concluded that the Wechsler tests do not adequately assess Native children who are from a socially, culturally, and linguistically different culture and that no item difficulty data were available upon which to compare their findings. Further, test items have different meaning for various minority groups, with mean averages reflecting relative difficulty of items

across group mean scores. No research has been published to determine item-difficulty patterns and how that affects the group mean scores (Irvine, 1985).

Although the Mueller et al. (1986) study was an excellent study, their analysis of item bias was based on subtests in the WISC-R that could be scored as correct or incorrect. That is, for the Similarities, Vocabulary, Comprehension, and Digit Span subtests, the Mueller et al. study admits to error obtained in restricting the data as such. Their analysis resulted in some loss of response variance and lower item-to-total correlations. However, for the subtest items that can be scored dichotomously as correct or incorrect, no error will be obtained in the item analysis.

Since the Mueller et al. study was conducted, the modern test theory approach for developing tests and assessing test bias has been developed. One of the most promising of these approaches is item response theory (Hambleton & Swaminathan, 1985). To date, there has not been a notable increase in the implementation of item response theory in developing tests (Yen, 1983). For example, Hambleton and Swaminathan (1985) have made reference to test publishers in the state departments of education (Pandey & Carlson, 1983) and various test publishers in the professional and industrial organizations (Guion & Ironson, 1983). Recent reviews cited by Rudner (1977) have claimed a rapid proliferation of such new methods for assessing bias in testing. Hambleton and Cook (1977) also provided a listing of various computer packages available to undertake item response model analyses. No studies have been undertaken

utilizing the item response theory with standardized tests, specifically, the WISC-R.

The Item Response Theory (IRT) model is based on accurately scaling the difficulty of test items which results in a test performance that can predict or explain traits or abilities (Hamblelton & Swaminathan, 1985). The test scores obtained can then be used to predict or explain item and test performance (Lord & Novick, 1968). Traits are not observable measures and, therefore, they are referred to as "latent traits" or "abilities" and the item response model designates a relationship between the observable subject's test performance on the test (Hambleton & Swaminathan, 1985).

In classical test theory the slope of the line predicting item response from latent capacity would be termed "item-total r" (regression) and is the foundation for test development which includes item selection, internal consistency, factor structure, and so forth. In modern test theory the difference is the presumed form of this relation (nonlinear versus classical theory's linear form) and in its correction for the simple linear additive model (e.g., error distributions) (Chrisjohn, Pace, Young, & Mrochuk, 1993).

CHAPTER III

PURPOSE AND OBJECTIVES

The purpose of this study was to examine the assumptions of unidimensionality underlying the WISC-R by utilizing the item characteristic curve technique and item response model. In addition this study also examines whether the Kainaiwa students' pattern of performance on the individual items of two WISC-R Verbal subtests and one Performance subtest conforms to or deviates from the pattern reflected in the standardized WISC-R norms. The other subtests will not be analyzed because scores from these subtests are not binary. Specifically, only subtests that are scored as 0 points or 1 point are considered for analysis. In order to analyze subtests that result in 2-point or 3-point item scores, as Mueller et al. (1986) did, the scores would have had to be transposed into a set of binary scores. This method would "massacre" the data analysis. For these reasons, it is feasible to consider only the subtests in the WISC-R that are scored dichotomously. The research questions are as follows:

1. Do the individual items on each of the Information, Arithmetic, and Picture Completion subtests reflect a unitary underlying dimension for the Kainaiwa respondents in the present study?

2. Are the individual items on each of the Information, Arithmetic, and Picture Completion subtests ordered in increasing difficulty for the Kainaiwa respondents in the present study?

CHAPTER IV

METHODOLOGY

Subjects

The population for this study included 332 Kainaiwa Indian children aged from 6 to 16 years who attended reserve schools in grades 1 through 11 and who resided on the Kainaiwa Indian Reserve in Standoff, Alberta, Canada. The three schools located on the Kainaiwa Indian reserve include Standoff Elementary (K-6), Levern Elementary (K-6), and Kainaiwa High (7-11). The WISC-R was administered to all children between the ages of 6 and 16 years. The Blood Indian children (now referred to as "Kainai") are all bussed to school.

Descriptive Characterstics of the Population

There are approximately 6,000 Blood Indians living on the reserve, with 90% unemployed and living on social assistance. Of the 332 Kainai students, 7.5% claim to speak the Blackfoot language, 31.2% understand the Blackfoot language, and 44.3% have minimal knowledge of the Blackfoot language (Chrisjohn & Towson, 1987). Prior to 1988, the three schools on the Kainaiwa Reserve in Canada were federally operated by the Canadian Federal Department of Indian Affairs. In 1988, the Blood Tribe assumed control of their educational system with funding support from the Canadian federal government. Previous to

Band Control of Education, the Kainaiwa Tribe undertook a comprehensive assessment of all students attending schools on the reserve to determine the level of functioning of all Kainai students in an effort to facilitate planning for effective educational needs. Community meetings were initiated by a team of testers, including one Native psychologist, Dr. Roland Chrisjohn, an Oneida Indian from the Six Nations Reserve in Ontario; Dr. Shelagh Towson, a psychologist from the University of Windsor; and 12 trained and supervised Kainai Native testers consisting of teachers, counselors, and six parents. This was an important component to the testing since it was felt that Kainai testers would be more sensitive to the language and cultural issues than non-Native testers, resulting in fewer errors associated with social situation of testing (Chrisjohn & Towson, 1987).

A number of community meetings were held to inform the public about the testing procedures. The test instrumentations were presented and the parents had the opportunity to ask questions about standardized testing. The information and feedback from the community provided the team with valuable information on possible items that may be biased within the Blood Tribe culture as well as to demystify in their minds the mystery in obtaining IQ scores. Many of the parents in the community were initially reluctant to participate because of past feelings of intimidation imposed by previous non-Native educators who had not taken the time to explain and discuss testing in general. After establishing rapport with the

community members, the assessment project was fully supported and their suggestions were incorporated into the administration of the testing.

First, the WISC-R was administered and scored according to standardized procedures in order to ensure the validity and comparability of results to other WISC-R research. Canadian items were substituted for American items. For example, in Information subtest item #24 "How tall is the average [Canadian] man?" These substitutions are common practice by Canadian psychologists in testing Canadian children (e.g., Crawford & Boer, 1985). Second, the students were given two more items on each test at the end, beyond the usual failure cutoff point to determine in further examination whether items were in order of proper difficulty. Third, the timed subtests were recorded according to protocol, but the testers allowed the students to finish if they were reasonably close, but no credit was given. This modification permitted analysis of the extent to which time limits impacted the results. Fourth, some "success" items based on Kainaiwa cultural knowledge were included at the end of each subtest. These items were scored separately and not included in the WISC-R scores.

Administration of the measures and collection of WISC-R data took place at the Blood Tribe Schools located on the Blood Indian Reserve in Standoff, Alberta, Canada in 1985-86. Parents and guardians agreed to have their children assessed as part of the assessment project for the planning and preparation of Blood Tribe Band control over education. Parents completed the consent forms before their children were assessed (see Appendix A).

Design

This is a descriptive survey utilizing various statistical techniques, including the modern test theory approach on item response theory method (Crocker & Algina, 1983) for subtests that are scored as binary items and by utilizing the "Noharm" program for fitting both unidimensional and multidimensional normal Ogive models of latent trait theory (Fraser, 1988). This is a descriptive survey utilizing two statistical techniques: the modern test theory approach and an item response theory method (Crocker & Algina, 1983) for subtests that are scored as binary items; and, secondly, including the "Noharm" program for fitting both unidemensional and multidimensional normal Ogive models of latent trait theory (Fraser, 1988).

Data and Instrumentation

The data consisted of the item scores on the 30 items included in the Information Subtest, 18 items in the Arithmetic Subtest, and 26 items in the Picture Completion Subtest. Information and Arithmetic subtests are part of the Verbal Scale. The Picture Completion Subtest is part of the Performance Scale. As noted earlier, the rest of the subtests in the WISC-R are not analyzed in calculating the item analyses because the scores are not binary.

Analysis

A descriptive statistical analyses technique employing the "Item Characteristic Curve" theory using a nonlinear approach was employed to summarize and describe the variables of subtest item difficulty and validity in each of the three subtests of the WISC-R. The following steps guided the analysis:

1. Collapsing the data into two groups–females and males.

2. Computing factor loadings on each item of each subtest to determine whether the subtest is measuring a single underlying dimension.

CHAPTER V

RESULTS

Subjects

The subjects included 332 students, 142 male and 190 female, ages 6 to 16 years, enrolled in grades 1 to 11 in the elementary and secondary schools on the Blood Indian Reserve in Alberta, Canada.

Information Subtest Factor Loadings

Separate analyses of the factor loadings on the 30 items of the Information Subtest were conducted for males and females. As indicated in Table 1, factor loadings are not generated for items which were answered correctly by all subjects (item 2 for males, items 1 and 2 for females) or items to which none of the subjects responded correctly (item 28 for males).

A factor loading of less than .500 indicates that the item is not measuring the underlying construct of verbal ability which the Information Subtest purports to measure. For the males, 7 of the 30 items failed to meet this criterion: (1) "What do you call this finger?" (14) "In what direction does the sun set?" (18) "Why does oil float on water?" (22) "What is the main material used to make glass?" (23) "What is the capital of Greece?" (29) "Who was Charles Darwin?" (30) "What does turpentine

Table 1

Information Subtest Factor Loadings

Item #	Male (N = 142)	Female (N = 190)
1	.202	—
2	—	—
3	.666	.517
4	.836	.679
5	.946	.871
6	.906	.697
7	.965	.919
8	.824	.854
9	.976	.803
10	.988	.914
11	.796	.770
12	.762	.828
13	.630	.636
14	.389	.618
15	.792	.662
16	.682	.759
17	.817	.486
18	.380	.429
19	.864	.817
20	.853	.475
21	.829	.758
22	.237	.448
23	.362	.747
24	.872	.731
25	.840	.668
26	.698	.649
27	.729	.587
28	—	.758
29	.476	.766
30	.120	.643

come from?" For females, 4 of the 30 items had factor loadings of less than .500: (17) "From what country did America become independent in 1776?" (18) "Why does oil float on water?" (20) "How many pounds make a ton?" (22) "What is the main material used to make glass?" As indicated, factor loadings on items 18 and 22 were low for males and females.

Information Subtest Difficulty Parameters

Difficulty parameters of each of the items on the Information Subtest for males and females are presented in Table 2. The items are scaled such that theoretically, the values range from negative infinity to positive infinity. In practice, most items fall within a range of ± 3, with negative values indicating easier items and positive values indicating more difficult items. No difficulty parameters are generated for items successfully completed or missed by all respondents.

Confirmation of the assumption that WISC-R Information items are ordered in terms of difficulty level requires that the rank of the difficulty parameters exactly parallels the item order. As indicated in Table 2, this was not the case in the present sample.

These data may be conceptualized in various ways. Given that the criterion for stopping testing on the Information Subtest is five consecutive failures, it is instructive to examine the difficulty parameters and rank orders

Table 2

Information Subtest Difficulty Parameters

	Males			Females		
Item #	Difficulty Parameter	Rank Order	Difficulty Increments	Difficulty Parameter	Rank Order	Difficulty Increments
1	-12.17	2	—	—	—	+1.5
2	—	1	-1	—	—	+1.5
3	-3.69	3	+2	-4.46	3	+1.5
4	-1.90	4	+1	-2.54	4	+1
5	-0.93	6	+2	-1.31	6	+2
6	-0.81	8	+2	-1.68	5	-1
7	-0.69	9	+1	-0.78	10	+5
8	-1.39	5	-4	-1.15	7	-3
9	-0.82	7	+2	-1.10	8	+1
10	-0.52	11	+4	-0.73	11	+3
11	-0.65	10	-1	-0.89	9	-2
12	1.05	14	+4	-0.73	11	+3
13	0.63	12	-2	0.21	13	-1
14	0.88	13	+1	-0.06	12	-1
15	1.28	16	+3	1.69	16	+4
16	2.09	22	+6	2.07	21	+5
17	2.02	20.5	-1.5	3.55	29	+8
18	2.24	23	+2.5	1.83	19	-10
19	1.25	15	-8	0.96	15	-4
20	2.02	20.5	+5.5	4.08	30	+15
21	2.65	24	+3.5	2.28	22	-8
22	8.57	28	+4	3.08	28	+6
23	6.78	27	-1	2.88	25	-3
24	1.69	18	-9	1.80	18	-8
25	1.59	17	-1	1.71	17	-1
26	1.74	19	+2	1.84	20	+3
27	3.01	25	+6	2.60	23	+3
28	—	30	+5	3.04	27	+4
29	5.16	26	-4	2.65	24	-3
30	18.34	29	+3	3.01	26	+2

in blocks of five. For both males and females, items 1 to 5, although out of order, contain only one item with a difficulty ranking greater than five (item 5). A similar pattern is apparent for items 6 to 10, in which item 10 is ranked more difficult than item 9, and items 11 to 15, in which item 15 has a difficulty ranking of 16. For items 16 to 20, however, the difference

between presumed and actual item difficulty becomes more extreme. For males, none of the items are ranked within the appropriate range. Four of the five items are more difficult than they should be, and one item (19) is easier than the test assumes. For females, one item (18) falls within the appropriate range. Three items are more difficult than they should be, and one item (19) is easier than presumed. The same disparity holds for items 21 to 25. For males, one item (21) is within the predicted range, two items (22 and 23) are too difficult, and two items (24 and 25) are too easy. For females, two items (21 and 23) are ranked approximately correctly, one item (22) is too difficult, and two items (24 and 25) are easier than expected. For items 26 to 30, the discrepancy for males is not as extreme as for previous five-item blocks, with two of the five items (26 and 27) being easier than predicted. For females, three of the five items (26, 27, and 29) are easier than expected.

Another way to conceptualize the data, which perhaps provides a better approximation for how a Native child would experience the test, is to examine increments in difficulty level from one item to the next. As the test is presumed to be constructed, each item is one "unit" more difficult than the last. Thus, the child is assured of a certain predictability as he or she proceeds through the test. This was obviously not the case for the Kainaiwa students in the present study. Examination of differences in difficulty level for males indicates that the assumed positive one-unit

increments from one item to the next occurred on only 3 of the 29 item-to-item progressions. Increments of 1.5 or 2 units to more difficult items occurred on five items, increments of 2.5 or to 3 units occurred on three items, and increments of more than 3 units occurred on eight items. Negative increments reflecting a progression from a more difficult to a less difficult item occurred on 10 of the 29 possible progressions; if the WISC-R Information Subtest items were ordered as assumed, no negative increments would occur.

Examination of differences in difficulty level for females reflects the same pattern. The assumed one-unit positive increments occurred on only 2 of the 29 item-to-item progressions. Increments of 1.5 or 2 levels of difficulty occurred on three items, positive increments of 3 units occurred on three items, and positive increments of more than 3 units occurred on eight items. Negative increments ranging from 1 to 10 units occurred on 12 of the 29 progressions.

Arithmetic Subtest Factor Loadings

As indicated in Table 3, the factor loadings obtained for both males and females on Arithmetic Subtest items suggest that these items probably do have the same underlying meaning for the respondents in this sample. For males, one item out of 18 had a factor loading of less than .500 (item 17: "Tony bought a second-hand bicycle for $28. He paid 2/3 of what the

Table 3

Arithmetic Subtest Factor Loadings

Item #	Male (N = 142)	Female (N = 190)
1	---	---
2	.674	---
3	.699	.973
4	.807	.718
5	.586	.308
6	.611	.876
7	.875	.867
8	.929	.836
9	.839	.933
10	.924	.916
11	.971	.920
12	.827	.825
13	.768	.775
14	.844	.855
15	.922	.783
16	.731	.758
17	.416	.537
18	.613	.771

bicycle cost new. How much did it cost new?"). Females also had only one item out of 18 with a factor loading of less than .500 (item 5: "If I cut an apple in half, how many pieces will I have?").

Difficulty Parameters

An examination of the difficulty parameters for the Arithmetic Subtest (Table 4) indicates that the actual difficulty of the items on the second half of the subtest was relatively close to the theoretically assumed difficulty

Table 4

Arithmetic Subtest Difficulty Parameters

	Males			Females		
Item #	Difficulty Parameter	Rank Order	Difficulty Increments	Difficulty Parameter	Rank Order	Difficulty Increments
1	—	1	—	—	1.5	
2	-3.64	2	1	—	1.5	
3	-3.14	5	+3	-2.09	5.	+3.5
4	-1.89	6	+1	-2.26	4	-1.0
5	-3.46	3	-3	-6.60	3.0	-1.0
6	-3.32	4	+1	-1.91	6.0	+3.0
7	-1.68	7	+3	-1.76	7.0	+1.0
8	-0.65	9	+2	-0.92	8.0	+1.0
9	-0.66	8	+1	-0.65	9.0	+1.0
10	-0.15	10	+2	-0.14	10.0	+1.0
11	0.07	11	+1	-0.09	11.0	+1.0
12	0.35	12	+1	0.24	12	+1.0
13	0.98	14	+2	0.68	14	+2
14	0.50	13	-1	0.43	13	-1
15	1.45	15	+2	1.23	15	+2
16	1.82	16	+1	1.61	16	+1
17	5.27	18	+2	3.79	18	+2
18	3.11	17	-1	3.22	17	-1

order. Administration of Arithmetic subtest items is discounted after three consecutive failures, so rank order discrepancies may be examined in groups of three items. For both males and females, the only discrepancies occurred for items 1 to 3 and items 4 to 6. For both male and female respondents, only one item in each group is inappropriately difficult (item 3) or easy (item 5).

Examination of positive and negative increments in difficulty level indicates that 6 of the possible 17 increments for males are +1 unit increments. For females, 7 of the 17 increments are +1 unit increments.

In no case is there a positive or negative increment of more than three difficulty levels from one item to the next for either males or females.

Picture Completion Factor Loadings

For male respondents, factor loadings of .501 or less (Table 5) on 11 of the 26 items (1, 3, 4, 5, 6, 7, 9, 12, 14, 22, 24) of the Picture Completion Subtest suggest that this subtest does not tap only one underlying dimension. For female respondents, 5 of the 26 Picture Completion items had loadings of less than .500 (3, 5, 6, 22, 24). It should be noted that these items had low factor loadings for both males and females.

Difficulty Parameters

Examination of the item difficulty indices in Table 6 reveals that, overall, the Picture Completion Subtest was a relatively easy one for both male and female respondents. For males, only the difficulty parameters for for items 20 to 26 were higher than 0. For females, difficulty parameters for items 22 to 26 were higher than this neutral point. This finding should be kept in mind when examining rank order and difficulty increment discrepancy.

Testing on the Picture Completion Subtest is discontinued after four consecutive failures. Therefore, the actual rank order of item difficulty is examined in four-item groupings, with the exception of items 21 to 26,

Table 5

Picture Completion Subtest Factor Loadings

Item #	Male (N = 142)	Female (N = 190)
1	-.139	.831
2	—	—
3	.367	.399
4	.152	—
5	.398	.491
6	.448	.422
7	.231	.799
8	.591	.657
9	.392	.793
10	.607	.646
11	.674	.731
12	.242	.656
13	.560	.687
14	.501	.549
15	.685	.766
16	.566	.635
17	.633	.650
18	.756	.683
19	.690	.633
20	.670	.726
21	.569	.617
22	.380	.355
23	.654	.743
24	.352	.288
25	.541	.609
26	.517	.610

which are discussed as one group. For both males and females, rankings for items 1 to 4 include only one relatively more difficult item, item 3 for males and item 1 for females. Items 5 to 8 include one easy item for males

Table 6

Picture Completion Subtest Difficulty Parameters

	Males			Females		
Item #	Difficulty Parameter	Rank Order	Difficulty Increments	Difficulty Parameter	Rank Order	Difficulty Increments
1	-15.84	3.0		-3.08	5.0	
2		1.0	-2		1.5	-3.5
3	-6.69	5.0	+4	-6.41	3.0	+1.5
4	-16.16	2.0	-3		1.5	-1.5
5	-4.79	7.0	+5	-3.52	4.0	+2.5
6	-3.54	8.0	+1	-2.96	6.0	+2.0
7	-8.26	4.0	-4	-2.43	7.0	+1.0
8	-2.33	12.0	+8	-2.10	9.0	+2.0
9	-3.40	9.0	-3	-1.69	14.0	+5.0
10	-1.83	14.0	+5	-1.33	16.0	+2.0
11	-1.51	16.0	+2	-1.80	12.0	-4.0
12	-6.32	6.0	-10	-2.21	8.0	-4.0
13	-2.54	10.0	+4	-2.01	10.0	12.0
14	-2.35	11.0	+1	-1.99	11.0	+1.0
15	-1.72	15.0	+4	-1.03	17.0	+6.0
16	-1.25	17.0	+2	-1.76	13.0	-4.0
17	-1.92	13.0	-4	-1.51	15.0	-2.0
18	-0.60	18.0	+5	-0.84	18.0	+3.0
19	-0.55	19.0	+1	-0.76	20.0	-2.0
20	0.21	20.0	+1	-0.77	19.0	-1.0
21	0.57	21.0	+1	-0.13	21.0	+2.0
22	1.80	24.0	+3	1.15	24.0	+3.0
23	1.08	23.0	-1	-.65	23.0	-1.0
24	2.28	26.0	+3	1.57	25.0	+2.0
25	0.88	22.0	-4	0.60	22.0	-3.0
26	1.96	25.0	+3	1.65	26.0	+4.0

and females (item 7 and item 5, respectively), and one relatively more difficult item, item 8, for males and females. For items 9 to 12, item 12 is inappropriately easy for both males and females. Items 10 and 11 are inappropriately difficult for males, as are items 9 and 10 for females. In the item 13 to 16 grouping, only one of the four items is within the appropriate difficulty range, item 15 for males and item 16 for females. Items 13 and 14 are ranked as lower difficulty levels for males and females, while item 16 for males

and item 15 for females are more difficult than the subtest assumes. Items 17 to 26 reflect almost perfect conformity to predicted difficulty levels, with only item 17 easier than appropriate for both males and females.

The positive and negative increments in difficulty from one item to the next range from the assumed +1 unit or 1.5 unit increments (on five of the 24 progressions for males and three of the 24 progressions for females) to extremes of +8 and -10 for the males, and +6 and -4 for the females.

Subtest Comparisons Factor Loadings

Comparison of the factor loadings on the three subtests reveals that the assumption of underlying unidimensionality is most problematic for the Information and Picture Completion Subtests. However, this assumption seemed to be supported overall for the Arithmetic Subtest. A comparison of male and female respondents indicates that more items had low factor loadings for males on both the Information and Picture Completion Subtests.

Item Difficulty

The discrepancy between assumed and actual item difficulty is most apparent on the Information Subtest. Discrepancies were also observed on the Picture Completion Subtest. However, in the latter case, the finding is qualified by the relatively low difficulty level of the entire test for male and female respondents. As was the case for the factor loading analysis, the Arithmetic

Subtest results conform most closely, although not perfectly, to the difficulty order established during WISC-R construction and standardization.

CHAPTER VI

DISCUSSION

The research questions that guided the present study focused on two possible sources of response bias on the WISC-R with reference to the assessment of Native children: (1) violation of the assumption of a unitary dimension underlying each WISC-R subtest, and (2) discrepancies between the presumed linear ordering of items in terms of difficulty and the actual difficulty of these items.

Although these two sources of response bias may be found on all WISC-R subtests, analysis in the present study focused on three subtests: the Information and Arithmetic Subtests from the WISC-R Verbal IQ Scale and the Picture Completion Subtest from the WISC-R Performance IQ Scale. The choice of these subtests was dictated by two considerations. First, analysis of these subtests is facilitated by the fact that responses are scored dichotomously. Second, past research on possible bias for minority children has examined results for these subtest (e.g., Mueller et al., 1986), thus providing the opportunity for some comparisons.

With regard to the question of the assumed unidimensionality of the subtests, the analysis of factor loadings on individual items within each subtest suggests that, for this sample of Kainaiwa children, this assumption may not be entirely valid. Of the three subtests, it could be argued that the Arithmetic

Subtest is least likely to be affected by cultural factors that could affect the meaning of individual items for children from different backgrounds. Consistent with this argument, the few low factor loadings on the Arithmetic Subtest items suggest that, for this sample of Native children, individual items do share a common unitary meaning.

On both the Information and Picture Completion Subtests, however, the number of items with low factor loadings for both males and females cast some doubt on the assumption of the unidimensionality. Examination of the individual items on the Information subtest with low factor loadings does not provide an obvious answer for the failure of these particular items to load more highly. One possibility for some of the questions is that their non-Canadian content (e.g., "How many pounds make a ton?") put them in a different meaning category for the respondents. However, this explanation does not work for more general knowledge items (e.g., "What is the main material used to make glass?"). The findings that males had more low factor loadings than girls is also difficult to interpret. Literature in other areas has suggested that girls, in general, are better students than boys; not because of intelligence difference, but because of their greater ease in conforming to classroom norms regarding good behavior and attentiveness. However, without information on such factors as the respondents' attendance and academic achievement records, this explanation for the observed differences between males and females is very tentative.

The highest proportion of low factor loadings for both males and females occurred on the Picture Completion Subtest, with more than twice as many low factor loadings for males as compared to females. An examination of those items that had low factor loadings does not suggest any possible explanations for this finding. The number of low loading items, especially for males, suggests the advisability of further research on the responses of minority children to this subtest.

A primary focus of the present research was the extent to which items on each of the subtests were not ordered in increasing levels of difficulty for the Kainaiwa sample. The results of the analysis of difficulty parameters suggest that the concern with order of difficulty for minority children is justified. Actual item difficulty did not correspond to assumed item difficulty on any of the three subtests. The severity of the problem varied across subtests. On the Picture Completion Subtest, discrepancies on order difficulty were more extreme on more difficult subtest items, and subject responses indicated that this was the easiest of the three subtests examined for the subject sample.

On the Arithmetic Subtest, the actual order difficulty did not deviate too much from the presumed order difficulty, and examination of the difficulty parameter values indicated that respondents found only the last four items particularly difficult. By contrast, examination of difficulty parameters for the Information Subtest indicates serious item order difficulty discrepancies and a relatively high degree of difficulty experienced relatively early in the test. It is of

interest to note that the move from negatively to positively valued difficulty parameters occurred when students encountered item 12, "Who discovered America?" The WISC-R instructions specifically disallow the answer, "Indians." Testers in the present study were instructed to accept that answer. In fact, few children chose that alternative; however, as suggested by the difficulty parameter, they were also less likely to produce the correct answer, "Columbus, Leif Ericson, Vikings, Amerigo Vespucci," than were the subjects on which the subtest was normed. The results of the present analysis provide strong support for their conclusions.

What are some probable sources of this response bias? First, it is probable that, for at least some of the subjects, the fact that the subtests were in English rather than Blackfoot constituted a barrier to responding correctly. Although, as noted in the Introduction, only 7.5% of the Kainaiwa students speak Blackfoot fluently, the fact that an additional 44.3% claim to understand it suggests that some of the respondents are being raised by parents or guardians whose first language is Blackfoot rather than English. If this is the case, then the adults with whom the children interact may have less facility with English than first-language English speakers. A second source of bias also alluded to previously has to do with the geographical characteristics of the respondents' home community. Being raised in a rural environment may give different meaning to some items than the understanding of an urban child. For example, several of the younger subjects, when asked "In what direction does the sun set," responded with "Over the

mountains." Although no data are available, the author of the present study would guess the child raised in fishing communities on the west coast of North American might answer, "Over the ocean." When children live close to the earth that sustains them, they see the world in different ways than their urban brothers and sisters.

Finally, it is probable that at least some of the order discrepancies are due to deep-rooted cultural differences. According to the data used to order the items on the WISC-R Information Subtest, "Who discovered America?" is an easier item than "What does the stomach do?" This was not the case for the respondents in the present study, because giving the "correct" answer to the former question requires an implicit rejection of their knowledge of themselves as Kainaiwa.

Obviously, these findings have implications for further research and for the assessment and education of Native children in Canadian and American school systems. First, however, it is necessary to address some of the weaknesses and limitations of the present study. First, although the sample size is larger than that used in much previous research on Native students, a larger sample would have increased the reliability of the data. A larger sample size would also have permitted more detailed analyses of subject responses by age. However, the sample of 332 is better than small samples as listed in the literature. Second, the present study focuses on only three subtests of the WISC-R. Analyses of the other subtests of the WISC-R completed by students in the present sample would have provided valuable information, but to utilize the data not scored dichotomously

would result in massacring the data. However, this analyses provides inquiry of tests developed from classical test theory in terms of validity and the need for more sophisticated psychometric studies.

In this study, there is evidence of violation of unidimensionality assumption based on the low factor loadings and violation of order of difficulty assumptions with the WISC-R for the Kainaiwa sample. There is a need to further explore the psychometric properties of the WISC-R in First Nations populations. In accordance to the American Psychological Association, the Standards for Educational and Psychological Testing (1985) state:

> Standard 3.10: When previous research indicates the need for studies of item or test performance differences for a particular kind of test for members of age, ethnic, cultural, and gender groups in the populations of test takers, such studies should be conducted as soon as feasible. (page 5)

From this study, we can see how the WISC-R behaved differently for measuring intelligence in Kainaiwa students and we can see the need to interpret the test with extreme caution.

REFERENCES

Beal, A. L. (1988). Canadian content in the WISC-R: Bias or jingoism. Canadian Journal of Behavioral Science, 20(2), 154-165.

Camilli, G., & Sheppard, L. A. (1987). The inadequacy of ANOVA for detecting test bias. Journal of Educational Statistics, 12(1), 87-99.

Chrisjohn, R. D., & Lanigan, C.B. (1986). Intelligence testing and Indian children: Review and prospects. In R. Anthony & H. McCue (Eds.), Proceedings of the 1984 MOKAKIT Conference (pp. 275-284). Vancouver: MOKAKIT.

Chrisjohn, R. D., Pace, D. F., Young, S., Mrochuk, M. (1993). Psychological Assessment and First Nations: Ethics, theory and practice. Mokakit Journal of Canadian Native Research, 4(1), 50-68.

Chrisjohn, R. D., & Towson, S. M. J. (1987). Kainaiwa Comprehensive Educational Assessment Project (Final report). Standoff, AB: Blood Tribe Education Board.

Chrisjohn, R. D., Towson, S., Pace, D., & Peters, M. (1988). The WISC-R in a Native context: Internal and external analyses. In J. Berry, R. Annis & R. Samuda (Eds.) Ethnic psychology: Research and practice with immigrants, refugees, Native peoples (pp. 14-18). Soujourners, Amsterdam: Swets & Zeitlinger.

Common, R. W., & Frost, L. G. (1988). The implications of measurement of Native students' intelligence through the use of standardized intelligence tests. Canadian Journal of Native Education, 15, 18-30.

Crawford, M. S., & Boer, D. P. (1985). Content bias in the WAIS-R information subtest and some Canadian alternatives. Canadian Journal of Behavioral Sciences, 17, 79-86.

Fraser, C. (1988). Noharm [Computer software]. Armidale, New South Wales, Australia: Author.

Guion, R. M., & Ironson, G. H. (1983). Latent trait theory for organizational research. Organizational Behavior and Human Performance, 31, 54-87.

Gordon, R. A., & Rudert, E. E. (1979). Bad news concerning IQ tests. Sociology of Education, 52, 174-190.

Hambleton, R. K., & Cook, L. (1977). Latent trait models and their use in the analysis of educational test data. Journal of Educational Measurement, 14, 75-96.

Hambleton, R. K., & Swaminathan, H. (1985). Item response theory: Principles and applications. Boston: Kluwer-Nijhoff Publishing.

Irvine, S. H. (1985). What does research have to say about the testing of minorities? In R. Samuda & A. Wolfgang (Eds.), Intercultural counseling and assessment: Global perspective (pp. 165-176). New York: C. J. Hogrefe Inc.

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

Kaufman, A. S. (1979). Intelligence testing with the WISC-R. New York: Wiley and Sons.

Lerner, J. (1981). Learning disabilities: A field in transition. In Learning disabilities: Theories, diagnosis and teaching strategies (3rd ed., pp. 3-25, 33-97). Boston, MA: Houghton-Mifflin.

Lord, F. M. & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

McShane, D. A. (1980). A review of scores of American Indian children on the Wechsler Intelligence Scales. White Cloud Journal, 1(4), 3-10.

McShane, D. A., & Plas, J. M. (1982). Wechsler Scale performance patterns of American Indian children. Psychology in the Schools, 19(1), 8-17.

Mishra, S. P. (1982). The WISC-R and evidence of item bias for Native American Navajos. Psychology in the Schools, 19, 458-464.

Mueller, H., Mulcahy, R., Wilgosh, L., Watters, B., & Mancini, G. J. (1986). An analysis of WISC-R item response with Canadian Inuit children. The Alberta Journal of Educational Research, 32(1), 12-36.

Pandey, T. N., & Carlson, D. (1983). Application of item response models to reporting assessment data. In R. K. Hambleton (Ed.), Application of item response theory (p. 8). Vancouver: Educational Research Institute of British Columbia.

Peters, H. (1963). Performance of Hopi children on four intelligence tests. Journal of American Indian Education, 2, 27-31.

Reynolds, J. R., & Reschly, D. J. (1983). An investigation of item bias on the WISC-R with four sociocultural groups. Journal of Consulting and Clinical Psychology, 51(1), 147-148.

Rudner, L. M. (1977, April). An approach to biased item identification using latent trait measurement theory. Paper presented at the annual meeting of the American Educational Research Association, New York.

Sachs, D. A. (1974). The WISC and the Mescalero Apache. Journal of Social Psychology, 92, 303-304.

Sattler, J . (1988). Assessment of children. San Diego, CA: Author.

Seyfort, B., Spreen, O., & Lahmer, V. (1980). A critical look at the WISC-R with Native Indian children. The Alberta Journal of Educational Research, 26, 14-24.

Standards for educational and psychological testing. (1985). Washingon, DC: American Psychological Association.

St. John, J., & Kricher, A. (1976). Northwestern Ontario Indian children and the WISC. Psychology in the Schools, 13, 4.

Thurber, S. (1976). Changes in Navajo responses to the Draw-a-Man test. The Journal of Social Psychology, 99, 139-140.

Vernon, P. (1977). Final report on modification of WISC-R for Canadian use. Canadian Psychology Association Bulletin, 7, 5-7.

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

Wilgosh, L., Mulcahy, R., & Watters, B. (1986). Assessing intellectual performance of culturally different, Inuit children with the WISC-R. Canadian Journal of Behavioral Science, 18(3), 270-277.

Wilson, M. L. (1981). Rural Alaska WISC-R norms. Anchorage: University of Alaska. (ERIC Document Reproduction Service No. ED 216 481)

Yen, W. (1983). Use of the three-parameter model in the development of standardized achievement test. In R. K. Hambleton (Ed.), Applications of item response theory (pp. 8-23, 167-307). Vancouver: Educational Research Institute of British Columbia.

APPENDIX

PARENT CONSENT FORMS

Dear_____

To help give _____ the best possible educational opportunities, we wish to give him/her an achievement test and/or a medical assessment in order to determine his/her academic abilities.

Would you please sign this form indicating your permission to do this testing. If you have any questions regarding this procedure, please phone _____. We welcome the opportunity to discuss the test results with you with the hope of providing your child educational opportunities which better meet his/her needs.

Sincerely,

_____

I hereby give my permission for _____ to be tested.

_____
Parents signature

_____
Date

Blood Tribe Education Board
P.O. Box 240, Standoff, Alberta T0L 1Y0
Telephone: (403) 737-3966 ◊◊ Fax: 737-2361