INFORMATION TO USERS

This manuscript has been reproduced from the microfilm master. UMI films the text directly from the original or copy submitted. Thus, some thesis and dissertation copies are in typewriter face, while others may be from any type of computer printer.

The quality of this reproduction is dependent upon the quality of the copy submitted. Broken or indistinct print, colored or poor quality illustrations and photographs, print bleedthrough, substandard margins, and improper alignment can adversely affect reproduction.

In the unlikely event that the author did not send UMI a complete manuscript and there are missing pages, these will be noted. Also, if unauthorized copyright material had to be removed, a note will indicate the deletion.

Oversize materials (e.g., maps, drawings, charts) are reproduced by sectioning the original, beginning at the upper left-hand corner and continuing from left to right in equal sections with small overlaps. Each original is also photographed in one exposure and is included in reduced form at the back of the book.

Photographs included in the original manuscript have been reproduced xerographically in this copy. Higher quality 6" x 9" black and white photographic prints are available for any photographs or illustrations appearing in this copy for an additional charge. Contact UMI directly to order.

UMI

A Bell & Howell Information Company
300 North Zeeb Road, Ann Arbor, MI 48106-1346 USA
313/761-4700 800/521-0600

COGNITIVE PERFORMANCE PATTERN UNDERLYING WJ-R TEST PERFORMANCE OF HISPANIC CHILDREN

by

Carla Ellen Hinton

A Dissertation Submitted to the Faculty of the

DEPARTMENT OF EDUCATIONAL PSYCHOLOGY

In Partial Fulfillment of the Requirements
For the Degree of

DOCTOR OF PHILOSOPHY

In the Graduate College

THE UNIVERSITY OF ARIZONA

1 9 9 4

THE UNIVERSITY OF ARIZONA
GRADUATE COLLEGE

As members of the Final Examination Committee, we certify that we have read the dissertation prepared by Carla Ellen Hinton

entitled Cognitive Performance Pattern Underlying WJ-R Test Performance of Hispanic Children

and recommend that it be accepted as fulfilling the dissertation requirement for the Degree of Doctor of Philosophy

Phitzle P. Hilson	12-9-94 Date
John K. Bergan	12/9/94 Date
Darrell Sabers	12/9/94 Date
	Date
	Date

Final approval and acceptance of this dissertation is contingent upon the candidate's submission of the final copy of the dissertation to the Graduate College.

I hereby certify that I have read this dissertation prepared under my direction and recommend that it be accepted as fulfilling the dissertation requirement.

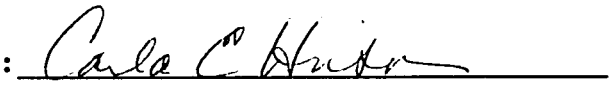
Phitzle P. Hilson — Dissertation Director

12-9-94 — Date

STATEMENT BY AUTHOR

This dissertation has been submitted in partial fulfillment of requirements for an advanced degree at The University of Arizona and is deposited in the University Library to be made available to borrowers under rules of the Library.

Brief quotations from this dissertation are allowable without special permission, provided that accurate acknowledgment of source is made. Requests for permission for extended quotations from or reproduction of this manuscript in whole or in part may be granted by the head of the major department or the Dean of the Graduate College when in his or her judgment the proposed use of the material is in the interests of scholarship. In all other instances, however, permission must be obtained from the author.

SIGNED: 

ACKNOWLEDGEMENTS

The author would like to express gratitude to Dr. Richard Woodcock for use of the WJ-R COG norming data and for his openness in working with a graduate student. Great appreciation is also expressed to Dr. Shitala Mishra for serving as dissertation chair and to all the committee members for their support, patient listening, and continued encouragement.

Special acknowledgement is also made to Lyn Lewis, a friend and teacher, who spent numerous hours helping to edit this dissertation and many other papers throughout the last six years. Special thanks to Richard Schwarz for his support with the LISREL program and statistical analysis.

I am very grateful to my family and friends for supporting me through all the long years that it took to complete this degree. Without the help of my mother, Alice Morris, who took care of the grandchildren when they where sick during finals, I could not have completed this work.

Finally, to my husband Rick and our two children, Megan and Sam, a special thanks for patiently listening and encouraging me through the ups and downs of graduate school.

TABLE OF CONTENTS

TABLE OF CONTENTS--Continued

LIST OF TABLES

ABSTRACT

The purpose of this study was to determine whether the Woodcock-Johnson-Revised Cognitive test is biased when used with a Hispanic population of school-age children. Norming data, provided by R. Woodcock, Ph.D., for grades three, five, eight, and eleven were used for the study. Three hypotheses were explored. The first hypothesis called for a comparison by gender. The second hypothesis called for a comparison of non-Hispanics and Hispanics. The third hypothesis called for comparisons between all combinations of grade levels using only the non-Hispanic subgroup.

Using the results of confirmatory factor analysis from LISREL VIII (1993), the chi-square difference test, and three goodness-of-fit indexes provided evidence of similarity in factor patterns between target groups.

Hypothesis 1 stated that there were no differences between male and female factor patterns. The results of the confirmatory factor analysis supported the acceptance of hypothesis 1.

Hypothesis 2 stated that there were no differences between non-Hispanic and Hispanic students. The results of the confirmatory factor analysis supported a qualified acceptance of hypothesis 2. The relationships between the latent variables are significantly different. Age, therefore, may have been a confounding variable in this study.

Hypothesis 3 stated that there were no differences in patterns between grades. Only one of the six grade comparisons, 3-5, found model 1 to be the preferred model. All other comparisons found model 3 to be the preferred model. The residual or error terms were variable in matrix patterns, indicating that a factor other than age may be influencing the relationships.

A fourth analysis was utilized and determined model 1 to be the preferred model. The results of the analysis indicate that differential patterns of processing, rather than age, may be the variable influencing the relationship of latent variables.

CHAPTER 1

INTRODUCTION

Increased public concern about the rights of minorities was reflected in legislation at the federal and state levels during the 1950's. The intent of the legislation was to improve educational and vocational opportunities for these groups (Anastasi, 1988). Further efforts to improve minority opportunities in education focused on ethical practices in the use of standardized testing.

Special interest groups were instrumental in raising discrimination issues that evolved from standardized testing (Bersoff, 1981; Prasse & Reschly, 1986). Standardized tests, they asserted, systematically excluded minority groups from higher levels of education (Prasse & Reschly, 1986; Samuda & Kong, 1989). Furthermore, the tests placed a disproportionate number of minority students in classes for the educable mentally retarded. Legislative actions and court orders during the 1960's and 1970's further increased professional examinations of these practices. Subsequently, educators became more sensitive about ways in which an individual's cultural environment could impact his or her ability to perform within educational institutions (Angoff, 1986; Weinberg, 1989).

A charge of bias can be easily directed against any available measurement. It is not so easy to obtain

incontrovertible evidence that either substantiates or refutes that charge. In the late 60's and 70's, psychometricians hastened to provide definitions of bias in terms of objective criteria. At the same time, they developed rigorous and precise methods for studying bias and conducted empirical investigations of bias in testing. Subsequently, there has been a proliferation of methods to detect bias. Studies to determine whether or not specific tests are biased continue (Berk, 1982).

Factor analysis and item response theory became the most frequently utilized methods in the search for bias within test instruments. Measurement of different constructs has been among the most frequently cited areas of possible bias in testing, especially in intellectual testing (Reynolds & Kaiser, 1988). Factor analysis and more specifically Confirmatory Factor Analysis (CFA) was found to be extremely effective in helping to identify underlying differential constructs for various types of populations (Anastasi, 1988; Gorsuch, 1983; Messick, 1989).

Because of the issue of bias in testing minority populations, professionals have had to revise older versions of standardized batteries. All of the revisions, in tests like the Wechsler Intelligence Scale for Children-III (Wechsler, 1990) and WJ-R (Woodcock, 1989), try to address the shortcomings identified by research (Roid, 1989; Woodcock, 1989). In the attempt to control and, when

possible, eliminate potential bias, the revisions took into account contemporary standards of test development (Roid, 1989; Woodcock, 1990).

A test's validity is threatened most frequently by nonrepresentative norming samples (Reynolds & Brown, 1984), culturally laden items (Barnett, 1983), and design in the absence of theory (Lowenthal, 1989; Yesseldyke, 1990). Revisions of intelligence tests, such as the WISC-III and the WJ-R, have taken steps to ensure that the norming samples are representative of the various groups that make up the United States population. In addition, statistical analysis of items in the revisions ensure that items on a test are not biased. The direct linking of a test with theory, however, involves a more complex and lengthy process, and is not so easily addressed.

Another step in the process of test validation is to draw the connection between the constructs of a specific test and specific population groups. Empirical evidence of construct consistency across differing population groups, different ecological settings, and different time periods indicates the level of an instrument's generalizabilty (Messick, 1989). It is the ability of an examiner to know whether she/he can generalize and interpret test results across these differing categories that is the crux of the bias and construct validation issue.

Construct validation has highlighted generalization across different populations as a central issue in the evaluation and revision of standardized intelligence tests (Roid,1989). The process of construct validation ensures generalizability and, thus, directly touches on the testing of minority groups. Research has indicated that the underlying constructs of an instrument may vary across culturally different populations. Therefore, construct validation of tests used with minority populations must provide evidence that the underlying constructs are the same as those found in the majority population.

Given the above concerns, the validity of utilizing standardized intelligence tests for placement in special education programs is questionable until empirical evidence is accumulated through the construct validation process. Without this validation process, it is impossible to generalize across differing population samples (Messick, 1989). Interpretations of data for one population are thus not necessarily generalizable to another (Taylor & Richards, 1991).

Purpose of the Study

Educators of bilingual populations need to find objective methods for evaluating their students. It is imperative that the instruments commonly used with this population be examined carefully for any construct differences (Bernal, 1975). The purpose of this study was

to investigate factorial pattern consistency between the norming sample and the Hispanic portion of the norming sample in the WJ-R Tests of Cognitive Abilities (WJ-R COG) (Woodcock & Mather, 1989). It has been suggested that multiple population samples are necessary to provide evidence of factorial consistency and generalizabilty (Messick, 1989). This study is only one step in gaining a better understanding of the populations that can be objectively and fairly tested with the WJ-R COG.

Importance of the Study

A number of factors, educational and political, impact the need for this type of study. There is a significant change occurring in the servicing of special education and regular education populations due to the Regular Education Initiative (REI). Regular Education Initiative (REI), mandated by the federal government, requires schools to provide special education support services to handicapped students within the regular education classroom. In addition, the initiative stipulates that any child who is having educational difficulties may be entitled to participate in this service.

Concurrently, state economic problems are forcing departments of education and local school districts to cut back on the disproportionately high levels of spending associated with special education. More services are being required of current special and regular education staff.

Resource teachers, who have in the past provided direct services to special education students, will be required instead to provide classroom consultation to regular teachers. However, they may still be able to provide a small amount of direct service within the context of the classroom setting. These trends will eventually culminate in the phasing out of resource programs except for the most serious of handicapping conditions.

Many districts see this move away from resource programs to support services in the regular classroom as a way of cutting back on money spent on special education, while at the same time providing special resources to low-functioning students who do not qualify for special education. Another aid in cutting costs is the utilization of tests that do not require the expensive services of school psychologists.

The WJ-R was developed to assess children's learning processes in light of this new public education mandate and tight economic budgets. The WJ-R Test of Cognitive Abilities (WJ-R COG) is not restricted to use by psychologists, as other intelligence tests are. It can be utilized by any professional who has received training in the instrument (Woodcock & Mather, 1989). Thus, schools can train their staffs to provide testing services without increasing the number of school psychologists. Technically, any qualified examiner can derive the IQ/achievement

discrepancy score required by law for a student's placement in a specific learning disabilities program.

In addition to children's right to be tested for handicapping conditions, the REI extends the right of all public school children to an individual education program (IEP). An IEP may require testing to identify and confirm educational and intellectual problems inhibiting academic progress. Consequently, larger numbers of students may be exposed to the cognitive testing that was once required only for special education placement.

Concurrently the public has called for increased accountability regarding the progress of students in the educational system (Hagerty & Abramson, 1987). The changes in federal law cited above put increased emphasis on testing. Tests that do not require special licensing of personnel administering them are more likely to be utilized in these situations. As a result of changes at the federal and state level, a new population may be exposed to increased testing by the public school system. This may result in an increase in testing of minority students, who currently do not qualify for special education services, and who constitute a large proporton of gray area students (students who are not found to be handicapped but who are not performing well academically). Tests that claim to provide information about a child's learning style and that

can be tied to the development of an I.E.P. will be the most likely to be targeted for use.

Minority children are a problematic population for testing (Martin, 1989). Instruments used with this population have generally been shown to be less reliable and less valid than those used for non-minority children (Martin, 1989). In consequence, there is a call for more stringent reliability and validity data on tests used for these populations (Bracken, 1987). This puts increased pressure on test developers to provide more careful studies of testing instruments. It is imperative, therefore, that the WJ-R be fully evaluated for construct validity with minority populations. It is also imperative that the WJ-R be examined for construct differences among various grade levels.

CHAPTER 2

REVIEW OF THE LITERATURE

Factor analysis, because it provides a methodology for studying complex areas of behavioral research, has become a significant tool in the field of educational psychology (Kerlinger, 1986). The fundamental purpose of factor analysis is to facilitate researchers' ability to identify dimensions, called "factors," behind measures or instruments. A factor is a construct, hypothetical entity, or latent variable assumed to underlie measures of almost any kind. The major purpose in identifying a factor is to simplify the description of behavior by reducing the number of variables to the smallest possible number (Daniel, 1989).

Gorsuch (1983) noted that a prime use of factor analysis has been in the development of both theoretical constructs and operational representatives for the theoretical constructs. The parsimony attained by identification of underlying factors helps researchers to provide more concise descriptions of constructs and better interpretations of complex sets of data.

Diverse theoretical perspectives on intelligence utilize factor analytic techniques when identifying types of processes or abilities that may compose intelligence. When used with existing intelligence tests, factor analytic methods promote homogeneity by providing evidence of

internal consistency (Sattler, 1988). Factor analysis is also frequently used when identifying factors underlying a particular test, which in turn provides empirical evidence linking identified factors to a particular theoretical framework. Finally, factor analysis is being used to determine whether construct consistency exists across differing populations. This particular use provides an additional tool for identifying differential construct validity in assessment instruments.

In order to discuss construct validity, it is important to define the meaning of construct. Kerlinger (1986) simultaneously defined and drew a distinction between the terms concept and construct, whose meanings are similar. "A concept expresses an abstraction formed by generalization from particulars" (p. 26). For example, the concept of achievement encompasses various behavioral observations.

A construct, on the other hand, has been deliberately invented to describe specific behavioral observations. Scientists define a construct in order to utilize the construct in specific ways. For example, intelligence is a construct, defined in a specified manner so that it may be observed and measured. A construct has an additional benefit. It can be used in theoretical schemes and, at the same time, relate in sundry ways to other constructs.

Construct validation is a process of evaluating a construct, utilizing observed and measured behaviors. This

validation process can refer either to a specific construct, and its specification through various measurement methods, or to the validation of multiple constructs that are tied into a specific theory.

A general discussion of validity, and construct validation in particular, would be useful at this point. The conceptualization of validity utilized in this study was based on Messick's chapter on validity in "Educational Measurement" (1989). Messick stated that "Validity is an integrated evaluative judgement of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores or other modes of assessment . . . therefore, what is to be validated is not the test or observation device, but the inferences derived from the test scores or other indicators." (p. 13). It is important to note here that validity is a matter of degree. It possesses an evolving property, and, as such, validation is a continuing process.

Because score-based inferences are supported by a variety of sources and mixes of evidence, validity is currently considered a unitary concept (Messick, 1989). The purpose of developing an empirical base for determining validity is to account for response consistencies that reflect underlying determinants of the behaviors or responses (Messick, 1989).

Traditionally, validity is broken down into three types with two subtypes. The three types are content validity, predictive and concurrent criterion-related validity, and construct validity. Current testing standards and guidelines still utilize these three types, although there has been some shift in emphasis.

This shift resulted from the conceptualization of construct validation as an umbrella concept, which includes the components content validity and criterion-related validity. Anastasi's fifth edition of Psychological Testing (1982) presents validation procedures for content, criterion, and construct validation, rather than presenting them as different types of validity. Construct validity is based on an integration of any evidence that influences the interpretation or meaning of the test scores. Construct validity is inclusive of all forms of validity evidence (Cronbach, 1984).

The way in which an instrument's structural composition is related to the instrument's validity is important to construct validation (Messick, 1989). Evaluating the structural components of an instrument involves examining the behavioral manifestations of the construct and their fit within the underlying theoretical constructs proposed (Messick, 1989). It is through a confirmatory factor analysis that empirical evidence is provided as to the fit

of the observed data with the theoretically hypothesized structure.

In addition to providing empirical evidence relevant to individual structures, confirmatory factor analysis allows for an empirically based determination of the generalizability of a factor structure across different population groups. This is effected by utilizing "goodness of fit" between factor structures (Mishra, 1981; Sattler, 1986). Group similarity in factorial structure provides evidence that the underlying constructs are the same for the different groups assessed (Sattler, 1986).

On the other hand, differential structures or patterns may indicate that the constructs tested are not the same for each group. Validation of similarity in the underlying constructs is central to education and psychology's perennial issue of generalizabilty across different populations groups (Messick, 1989; Reynolds & Brown, 1986).

Reschly (1978) noted that patterns of factors may vary significantly among groups, although overall IQ scores may not. For example, Hispanics may score higher on the performance (non-language dependent) subtests, while Anglos may have higher scores on the verbal portions (Berliner, 1988). Although overall ability appears the same, the resulting interpretation of abilities is considerably different (Berliner, 1988).

Persistent variability in factor patterns has been found for minority groups (Mishra, 1981; Valencia & Rankin, 1986). Furthermore, there is mounting evidence that patterns may vary across age groups (Shinn, Algozzine, Marston, & Ysseldyke, 1982; Taylor & Richards, 1991).

Reschly (1979) noted that similarity in factorial patterns is a necessary condition before interpretations of scores can be used in a global manner. Publishers of standardized tests, however, claim that the cost of the analysis prevents testing companies from performing factor analysis among differing age and ethnic groups (Drasgow, 1987). Despite this claim, both ethnicity and age must be examined in determining whether the underlying constructs are the same for various groups. Furthermore, it is important that tests be examined carefully for construct validity among all groups receiving the test. This will ensure that generalizability of interpretation of resulting scores is valid.

Factor analytic techniques were employed during the revision process of the WJ-R. Initially, exploratory factor analysis helped to identify the Horn-Cattell model of intelligence as a framework that best fits with the WJ (McGrew, Werder, & Woodcock, 1989). Factor analysis was also used to develop homogeneity within specific subtests in order to minimize extraneous constructs not identified with the theoretical framework. Factor analysis has been used to

indicate construct convergence between specific sets of subtests with high correlations and divergence of constructs or factors as indicated by low correlations between other sets of subtests.

Confirmatory factor analysis provided evidence that the revised edition of the test provided a good fit with the tests' hypothesized theoretical underpinnings. Further analysis provided information as to the consistency of factor loadings on subtests for all of the age groups within the tests' scope of use.

Analyses for specific minority groups have not currently been performed. The need for further construct validation in the area of construct similarity with a Hispanic population was the focus of this study. Confirmatory factor analyses were utilized to empirically test the similarity between factorial patterns of the Hispanic population with that of the general theory and more specifically a portion of the norming sample. Also patterns of goodness-of-fit were compared among the four grade samples extracted from the norming population.

Validity

Many psychological theories are based on understanding differences between individuals and groups through the measurement of traits and theoretical constructs. Therefore, the construct validity of a test is of

significant concern to both practitioners and researchers in psychology.

An example of this problem can be found in the use of an IQ test, such as the Wechsler Intelligence Scales for Children-Revised (WISC-R, 1974), to place bilingual children in special education programs. Consistent mean score differences have been found between Hispanic minority students and majority students. Claims of bias have led researchers to examine the construct validity for the WISC-R. A number of studies concluded that the test did not measure the same traits or constructs for bilingual children as it did for majority children (Cummins, 1984; Mishra, 1981).

The WISC-R may provide insights into a majority child's cognitive processes, but may be only a test of vocabulary for a child for whom English is a second language (Wilen & van Maanen Sweeting, 1986). In short, the standard interpretations of scores used for majority children may be inappropriate when used for bilingual children (Figueroa, 1985). The consequences of using the WISC-R for assessment of bilingual children may be inappropriate labeling, inappropriate placement, and inappropriate instruction (Figueroa, 1985).

The complexity of a concept such as construct validity has required a variety of methods to examine existing tests for bias. The most frequently used approach is factor

analysis (Gorsuch, 1983), a procedure that makes it possible to identify clusters of test items or subtests that have similar factor loadings. Consistent factor analytic results across populations suggest that whatever is being measured by the testing instrument is being measured in the same manner and is the same construct within each group (Sattler, 1986).

Confirmatory factor analysis uses the constructs from a specific theoretical framework to examine the factor loadings for the items or subtests (Reynolds, 1982). These factors, paired with the theoretical framework, help validate the test in direct reference to the specific theory.

CHAPTER 3

METHOD

The purpose of this chapter is to discuss the procedures that were used in the study. Included are a description of the sample, a description of the instrument used, a rationale, and a description of the statistical procedures for analyses of the data.

Theoretical Underpinnings of the WJ-R

Prior to describing the Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R) (Woodcock & Mather, 1989), it is important to provide a perspective regarding the changes that were made in the revised version. This will be accomplished by a description of the structural and theoretical base of the Woodcock-Johnson Psycho-Educational Battery (WJ) (Woodcock & Johnson, 1977).

The norming of the WJ was extensive (Woodcock, 1977). The sample was based on the 1970 U.S. Census and was gathered from 4,732 subjects in 49 communities (McGrew, 1986). The WJ was composed of three distinct parts: Tests of Cognitive Ability, Tests of Achievement, and Tests of Interest.

The Test of Cognitive Ability (WJTCA) had 12 subtests that could be used to assess both cognitive ability and specific aptitudes. It yielded three scores: a Full Scale score, a Preschool Scale score, and a Brief Scale score. It

was used to provide information relevant to placement in special education classes and for the development of an individual educational program.

The WJ was designed to address specific assessment criteria for evaluating and placing children into special education programs (McGrew, 1986). The WJTCA, or Part I of the battery, was used to provide the ability score in the ability/achievement formula used for a special education placement.

The WJTCA was based on a pragmatic decision-making model that was not tied to a specific theoretical model of intelligence (McGrew, 1986; Woodcock, 1984; Ysseldyke, 1990). This model focused on identifying important placement and instructional decisions that needed to be addressed by the educator (Woodcock, 1984). The lack of a theoretical basis had been a source of frequent criticism of the WJTCA (Kaufman, 1984; Ysseldyke, 1990).

Twelve subtests made up the Broad Cognitive Ability scales. Differing combinations of the subtests provided clusters for the Full Scale, the Preschool Scale, and the Brief Scale. The author claimed that the scales under the Broad Cognitive Ability section provided the scores that were to be used in the ability/achievement discrepancy formula (McGrew, 1986), an essential comparison required for determining eligibility for a learning disabilities program (McGrew, Mather, & Woodcock, 1990).

Although the WJTCA was not constructed to be an intelligence test, it was viewed as a potentially useful indicator of ability or intelligence (McGrew, 1986). Subsequently, an extensive amount of research was compiled comparing the WJTCA to other intelligence tests (Kaufman,1985). The most frequent comparison (McGrew, 1986) was with the Wechsler Intelligence Scale for Children - Revised (WISC-R, 1974).

A study by Reeves, Hall, and Zarkeski (1982) sampled 51 children identified as learning disabled. The scores obtained by the WJTCA were almost one standard deviation below their scores on the WISC-R. Two subsequent studies found similar differences (Ysseldyke, Shinn, & Epps, 1981; Estabrook, 1984). Only McGrew's (1983) study did not find a significant mean discrepancy. In general, it was concluded that, for the same subjects, the WJTCA scores were one-half to one standard deviation below the WISC-R scores.

Shinn, Algozzine, Marston, and Ysseldyke (1982), as well as an earlier study by Ysseldyke, Shinn, and Epps (1981), also provided evidence that the WJTCA was not testing the same constructs as the WISC-R. These studies suggested that problems with sampling of the norming population (Salvia & Ysseldyke, 1981) and construct validity (Thompson & Brassard, 1984; Ysseldyke, 1985) were the two most likely explanations for the differences between the two tests (Reeve, Hall, & Zakreski, 1979).

Since the research indicated that the WJTCA was not testing the same underlying constructs as the WISC-R, the question became: what was it testing?

Algozzine, Marston, and Ysseldyke (1982), using Cattell's (1963) model of fluid and crystallized abilities, examined the underlying constructs of both the WJ and the WISC-R in order to account for the mean differences. They proposed that the WISC-R mainly tapped fluid abilities, while the WJTCA measured crystallized abilities. This was partially supported by the study.

It became apparent, then, that further studies were needed to test for the type of constructs underlying the WJTCA. The constructs were usually identified by the theoretical underpinnings of the test. But since the WJTCA was not originally grounded in theory, it was unclear what constructs needed to be tested (Ysseldyke, 1985).

The identified task became to find the theoretical model that best fit the WJTCA. As a result, exploratory factor analysis of the WJTCA was performed, based on a variety of models of intelligence (McGrew, 1986). The most promising comparison was with Horn-Cattell's (McGrew, 1986) model of cognitive abilities.

Analysis of the WJTCA norming sample consistently identified four factors that related to the Horn-Cattell model. Verbal ability (Gc) was defined primarily by the Picture Vocabulary and Antonyms-Synonyms tests and, to a

lesser extent, by the Analogies and Memory for Sentences (O'Neal, 1988; McGrew,1986). Reasoning (Gf) was primarily defined by the Analysis-Synthesis, Concept Formation, and Analogies subtests. Perceptual speed (Gs) was defined by two timed tests: Spatial Relations and Visual Matching (McGrew, 1987; Rosso & Phelps, 1988). The fourth factor, memory (Gsm), was related to the Memory for Sentences and Numbers Reversed subtests (McGrew, Werder, & Woodcock, 1990).

By combining cognitive and achievement subtests, the auditory processing (Ga) and quantitative skills (Gq) factors were defined. The results of the factor analytic research indicated that 9 of the 12 WJTCA subtests could be classified consistently according to four factors in the Gf-Gc theory. This research provided the basis for the development of the WJ-R COG subtests.

The Woodcock Johnson-Revised

The Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R) (1989) was developed in order to address earlier criticisms of the WJTCA (1977). Twenty-one subtests (1-21) were developed for the Woodcock-Johnson Tests of Cognitive Ability (WJ-R COG). Subtests 1-7 formed the Standard Battery, while subtests 8-21 formed the Supplemental Battery. Thirteen additional subtests (22-35) formed the Woodcock-Johnson Tests of Achievement (WJ-R ACH). Both the Technical Manual (McGrew, Werder, & Woodcock, 1990) and

Ysseldyke's (1990) critique of the test refer to subtests 1-14 in their discussions of the WJ-R COG (see Table 1).

Three goals guided the revision. The first was to revise the existing subtests so that they reflected purer measures of the Gf-Gc factors they represent. The second goal was to increase the number of tests representing the Gf-Gc factors. The third was to enlarge the set of Gf-Gc factors measured.

The goals were addressed by a series of steps. The first step required a complete analysis of the factorial structure of the WJTCA. The second required a preliminary exploratory and confirmatory analysis of subsamples taken from the WJ-R norming population. The third step consisted of a thorough logical and psychological analysis guided by the Gf-Gc theory (McGrew, Werder, & Woodcock, 1990).

The preliminary exploratory and confirmatory analysis was based on a quarter of the norming sample, with the primary objective being to determine how the new subtests would load for the Gf-Gc factors. This led to some modifications of the subtests. Subsequent to the collection of the norming data, a series of exploratory and confirmatory factor analysis studies were conducted. Employing multiple regression, the sample was analyzed for an age-dependent effect. Sixteen subtests were analyzed. Two of the sixteen were taken from the achievement portion while #1-#14 were from the cognitive portion. The subtests

Table 1

Description of Tests of Cognitive Ability

Test 1: Memory for Names

Memory for names measures the ability to learn associations between unfamiliar auditory and visual stimuli. This test is a measure of long-term retrieval (Glr).

Test 2: Memory for Sentences

Memory for sentences measures the ability to remember and repeat simple words, phrases, and sentences presented auditorily. This test is a measure of short-term memory (Gsm) and, to a lesser extent, comprehension-knowledge (Gc).

Test 3: Visual Matching

Visual matching measures the ability to locate and circle the two identical numbers in a row of six numbers. This test is a measure of processing speed (Gs).

Test 4: Incomplete Words

Incomplete words is a test that measures auditory closure. This test is a measure of auditory processing (Ga).

Test 5: Visual Closure

Visual closure measures the ability to name a drawing or picture of a simple object that is altered. This test is a measure of visual processing (Gv).

Test 6: Picture Vocabulary

Picture vocabulary measures the ability to recognize or to name pictured objects. This test is a measure of verbal comprehension or crystallized intelligence (Gc).

Test 7: Analysis-Synthesis

Analysis-Synthesis measures the ability to analyze the components of an incomplete logic puzzle and to determine and name the missing components. This test is a measure of reasoning or fluid intelligence (Gf).

Table 1--Continued

Tests of Cognitive Ability - Supplemental Battery

Test 8: Visual-Auditory Learning

Visual-auditory learning measures the ability to associate new visual symbols. This test is a measure of long-term retrieval (Glr).

Test 9: Memory for Words

Memory for words measure the ability to repeat lists of unrelated words in the correct sequence. This test is a measure of short-term memory (Gsm).

Test 10: Cross Out

Cross out measures the ability to quickly scan and compare visual information. This test is a measure of processing speed (Gs).

Test 11: Sound Blending

Sound blending measures the ability to integrate and then say whole words after hearing parts of the words. This test is a measure of auditory processing (Ga).

Test 12: Picture Recognition

Picture recognition measures the ability to recognize a subset of previously presented pictures within a field of distracting pictures. This test is a measure of visual processing (Gv).

Test 13: Oral Vocabulary

Oral vocabulary measures knowledge of word meanings. This test is a measure of comprehension-knowledge or crystallized intelligence (Gc).

Test 14: Concept Formation

Concept formation measures the ability to identify and state the rules for concepts. This test is a measure of reasoning or fluid intelligence (Gf).

of Calculation (#24) and Applied Problems (#25) make up the quantitative ability (Gq) factor in the Gf-Gc model, measuring the eighth of the nine cognitive ability processes represented by the WJ-R.

Besides the Technical Manual, to date only those articles by Ysseldyke (1990), Reschly, 1990, and Woodcock (1990) had been published describing the WJ-R COG. In addition to the Technical Manual (McGrew, Werder, & Woodcock, 1990), two of these articles dealt with the process of standardization and the fit of the Woodcock-Johnson Test of Cognitive Ability - Revised (WJ-R COG, 1990) to the Horn-Cattell model. Ysseldyke (1990) asserted that the analyses supporting construct validity showed that the fit of the WJ-R COG to the Horn-Cattell model was impressive and convincing.

Reschly's (1990) article recognized the advances that were made in the WJ-R COG with regard to its links to theory and the sophisticated statistical analyses used in the construct validation of the test (Reschly, 1990). He did have concerns regarding treatment validity, which referred to the relationship of the information provided by the test to treatment plans.

The WJ-R COG provides two subtests for each of the eight factors in the Gc-Gf theory, with all sixteen subtests showing clean, high factor loadings (Woodcock, 1990). Another analysis indicated that each of the subtests had low

correlations with the others, signifying a low degree of overlap (McGrew, Werder, & Woodcock, 1990; Woodcock , 1990).

Culture and Language

The Hispanic population is the most rapidly growing minority in the United States (Sattler, 1988); Hispanics are over-represented in learning disabilities programs (Figueroa, 1985). It is important that tests used for evaluating the Hispanic population be free of test bias, and specifically of bias against culturally and linguistically different groups.

As a result of the revision, the underlying constructs of the WJ-R COG subtests have been clearly defined and grounded in theory. It is now feasible to examine the possible existence of construct differences between the Hispanic and majority populations.

Sample

The original norming sample for the WJ-R was based on the 1980 and later U.S. census (McGrew, Werder, & Woodcock, 1990). Norming sample subjects were randomly selected using a stratified sampling design that controlled for both geographic and personal variables such as sex, race, and indices for socioeconomic status (SES). The SES indices were derived from measures of education, income, and occupation. Geographic regions used for the sampling were defined by the 1980 census regions as northeast, northcentral, south, and west. From within these regions,

three types of communities were selected: central city, urban fringe, and outside urban and rural. The intent was to ensure a sample that approximated the race, sex, and SES characteristics of the total U.S. population.

The subjects in this study were a subsample of the original norming population for the WJ-R. The subsample, referred to as the composite group, consisted of students from grades 3,5, 8, and 11, for a total of 1,393 subjects. The composite group reflected the same demographic cross section of the U.S. as found in the overall norming population.

The sample reflected a cross section of the various Hispanic groups within the U.S. Although Hispanics make up 9% of the total U.S. population, they make up 13% of the norming sample. Weighting was used to compensate for the discrepancy (Woodcock & Mather, 1989). Hispanic students numbered approximately 47 subjects in grade three, approximately 46 subjects in grade 5, approximately 47 subjects in grade 8, and approximately 43 subjects in grade 9, for a total sample size of 183 Hispanic subjects.

Instrumentation

Based on past criticism of the WJTCA's lack of a theoretical foundation, Woodcock and Mather (1989) developed the WJ-R COG to be an operational representation of Horn-Cattell's Gf-Gc theory (Horn, 1968). Due to this direct linking, the WJ-R COG ability subtests scores may be useful

in providing documentation for the aptitude portion of the aptitude and achievement discrepancy formula (McGrew, Werder, & Woodcock, 1990). Woodcock labeled this a Type I discrepancy (Woodcock, 1990). As with the model, the WJ-R COG was constructed to assess a hierarchy of cognitive functions.

Specific measures, in the form of subtest clusters, provide information about eight of the nine cognitive processes that the Horn-Cattell theory associates with cognitive ability (see Table 2 for a description of the nine processes). These cognitive processes consist of two major factors, Fluid Reasoning (Gf) and Crystallized Reasoning (Gc). Six other forms of processing are described by Horn (1968) as specific cognitive ability factors. These include Visual Processing (Gv), Auditory Processing (Ga), Processing Speed (Gs), Long-Term Retrieval (Glr), Short-Term Memory (Gsm), and Quantitative Ability (Gq).

The Quantitative Ability process is identified by two subtests found in the Achievement battery. The other seven processes are identified by fourteen subtests in the Cognitive battery. A subtest has not yet been developed for the ninth cognitive process, Correct Decision Speed (CDS).

The WJ-R COG, unlike other testing instruments currently in use, is based on a developmental pattern of cognition. Many instruments provide a view of a child's

Table 2

Description of the Eight Processes

Glr - Long-Term Retrieval

Retrieving information stored over extended periods of time. This involves the storing of information and the fluency of retrieving it. Time is not the essence of Glr as much as intervening tasks that have engaged working memory.

Gsm - Short-Term Memory

Apprehending information and utilizing it within a short period of time. This is the ability to hold information in the immediate situation and then use it within a few seconds.

Gs - Processing Speed

Performing relatively trivial cognitive tasks quickly. "Clerical Speed."

Ga - Auditory Processing

Comprehension and synthesis of auditory patterns. Auditory processing is the ability to fluently comprehend patterns among auditory stimuli.

Gv - Visual Processing

Perceiving and thinking with visual patterns. "Broad visualization" requires fluent thinking with stimuli that are visual in the mind's eye.

Gc - Comprehension-Knowledge

Breadth and depth of knowledge and its effective application. Often referred to as "crystallized intelligence." Represents a person's breadth and depth of knowledge of a culture.

Gf - Fluid Reasoning

Capability to reason in novel situations. Tasks intended to measure this ability do not depend heavily on previous acquired knowledge or previously learned problem-solving procedures.

Table 2--Continued

Gq - Quantitative Ability

Capability to comprehend quantitative concepts and relationships and to manipulate numerical symbols. A measurement of mathematical skills.

cognition only as a downward extension of an adult's (Horn, 1979). The hierarchy of processes found in the Gf-Gc model describes the developmental pattern of cognition that occurs from infancy to adulthood (Woodcock & Mather, 1989; Horn & Cattell, 1967). In other words, the description of the nine processes and the period of development that they start and end is based on qualitative and quantitative changes that occur in a child's cognitive structures (McGrew, Werder, & Woodcock, 1990).

Content and Format of the WJ-R COG

The core of the WJ-R COG is composed of 14 subtests for cognitive ability. There are two subtests for each of the seven broad intellectual categories hypothesized under the Horn-Cattell model. Secondary measures of the seven factors can be found within the same 14 subtests. Table 3 provides a description of the relationship between the subtests and the basic factors described in the Horn-Cattell theory (Woodcock & Mather, 1989).

For both the WJTCA and the WJ-R COG, clusters of subtests, rather than individual subtests (Cummings & Moscato, 1984; McGrew, 1986; McGrew, 1990), provide the basic unit of interpretation. Utilizing cluster profiles increases the reliability of the scores and prevents practitioners from interpreting students' strengths and weaknesses on the basis of individual subtests (Salvia & Ysseldyke, 1981; Woodcock & Mather, 1989). For the WJ-R

Table 3

Hypothesized Model of the WJ-R COG Factor Structure Based on 14 of the 16 Primary Tests Used in Calculation of 7 of 8 Gf-Gc Factor Scores, K-Adult sample (McGrew, Werder, & Woodcock, 1990)

Ability	WJ-R COG Subtest
Gf: Fluid Reasoning	#14 Concept Formation #7 Analysis Synthesis
Gc: Comprehension-Knowledge	#6 Picture Vocabulary #13 Oral Vocabulary
Gv: Visual Processing	#5 Visual Closure #12 Picture Recognition
Ga: Auditory Processing	#4 Incomplete Words #11 Sound Blending
Gs: Processing Speed	#3 Visual Matching #10 Cross Out
Gsm: Short-Term Memory	#2 Memory for Sentences #9 Memory for Words
Glr: Long-Term Retrieval	#1 Memory for Names #8 Visual-Auditory Learning

COG, it is not only the clusters of subtests but also the factors to which they are related that provide the basic units of interpretation (Woodcock & Mather, 1989; McGrew, 1990).

Reliability and Validity

Extensive reliability data had been gathered on the WJ-R COG scales. Statistics had been calculated for reliability of difference scores, test reliability, cluster reliability, discrepancy reliability, test-retest reliability, and interrater reliability (McGrew, Werder, & Woodcock, 1990). Due to the extensive reliability statistics, the specifics of each analysis are not reported. The reader is referred to the technical manual (McGrew, Werder, & Woodcock, 1990).

The internal consistency test data indicated that reliabilities range from the .80's for the individual subtests and in the mid .90's for clusters. The calculation of internal consistency statistics for each age or grade level was based on the data for all subjects at that level in the norming sample. Except for Visual Matching, all reliabilities were calculated by the split-half procedure, using odd and even raw scores, and were corrected for length by the Spearman-Brown formula. The Visual Matching is a timed test with reported reliabilities derived from test-retest data. The cluster reliabilities were calculated by

Mosier's procedure (1943). All other reliability checks were considered acceptable (Ysseldyke, 1990).

The WJ-R COG was examined with respect to content and concurrent and construct validity. Content validity represents the extent to which a test's items sample the ability to be measured. Extensive analyses were conducted to examine items for their appropriateness and comprehensiveness. Clusters were also examined in their relationship to the breadth of their contribution to the cognitive model. From this standpoint, content validity was considered more than adequate (Ysseldyke, 1990).

Concurrent validity was evaluated by comparing the scores on the WJ-R COG with other standardized cognitive tests such as the Kaufman Assessment Battery for Children (KAB-C) (Kaufman & Kaufman, 1983), WISC-R (Wechsler, 1974), Stanford-Binet IV (Thorndike, Hagen, & Sattler, 1986), and the Wechsler Adult Intelligence Scale-Revised (WAIS-R) (Wechsler, 1981). Studies were conducted at different sites, across selected age groups, and with both "normal" and handicapped populations. Data were available that compare the tests to the Horn-Cattell model, while other data provide comparisons of the tests to each other.

The WJ-R COG demonstrates its highest correlations with the respective verbal knowledge or comprehension measures from the other intelligence batteries (McGrew, Werder, & Woodcock, 1990). Correlations are in line with previous

correlational studies that compare the other tests to each other (Keith, 1985; Keith & Dunbar, 1984), and show a significant increase in correlations from studies done with the WJTCA (McGrew, Werder, & Woodcock, 1990).

The construct validity of a test is the extent to which the test may be said to measure a theoretical construct or trait. Ysseldyke (1990) found that the construct validity data strongly supported the test's relationship to the Horn-Cattell model. Extensive testing of construct validity showed not only that the subtests were independent measures, but also that they were clearly related to the Horn-Cattell Model (McGrew, Werder, & Woodcock, 1990; Ysseldyke, 1990).

Procedures

Confirmatory Factor Analysis (CFA) is a tool that the researcher can use in two types of data analysis. Both types assume that the hypotheses tested have either a theoretical base or an empirical base (Long, 1983a). One type compares hypothesized patterns of relationships with a specific data sample. The hypothesized patterns may involve a number of factors, factor structure coefficients for some or all of the variables, or correlations among factors. The second type of analysis compares two data sets derived from actual samples (Kim & Mueller, 1978a). Only research employing the latter type of data analysis is considered in this study.

In this study, CFA provides evidence of consistency or invariance of factor patterns, loadings, and structures (Kim & Muller, 1978b), thus providing a basis for comparison of the various groupings of data. Typically, only the loadings of the latent variables on the manifest variables are examined when discussing possible bias in a test. This is a very limited use of CFA, which also allows for more specific analyses to be made regarding the relationships of the latent variables to each other and the relationship between residual/error patterns. Considering the latent variable relationships and the residual/error patterns decreases the chance that specific hypotheses were supported by a given covariance structure.

For purposes of this study, the two relationships defined by the Lambda and Theta matrices were utilized for judging the potential bias in the WJ-R COG. The Lambda matrix is the most frequently reported result for determining potential bias in a testing instrument, while the Theta matrix is reported only on occasion (Loehlin, 1992). The Phi matrix, which indicates the relationship of the latent variables to each other, is not discussed in any of the studies conducted for examining potential bias in an instrument. This study did not use the Phi matrix in determining potential bias in the WJ-R COG. The results of the Phi matrix are discussed and related to possible uses of the WJ-R COG.

A variety of other methods have been utilized by researchers for assessing model fit. Over all, researchers have been urged not to judge model fit solely on either the basis of chi-square values (Bentler & Bonett, 1980; Marsh, Balla, & McDonald, 1988) or alternative fit indices. Rather, assessments should be based on multiple criteria, including substantive theoretical and conceptual considerations (Joreskorg, 1975; Kirby, 1987). There must be adequate theoretical and empirical support for hypothesizing relationships among variables (Kirby, 1987). The sample must be shown to be representative of both the target population and a multivariate normal distribution (Kirby, 1987).

The WJ-R COG meets the first criterion on a number of accounts. The Horn-Cattell model of intelligence, which provided the theoretical basis of the WJ-R COG, has good empirical support for its theoretical hypotheses concerning intelligence (Horn, 1979). Both exploratory and confirmatory factor analysis done on the WJ-R COG provide a strong link between the Horn-Cattell theory of intelligence and the WJ-R (McGrew, Werder, & Woodcock, 1990). The second criterion, that the data have a multivariate normal distribution, is supported by the data provided in the Technical Manual (McGrew, Werder, & Woodcock, 1990).

The chi-square test is the only method for statistically testing the variance between two groups in CFA

(Byrne, Shavelson, & Muthen, 1989). The other indexes are considered only descriptive in nature. The chi-square test presents a problem that makes it difficult to use as a goodness-of-fit measure in CFA. Chi-square tests are reactive to sample size. Large sample sizes frequently produce a significant chi-square; even a trivial discrepancy in the model requires rejection. By contrast, with small sample sizes a nonsignificant chi-square may be found with a poorly fitting model (Marsh, Balla, & McDonald, 1988; McGrew, Werder, & Woodcock, 1990). Because of the substantial influence of sample size on the chi-square, therefore, researchers have developed a variety of goodness-of-fit indexes (Bollen, 1990; Marsh, Balla, & McDonald, 1988).

In order to satisfy the requirement of using multiple methods for determining goodness-of-fit, LISREL VII provided a number of indexes that indicated the degree of overall fit. Of the various indexes made available by LISREL VII, only three were selected for use in the Technical Manual (McGrew, Werder, & Woodcock, 1990). The selection was based on Marsh, Balla, and McDonald's (1990) comprehensive article, which identified 30 different goodness-of-fit indexes. The three indexes utilized and presented by McGrew, Werder, and Woodcock (1990) in the manual are the Goodness of Fit Index (GFI), the Adjusted Goodness-of-Fit Index (AGFI), and the Root-Mean-Square Residual (RMR) .

The technical manual reports that all confirmatory factor models were estimated by utilizing the maximum likelihood fitting function in the LISREL VIII (Joreskog & Sorbom, 1993) computer program. The reported data were based on a 16-variable (subtests) model with two subtests associated with each of the eight factors.

This study used only 14 of the subtests and seven factors associated with the subtests. Chi-square difference tests, along with similar types of indexes reported in the Technical Manual (McGrew, Werder, & Woodcock, 1990), are used to determine similarity of latent and manifest factorial patterns and structures.

Data in the study include all WJ-R COG scores for all subjects tested in grades 3, 5, 8, and 11. Chi-square difference tests were performed on the data testing for all three null hypotheses. The chi-square difference test provides goodness-of-fit information by helping to identify the preferred model for each set of data. Three goodness-of-fit indexes generated by LISREL VIII confirmatory factor analyses were used in conjunction with the chi-square difference test.

Null hypothesis 1 states that there are no significant pattern differences between genders. Analysis for this null hypothesis was based on the total sample, which includes Hispanics and non-Hispanics across all four grade levels.

Null hypothesis 2 states that there are no significant pattern differences between non-Hispanics and Hispanics. The first analysis for this null hypothesis consisted of the total non-Hispanic portion of the sample and the total Hispanic portion of the sample. Therefore the sample sizes were unequal for the two comparison groups. In order to control for the possible effect of unequal sample sizes, an analysis of equal size groups was compiled from the data. The ten non-Hispanic samples, equal to the size of the Hispanic sample, were randomly selected for comparison with the total Hispanic sample.

Null hypothesis 3 states there is no significant difference in patterns between age groups, based on the four grade levels.

Data Analysis

This study utilized the chi-square difference test and three indexes generated by LISREL VIII (Joreskog & Sorbom, 1993a). It also used the three matrices (Lambda, Theta, and Phi) to analyze the results of the CFA. The chi-square difference test is a good alternative to the chi-square test when determining which model in a hierarchical analysis represents the preferred model or the best-fitting model (Byrne et al., 1989; Long, 1983b). The three indexes selected for the analysis were the Goodness-of-Fit Index (GFI), the Standardized Root Mean Square Residual (SRMR), and the Parsimony Normed Fit Index (PNFI).

The initial step taken in determining goodness-of-fit was to examine two matrices produced by LISREL VIII (1993a). One of these, the Lambda matrix, estimates the invariance of factorial loadings between two target comparison groups. This estimate helps to address the question of whether or not the two samples' latent variables (in this case cognitive processes defined by the Horn-Cattell model of intelligence) show similar loading patterns on the manifest variables (WJ-R Cognitive subtests). As noted before, it is the Lambda matrix that supplies most studies with the estimate of fit, which is then scrutinized when searching for bias.

The Theta matrix, which is based on the error/uniqueness variance, estimates the pattern variance between the two sample groups. This step utilized the residuals of the latent variable loadings on the manifest variables. The discrepancies identified by the Theta matrix provide an indication of the presence of those moderating variables not already identified.

The Lambda and Theta matrices make up the measurement aspect of the model (Byrne, Shavelson, & Muthen, 1989). The Phi matrix describes the relationship between latent variables thereby determining the structure of the model (Byrne, Shavelson, & Muthen, 1989).

Three different samples, extracted from the overall data, provided the basis for the three previously presented

null hypotheses. The analysis done, to test for null hypothesis 1, called for the use of the total sample divided by gender only. This resulted in an analysis that was used as a baseline for comparison with other groups.

In order to test for Null Hypothesis 2, a sample was drawn for comparison of the non-Hispanic group with the Hispanic group. This portion of the analysis was broken into parts "A" and "B". Part A was composed of the total sample of non-Hispanics (1,393 subjects) compared to the total number of Hispanics (183). Since the difference in size may have an effect on the results of the CFA, Part "B", composed of ten groups, was utilized to control for the unequal sample size. Thus, in Part "B" non-Hispanic samples were held equal in size to the Hispanic sample. This required that the non-Hispanic sample be randomly selected from the total non-Hispanic group.

In order to test for null hypothesis 3, a third sample was drawn only from the non-Hispanic group. The third sample was used to make comparisons between six combinations of grade levels. The intent here was to control for differing cognitive structures that may be associated with different grades (ages).

Tests of factorial invariance involved both measurement of the model and structural components of the model. This means that, in LISREL VIII, factor (Lambda) and error (Theta) matrices are of primary interest.

The null hypotheses, based on these matrices, relates to the invariance of hypothesis a, the factor loading pattern (i.e., null hypothesis a: lambda 1 = lambda 2 = ... lambda g). The tenability of null hypothesis a is a logical prerequisite to testing hypothesis b, the error/uniqueness (i.e., null hypothesis b; theta 1 = theta 2 = ...theta g).

Failure to reject the null hypothesis is interpreted as evidence of invariance across groups. Rejection of the null hypothesis, however, leads to testing a series of decreasingly restrictive hypotheses in order to identify the source of nonequivalence.

The procedures for testing the invariance hypotheses are identical to those used in model fitting; that is, a model in which certain parameters are constrained to be equal across groups is compared with a less restrictive model, in which these same parameters are free to take on any value (Byrne, Shavelson, & Muthen, 1989). For example, the hypothesis of an invariant pattern of factor loadings (Lambda) can be tested by first constraining all corresponding Lambda parameters to be equal across groups, and then comparing this model with one in which the number of factors and the pattern of loadings are invariant but not constrained to be equal. If the difference in chi-squares is not significant, then the hypothesis of an invariant pattern of loadings is considered tenable.

Byrnes, Shavelson, and Muthen (1989) found that testing for factorial invariance alone is not sufficient evidence for supporting a particular model. They recommend that further steps be taken to test the invariance of the Theta and Phi matrices. This should be done despite findings of an invariant Lambda matrix. Model testing needs to include multiple indicators of model invariance. This study, therefore, includes the testing of these three matrices through a hierarchical or nested sequence analysis.

The model of the hierarchy or nested sequences is based on the three matrices Lambda, Theta, and Phi. The hierarchy is based on nesting the models by the number of covariance matrices that are fixed (Mulaik, James, Van Alstine, Bennett, Lind, & Stilwell, 1989). The models are numbered one through four according to the constraints imposed upon them.

In the nested sequence of models, model 1, the null model, is the most restricted, requiring that all three matrices be constrained. It provides a baseline for statistically evaluating increments in goodness-of-fit. Model 2 is composed of fixed Lambda and Theta matrices while the Phi matrix is freed. Model 3 has only the Lambda matrix fixed with the Theta and Phi matrices freed. Finally, model 4, the saturated model, is defined by freeing all three matrices.

In order to identify the preferred model, the four nested models are compared utilizing the chi-square difference test (Byrne, Shavelson, & Muthen, 1989; Joreskog & Sorbom, 1989; Loehlin, 1992).

The chi-square difference test is utilized in order to describe the lack-of-fit between adjacent models within the nested model sequence (Byrne, Shavelson, & Muthen, 1989; Loehlin, 1992; Long, 1983b; Mulaik et al., 1989). This analysis may also allow identification of problematic areas, thus pointing the way for future research (Marsh, Balla, & McDonald, 1988).

Joreskog and Sorbom (1993b) provide several goodness-of-fit indices for LISREL VIII that are based, more or less directly, on discrepancies between the observed covariance matrix and the implied matrix. This provides an altogether different approach from that found when using chi-squares (Long, 1983b). The simplest of their measures is just the square root of the mean of the squared discrepancies between the observed and implied matrices, or the root-mean-square residual (RMR) (Loehlin, 1992). When using this measure with covariance matrices, the standardized root-mean-square residual (SRMR) form provides the most interpretable information.

On the other hand, the Goodness-of-Fit Index (GFI) provided by LISREL VIII (Joreskog & Sorbom, 1993b) seeks to measure the fit of the model to the entire covariance matrix

(Mulaik et al., 1989). The subsequent fit, therefore, is based more or less directly on discrepancies between the covariance matrices. It is problematic in that GFI has been shown to be affected by sample size. In addition, it does not take into account parsimony (Mulaik et al., 1989; Marsh, Balla, & McDonald, 1988).

Mulaik et al. (1989) noted that the goodness-of-fit of a model should never be taken into account without also taking into account the parsimony of the model. The principle of parsimony with its roots in Occam's razor insists that an experience be as unified or simply described as possible, by using the smallest number of concepts. Parsimony is an important facet of evaluating the merits of a hypothesis before and after it is subjected to empirical testing.

Parsimony is tied to hypothesis testing because a hypothesis with few freely estimated parameters may be subjected to more tests of possible disconfirmation than a hypothesis containing numerous freely estimated parameters. Each parameter freed for estimation results in not only one fewer constraint on the final solution but also a better fit between the models being compared (Mulaik et al., 1989). Therefore, it becomes important to include an index that takes into account the degrees of freedom. The goodness-of-fit of the model should never be taken at face value without

also taking into account the parsimony of the model (Mulaik et al., 1989).

Accordingly, the Parsimony Normed Fit Index (PNFI) takes parsimony into consideration and should be used in conjunction with the GFI. PNFI utilizes a normed-fit index that is adjusted for loss of degrees of freedom rather than the AGFI that was used for the WJ-R studies (Mulaik et al., 1989). The AGFI produced can be over unity, making the index difficult to interpret. The PNFI, by contrast, is bound by zero to unity and is more easily interpreted (Mulaik et al., 1989).

Discrepancies between the goodness-of-fit indices and the parsimonious normed-fit indices have made researchers reluctant to report these indices because of the apparent implication that something is wrong with their model. It is not unusual for the goodness-of-fit indices to be in the high .90's while the parsimonious normed-fit indices are in the low .50's. Loehlin's (1992) interpretation of the research indicates that GFI's in the .90's and PNFI's above .50 provide adequate assurance that the model is a good fit.

The first CFA analysis compares the goodness-of-fit between genders. This analysis acts as a baseline for the other comparisons. According to a cross-cultural meta analysis study done by Born, Bleichrodt, and Flier (1987), most of the significant gender-related differences turned

out to be rather small, indicating that, in general, little practical value should be attached to gender-related cognitive differences. The expectation is that between the male and female models the CFA will yield high GFI's in the .90's and moderate PNFI's in the .50's. The preferred model is the null model.

The second analysis compares the non-Hispanic and Hispanic samples. A CFA was run that included a comparison between all non-Hispanics and all Hispanics. These analyses included students from the four grade levels previously noted. Because there was concern as to how the differing sizes of the two groups affected the results of the CFA's, a set of 10 CFA's was run. These 10 analysis used the total Hispanic group compared to an equal number of non-Hispanics. The non-Hispanic groups were derived from 10 random samples taken from the total non-Hispanic group.

McGrew, Werder, and Woodcock (1990) had statistically controlled for age differences when reporting the factor loadings at one-year intervals. No information was provided by the technical manual about the effects of development on the factor structure (McGrew, Werder, & Woodcock, 1990). The analysis done in this study allows inspection of data to determine whether the factor structure is invariant across age groups.

Theoretically, factorial structures would be expected to change as cognitive structure changes across age (Horn &

Cattell, 1967). This variance, reflected in the Phi matrix, would be indicated by poor goodness-of-fit indices. Since age is not controlled for in the non-Hispanic/Hispanic comparisons, it is considered a confounding factor that may contribute to the variance found.

In order to gain some perspective on the possible effect of age-related variance as it impacts cognitive structures, a third analysis was performed. The third analysis was performed on the non-Hispanic segment separated into four grade levels. Six combinations of pairs based on grade were analyzed.

CHAPTER 4

RESULTS

The purpose of this study was to identify whether the WJ-R COG is biased for a Hispanic population. CFA allowed the non-Hispanic and Hispanic samples to be evaluated for invariance in factorial loadings and error/uniqueness patterns. The results were summarized, and the three null hypotheses were examined in light of the obtained results.

The recently revised version of the WJ-R COG (1989) has not yet accumulated a data base adequate to support its use with a Hispanic population. However, it is currently used as part of a regular assessment battery for determining placement of Hispanic children in special education programs. This study was an attempt to add information to a body of data necessary for validating WJ-R COG's use with Hispanic students.

The chi-square statistic, which ranges from zero to infinity with the zero indicating a perfect fit, is sensitive to sample size and departures from multivariate normality of the variables. Because of this sensitivity, it is suggested that the chi-square statistic not function as an inferential statistical test but instead function as a comparative measure (Bentler & Bonnet, 1980; Marsh et al., 1990; Loehlin, 1992). It is therefore deemed more

appropriate to use the chi-square difference test as the basis for determining model fit.

The difference in chi-square for competing models is itself chi-square distributed, indicating whether the re-estimated model represents a statistically significant improvement in fit. The chi-square difference test provides the comparative measure for the four hierarchical models and determines which of the models in the nested sequence is the preferred model. They range from the most constrained null model (Model 1), which requires all three matrices to be constrained, to the least constrained or saturated model (model 4), which allows for all three matrices to be freed (see Table 3 for description of Models).

The three matrices provide information about the relationships found in the model. The most desirable fit occurs when the preferred model is the most constrained model (null model). All of the three matrices utilized in the CFA exhibit similar patterns. Typically, reporting of the goodness-of-fit between theoretical models and tests places a heavy emphasis on factorial loadings.

The information supplied by the Technical Manual (McGrew, Werder, & Woodcock, 1991) includes only the factor loadings and does not include specific discussions of the Theta or Phi matrix fits.

As suggested by Marsh et al., (1988) three goodness-of-fit indexes were utilized lending further support to the

results of the chi-square difference analysis. These three indexes are supplied by the Lisrel VIII (Joerskog & Sorbom, 1993) program to support model fit as defined by the chi-square difference test. They include the Goodness of Fit Index (GFI), the Parsimony Normed-Fit Index (PNFI), and the Standardized Root Mean Square Residual (SRMR).

Cole (1987) recommends that GFI values be equal to or above .80 and that SRMR values be below .10. Loehlin (1992), however, recommends that the GFI values be equal to or above .90 and that the PNFI values be equal to or above .40. For this study, the higher GFI value of .90 was utilized.

Presentation of Findings

The purpose of this study was to determine whether bias is present for the Hispanic sample. The gender analysis provides a baseline estimate of the level of fit between groups where differences were expected to be minimal. Age as a confounding factor was also taken into consideration. Based on previous research, there is an expectation that age may contribute to differences in structural patterns. The Technical Manual specifically indicates that age will influence the structural patterns of the cognitive processes involved in intelligence. The fact that the WJ-R COG has the capacity to test the development of these processes over time is seen as a distinct advantage of the test.

Null Hypothesis 1

Null hypothesis 1 states that there are no significant differences in underlying construct patterns between the genders. Confirmatory factor analysis is used to support or disprove conclusions about the invariance of construct patterns between the genders.

The initial CFA, conducted to determine gender pattern similarity, indicates that there is a high degree of invariance across all four models (see Tables 4, 5, 6, & 7). The chi-square difference test, based on the initial CFA, is used to determine the preferred model for gender. Model 1, the null model, is identified as the preferred model for this analysis. This indicates that factorial loadings for latent variables on the manifest variables, residual variables, and factorial structure are invariant between the genders.

Several fit indices are utilized in the study as suggested by Marsh et al. (1988) and Byrne, Shavelson, and Muthen (1989). In accordance with this finding the three indexes based on model 1 provide a GFI score of .97, a PNFI score of .86, and a SRMR score of .031. All scores are within acceptable limits, representing good fits for model 1. The indication is that there are no significant differences found between factor loading patterns, error/uniqueness residuals, and structural pattern for the

Table 4

Description of Hierarchical Models Utilized

Model 1	Matrix(ices) Constrained	Description
1	LXTDPH	Null Model: all matrices are constrained
2	LXTD	Phi matrix free
3	LX	Phi and Theta matrices are free
4	PS	Saturated Model: all matrices are free

Table 5

Chi-Square Difference Test for Gender Analysis

Model*	df	x^2	df	x^2	p
1	161	313.99			
2	133	285.34	28	28.65	>.250
3	119	274.03	42	39.96	>.250
4	112	268.78	49	45.21	>.500

* Preferred model identified by analysis is model 1.

Table 6

Goodness-of-Fit Indexes from LISREL VIII for Gender

Model	GFI	PNFI	SRMR
1	.97	.86	.031
2	.97	.72	.022
3	.97	.64	.021
4	.97	.60	.020

GFI = Goodness of Fit Index
PNFI = Parsimony Normed Fit Index
SRMR = Standardized Root Mean Square Residual

Indexes were produced by LISREL VIII (Joreskog & Sorbom, 1993).

Table 7

Lambda Matrix for Gender

ST	Gf	Gv	Gs	Glr	Gc	Ga	Gsm
1	-	-	-	1.0	-	-	-
	-	-	-	-	-	-	1.0
2	-	-	1.0	-	-	-	-
	-	-	-	-	-	1.0	-
3	-	1.0	-	-	-	-	-
	1.0	-	-	-	-	-	-
4	-	-	-	1.0	-	-	-
	-	-	-	-	-	-	.94
5	-	-	-	-	-	1.8	-
	-	-	-	-	1.0	-	-
6	1.2	-	-	-	-	-	-
	-	-	-	-	.82	-	-
7	-	-	.89	-	-	-	-
	-	.89	-	-	-	-	-
8							
9							
10							
11							
12							
13							
14							

ST = Subtests on the Cognitive Abilities portion of the WJ-R

two genders. Therefore, null hypothesis 1 can be retained and used as a baseline indicator.

Null Hypothesis 2

Null hypothesis 2 provided the basis for the second series of CFA's performed. Null hypothesis 2 states that there is no significant difference between the non-Hispanic and Hispanic populations. Two separate CFA's were performed. The first, A, was based on the sample containing 1,393 non-Hispanics and 183 Hispanics. The second CFA, B, was based on ten samples, each containing 183 non-Hispanics and 183 Hispanics. The non-Hispanics for each of the ten groups were selected at random from the full group of 1,393.

Analysis of Unequal Samples

A chi-square difference test was used to identify the preferred model for the unequal sample. An alpha level of .05 was used to determine significance. Model 3 is identified as the preferred model. Identification of model 3 as the preferred model indicates that the factorial loading pattern is invariant. By contrast, the Theta and Phi matrices are identified as significantly different, indicating that the error/uniqueness and latent factor relationships vary significantly for the two groups (see Table 8).

The goodness-of-fit indexes do not coincide with the preferred models derived from the chi-square difference tests. Examining only model 1, the CFA run on LISREL VIII

Table 8

Chi-Square Difference Test for Non-Hispanic and Hispanic For an Unequal Sample

Model	df	x^2	df	x^2	p
1	161	422			
2	133	.24	28	57.80	<.001
3	119	364	14	31.59	<.001
4	112	.44	7	5.62	<.500

Preferred model is Model 4, the saturated model.

Alpha level is set at .05.

program provided a GFI score of .98, a PNFI score of.86, and a SRMR score of .021. These scores are associated with high levels of fit for model 1. Table 9 provides the index scores for all models.

Analysis of Ten Equal Size Samples

To control for unequal sample size, ten random samples of 183 were selected from the non-Hispanic sample and paired with the total Hispanic sample of 183. Ten samples were utilized to enhance the credibility of the equal sample results.

Chi-square difference tests were done for each of the ten equal-sized samples. In order to control for alpha slippage when using repeated measures, an alpha level of .005 was used to determine significance (Keppel & Zedeck, 1989). The tests identified model 2 as the preferred model for all ten equal-sized samples. Table 10 provides the chi-square difference test for sample 1 of the ten samples. The finding there indicates that, for all ten equal-sized samples, both the factor loadings and the error/uniqueness residuals are invariant.

The chi-square difference test also indicates that a significant source of variance is identified by the Phi matrix. This indicates that the pattern of the relationship between the latent variables is not invariant for the two groups. Examining only model 2, the CFA provides GFI scores ranging from .91 to .94, PNFI scores ranging from .68 to

Table 9

Indexes Reported for Non-Hispanic and Hispanic for the Unequal Sample

Model	GFI	PNFI	SRMR
1	.98	.86	.021
2	.98	.71	.020
3	.98	.64	.019
4	.98	.60	.019

Table 10

Chi-Square Difference Test for Sample #1

Model	df	x^2	df	x^2	p
1	161	364.03			
2*	133	280.63	28	83.40	<.001
3	119	249.61	14	31.02	<.005
4	112	239.68	7	40.95	<.005

Significance based on an alpha level of .005

*Preferred Model

(All chi-square difference tests are listed in Appendix B)

.69, and SRMR scores ranging from .038 to .052. Almost equally high levels of goodness-of-fit were found for model 1 (null model). For model 1, the GFI scores ranged from .90 to .92, the PNFI scores ranged from .81 to .82, and the SRMR scores range from .063 to .091. The goodness-of-fit index scores across all ten samples meet the criterion for finding a good-fitting model in model 1.

A conservative approach to null hypothesis 2, based on the chi-square difference tests, indicates that the Phi matrix shows variant patterns between latent variables. The factor loadings and the error/uniqueness residuals, on the other hand, are invariant. Examination of the three indexes of goodness-of-fit suggest that there are no differences between non-Hispanic and Hispanic populations (see Tables 11 and 12).

Rather than reject null hypothesis 2, it might be more reasonable to examine possible causes for this variance in the Phi matrix. As noted previously, there may be a confounding factor of age being expressed in the Phi matrix.

Null Hypothesis 3

Since age plays an important role in Horn-Cattell's theory of intelligence (Horn & Cattell, 1978), its effects must be controlled for in this study. If they are not, then age may be considered a confounding variable. Although Horn and Cattell almost exclusively examine changes that occur as a result of adult aging, their writings also note that

Table 11

Index Scores for Sample #1

Model	GFI	PNFI	SRMR
1	.92	.81	.091
2	.93	.69	.045
3	.94	.62	.041
4	.95	.58	.026

GFI: Goodness of Fit Index

PNFI: Parsimony Normed Fit Index

SRMR: Standardized Root Mean Square Residual

cognitive structural changes can be expected to occur during the course of childhood (Horn, 1968). In addition, Yesseldyke (1990) notes that because the WJ-R COG is based on a developmental theory of intelligence, its use is warranted when examining developmental changes in children. Age, then, should be examined if the structure of the latent variables is found to be significantly different.

Before isolating age as a variable, two problems must first be addressed. First, the data available do not allow for age to be examined specifically. The sample of children, therefore, is divided not by age but by grade. Typically, the age range within each grade level is not more than a year. Furthermore, the grades reported for the study are two years apart, which allows for no overlap in age groups. Consequently, this study utilizes specific grade levels in place of specific ages.

Second, there is an insufficient number of children in the Hispanic sample to separate them by grade. Therefore, although a Hispanic sample would have provided the best control for both ethnicity and age factors, the non-Hispanic sample, because it contains sufficient numbers, was utilized for making grade (age) comparisons.

The CFA's involved six combinations of grade level comparisons. The non-Hispanic data were separated into four grade levels. Comparisons were made between all possible combinations of grade levels. CFA's were run to determine

similarity between factor, error/uniqueness, and structural patterns for the six age comparisons. To determine the preferred model for each group, chi-square difference tests were performed on all sets of comparisons. The preferred model for each set of grade comparisons is presented in Table 12 (the alpha level used is .01 to adjust for alpha slippage).

Information obtained from the CFA indicated that Model 3 was the preferred model in five of six groups, while model 1 was identified as the preferred model in the comparison between grades three and five. All other grade level comparisons had dissimilar patterns for both the Theta matrix (uniqueness/error terms) and the Phi matrix (relationship of latent variables to each other) (Table 13). In particular, the Theta matrix showed significantly different patterns of uniqueness/error for the two groups. James, Mulaik, and Brett (1982) noted that a Theta matrix indicating variability between two groups' patterns may indicate that there is an as yet unidentified variable at work.

McGrew, Werder, and Woodcock (1991) note only that the factor loadings are consistent across one-year intervals. In addition, this is after an adjustment for age was made. The Technical Manual does not indicate whether the other matrices indicate invariance. Thus, it is impossible to conclude whether other matrices are invariant across ages.

Table 12

Chi-Square Differences for the Six Grade Comparisons Based on the Non-Hispanic Sample. (Alpha level to determine significance .01)

3/5 Grade Comparison

Model	df	x^2	df	x^2	p
1	161	210.72			
2	133	172.24	28	38.58	<.050
3	119	153.89	42	56.83	<.050
4	112	142.92	49	67.80	<.500

3/8 Grade Comparison

Model	df	x^2	df	x^2	p
1	161	299.18			
2	133	227.26	28	71.92	<.001
3	119	182.58	14	44.68	<.050
4	112	174.85	7	7.73	<.250

3/11 Grade Comparison

Model	df	x^2	df	x^2	p
1	161	335.63			
2	133	210.92	28	124/81	<.001
3	119	163.97	14	46.95	<.001
4	112	155.09	7	7	<.250

Table 12--Continued

5/8 Grade Comparison					
Model	df	x^2	df	x^2	p
1	161	229.07			
2	133	200.10	28	28.97	<.250
3	119	169.67	14	30.43	<.001
4	112	163.97	7	5.70	<.050

5/11 Grade Comparison					
Model	df	x^2	df	x^2	p
1	161	291.28			
2	133	214.13	28	77	<.001
3	119	364	14	53	<.001
4	112	144.21	7	16.76	<.025

8/11 Grade Comparison					
Model	df	x^2	df	x^2	p
1	161	268.51			
2	133	222.78	28	45.73	<.010
3	119	186.90	14	35.88	<.001
4	112	176.14	7	10.76	<.100

Preferred model is model #4, the saturated model.

Alpha level is set at .01.

The goodness-of-fit indexes, provided by LISREL VIII (1993), did not make the same discriminations between the grade levels. All grade level combinations identified Model 1 as the model that meets the criterion for a good fit. Table 13 provides the three sets of index scores for both models 1 and 3. Model 1 obtained a GFI score of .95, a PNFI score of .82, and a SRMR score of .071. Model 3 obtained a GFI score of .96, a PNFI score of .72, and a SRMR score of .044.

Upon consideration, another CFA was performed utilizing the non-Hispanic sample. This time non-Hispanic samples were compared in a manner similar to the one employed with the Hispanic/non-Hispanic samples in the analysis done for null hypothesis 2. The assumption tested is that if age is a confounding variable, then model 2 will be the preferred model for a non-Hispanic/non-Hispanic comparison, as it was for the non-Hispanic/Hispanic comparison. However, if age is not a confounding variable, then model 1 will be the preferred model. Furthermore, if age is not culpable, then some other variable must be causing the variance in the Phi matrix.

This non-Hispanic/non-Hispanic comparison was based on twenty groups selected at random from the non-Hispanic sample. Each sample consisted of two non-Hispanic groups, each equal in size to the Hispanic group. This allowed for

Table 13

Indexes for Six Grade Combinations.

Model	GFI	PNFI	SRMR
Grade 3/5 Comparison			
1	.95	.82	.071
2	.96	.69	.044
3	.97	.62	.039
4	.97	.59	.031
Grade 3/8 Comparison			
1	.93	.79	.085
2	.94	.67	.052
3	.95	.61	.043
4	.96	.58	.039
Grade 3/11 Comparison			
1	.92	.78	.088
2	.95	.68	.050
3	.96	.62	.040
4	.96	.58	.035

Table 13--Continued

5/8 Comparison

Model	GFI	PNFI	SRMR
1	.94	.81	.052
2	.95	.68	.046
3	.95	.61	.041
4	.96	.58	.039

5/11 Comparison

Model	GFI	PNFI	SRMR
1	.93	.79	.080
2	.95	.67	.045
3	.96	.62	.041
4	.96	.57	.035

8/11 Grade Comparison

Model	GFI	PNFI	SRMR
1	.94	.80	.069
2	.95	.67	.041
3	.96	.61	.038
4	.96	.57	.035

ten comparisons to be made between groups equal in size to the non-Hispanic/Hispanic comparisons. If age is the factor affecting the latent variable relationships, the preferred model pattern should be similar to that of the comparisons between the non-Hispanic and Hispanic groups. If age is not the variable influencing the latent variable relationships, the pattern would be different.

The results of the fourth analysis found model 1 to be the preferred model for all ten comparisons between non-Hispanics. Consequently, the fourth analysis indicated that the differential relationships between latent variables is probably due not to age, but to some other mediating variable.

Summary

This chapter presents the results derived from the reanalysis of norming data supplied by Richard Woodcock, Ph.D. Both factorial loadings and latent factor structures are considered in establishing whether similar patterns exist between non-Hispanic and Hispanic groups.

In order to provide a baseline for discussing goodness-of-fit, a CFA, based on gender, was performed. The CFA established that, at the most constrained level, model 1 is the preferred model. Therefore null hypothesis 1 is accepted as establishing that the genders have an invariant pattern of factor loadings, error/uniqueness, and latent structural relationships.

A CFA based on the ten equal-sized samples was used for testing null hypothesis 2. The resulting analyses establish that there is invariant factor loading and error/uniqueness found between the non-Hispanic and Hispanic groups. The ten samples all obtained index scores indicating high levels of goodness-of-fit (Table 11). Thus, conditional support is found for acceptance of null hypothesis 2. That is to say, there are no significant differences that would indicate test bias for Hispanic groups.

It must be noted that the Phi matrix, for the ten equal size samples, indicates there is a significant difference in the relationship of the latent variables between the two groups. If the developmental nature of the Horn-Cattell theory of intelligence is taken into consideration, age differences can be considered a possible source of the discrepancy found in the patterns of the relationships between latent variables. This assessment, then, required that an additional analysis to be conducted.

Null hypothesis 3 states that there are no differences between age for the non-Hispanic group. A conditional acceptance of null hypothesis 3 is considered valid. Both between-grade level comparisons and across-grade level comparisons of the non-Hispanic samples indicate that the manifest variables (subtests) have the same patterns of latent variable loadings across groups.

An examination of the comparisons done by grade shows that all but one of the groups (3-5) have significantly different Theta and Phi matrix patterns. Not only is there a difference between the latent variable patterns; there is also a significant difference in the error/uniqueness patterns for these groups. Based on this analysis, age can still be considered a confounding variable. But, due to the differences found in the Theta matrix patterns, there is sufficient evidence to suggest that some other variable may also be affecting the results.

A fourth analysis examines the role of age as a confounding variable by using non-Hispanic by non-Hispanic comparisons as a baseline for comparing non-Hispanic by Hispanic groups. The size of the samples for the non-Hispanic versus the non-Hispanic were the same as those for the non-Hispanic versus the Hispanic. This controlled for the effects of sample size.

If age is a factor in producing variant Phi matrices for the two differing ethnic groups across grade levels, then age should be present as a factor when comparing the same ethnic groups across grade levels. The ten comparisons based on the comparisons of non-Hispanic samples found that model 1 is the preferred model in every case.

This indicates that the differential Phi matrix patterns are not found in comparisons of same-ethnicity groups, with age continuing to be a possible confounding

variable. The results of this analysis are clearer than those of analysis 3 in ruling out the effects of age based on differential Phi matrix patterns.

While the loadings for the latent variables, as measured by the manifest variables, are the same for each ethnic group, the differential development or relative use of various latent variables may be different between ethnic groups. This is consistent with the theory that different ethnic groups have differing learning styles. The assumption is that ethnic groups, due to environmental and cultural differences, develop differing levels of the same perceptual processes (Dunn, 1988; Dunn, Gemake, Jalali, & Zenhausern, 1990; Kleinfeld & Nelson, 1991).

In reporting evidence of bias for specific tests, only the factor loadings of the manifest variables on the latent variables are considered. Therefore, evidence that the subtests of a particular test describe the same latent variables for both groups is taken as evidence that a test is not biased. Typically the results of pattern invariance on the Theta and Phi matrices are not examined.

Using the above criterion for judging whether or not the WJ-R COG is biased, one finds that it is not. This battery of subtests consistently appears to be testing the same factors across non-Hispanic and Hispanic groups. The goodness-of-fit indexes indicate adequate fit between all the age comparisons. Thus, acceptance of null

hypothesis 3 is indicated; age does not seem to have a significant effect on resulting variance found within the Phi matrix. The variation in the Phi matrix pattern may be a result of the differential development of latent variables. The consideration of the differential development of latent variables is discussed in the following chapter.

CHAPTER 5

DISCUSSION

The present investigation examines the WJ-R COG in an attempt to identify bias when testing Hispanic students. A series of CFA's were performed to evaluate the similarity between factorial, residual, and structural patterns for non-Hispanic and Hispanic students.

Due to the short period of time the WJ-R COG has been available, few research studies have been published. The sparse research in this area may also be due to the changing educational trends in service delivery (Mather & Roberts, 1994). The few studies that have focused on the cognitive portion of the WJ-R are derived from analyses based on the norming sample of the WJ-R, allowing for no independent sample analyses. Independently collected data samples need to be analyzed. This investigation constitutes only a preliminary step toward establishing that the WJ-R COG is free from bias when used with a Hispanic population.

Comparisons made between genders indicate that the patterns of fit for all three matrices and indexes are highly similar. This result supports Hypothesis 1: that gender is not a factor contributing to pattern variance. The results lend added plausibility to the expectation that the WJ-R COG provides a good fit with the proposed model.

Non-Hispanic/Hispanic comparisons are the focus of the study. Ten equal size samples are utilized for determining similarity between the non-Hispanic and Hispanic populations. The analyses indicate stable Lambda matrix patterns for all ten samples. This stability, in turn, lends credence to the use of the ten samples as the basis for interpreting the results.

The chi-square difference tests, performed on the ten samples, identify model 2 as the preferred model. This indicates a good fit between the Lambda and the Theta matrices. In addition, these indexes indicate that, even with model 1, there are no significant differences between the Hispanic and non-Hispanic groups.

However, results of the chi-square difference tests for the same ten samples indicate there are significant discrepancies when the Phi matrix is constrained. The implication is that the relationship between the latent variables is significantly different for the non-Hispanic and Hispanic groups.

The difference in patterns between the non-Hispanic and Hispanic groups, as reflected in the Phi matrix, may be explained by differential development of cognitive structures due to environmental influences and/or age. Literature on the Horn-Cattell model of intelligence explicitly states that cognitive structural differences are expected at different ages (Horn, 1972). Furthermore, these

differences may be especially evident during a child's development (McGrew, Werder, & Woodcock, 1991; Ysselydke, 1991). Structural differences associated with age are a reflection of differential development and specialization of the cognitive abilities identified by Horn and Cattell (Cattell, 1963; Horn, 1968; Horn, 1972).

A prime example of differential development of abilities can be found in the charting of growth curves for Gf and Gc abilities (Horn, 1972). Horn specifically cites the development of Gf and Gc abilities as an example of the types of growth curves that result from the interaction of innate and acculturated development occurring over a person's lifetime. The theory of fluid and crystallized intelligence concerns the development of abilities and, more particularly, the development of a gradual distinction between the broad patterns of abilities, Gf and Gc.

In the earliest period of a child's development, a distinction cannot be drawn between fluid intelligence and crystallized intelligence. The separation between Gf and Gc occurs as a result of both development and acculturation. Furthermore, acculturation assists in forming the shape of crystallized intelligence. As development proceeds, individual differences and the extent of acculturation increase. This allows a more distinct pattern to develop for crystallized intelligence (Horn 1972).

McGrew, Werder, and Woodcock (1990) recognize the differential growth curves of the abilities that they term cognitive factors. These cognitive factors represent the latent variables in this study's analyses. The Technical Manual (McGrew, Werder, & Woodcock, 1991) provides information about the pattern of growth curves based on the latent variables. Each latent variable has a point on the growth curve that is associated with a particular age.

This curve portrays the rise and decline of median performance as it occurs across age within the general population at the time the WJ-R COG was normed. It notes, however, that the data utilized are derived from a cross-sectional sample, not from data obtained longitudinally. Because of the use of a cross-sectional sample, the curves presented may not portray accurately the longitudinal rise and decline patterns for individuals (McGrew, Werder, & Woodcock, 1991).

The expectation is that structural differences of the latent variables, based on age, will cause significantly different patterns between age groups. The latent variables identified in the study are the underlying structures for the WJ-R COG (1990) and are based on the Horn-Cattell model of intelligence. It is the resulting Phi matrix of the CFA that represents the relationship of these latent variables to each other. Hence, it is expected that the changes

occurring in the relationships between the cognitive factors are manifested within the Phi matrix.

Because of the proposed structural nature of the cognitive changes, one might expect the Phi matrix to reflect age changes. The second analysis used to support null hypothesis 3 found that age was not a significant factor. Although age did not appear to be a moderating variable in this study, it is possible that it may still have an influence on the relationship of the latent variables. The age spread across this study may not be sufficient to identify differences based on age; a study examining data from a broader range of ages might indicate age effects.

Horn and Catell discuss developmental changes from birth to late adulthood. It may be that the span assessed by this study, approximately ten years, does not include a sufficient portion of the growth curve to show age-related changes. This study finds it questionable, however, that age may be a contributing factor to Phi matrix pattern variability, though the possibility cannot be ruled out completely.

Analysis provided by the Technical Manual (McGrew, Werder, & Woodcock, 1990) suggests that factor loadings, taken from the Lambda matrix and the goodness-of-fit indexes, are the primary indicators for determining pattern similarity. Reynolds (1991) notes that factorial similarity

is essential when evaluating for bias in a test. The CFA, therefore, should focus on the Lambda matrix, which describes factor loadings. The Theta matrix can also be considered in determining the amount of potential error found within the data.

Both the chi-square difference test and the goodness-of-fit indexes generate data that provide evidence of highly similar factor loading patterns and residuals between non-Hispanic and Hispanic groups. It follows, then, that the WJ-R COG does not produce biased results when used with a Hispanic population.

Incorporating the theoretical underpinnings of Horn-Cattell's concept of intelligence may shed some light on the results of the ten Hispanic/non-Hispanic and the non-Hispanic/non-Hispanic analyses. The Horn-Cattell model of intelligence explicitly includes both physiological developmental changes and their interactions with the acculturation process. Doing so implies that acculturation, particular to specific minority groups as well as subdivisions within a particular minority, may influence growth curve patterns (Horn, 1968; Horn 1978).

If acculturation differences lead to differential patterns of cognitive factors, the evidence for these differences may lie within growth curve patterns. The WJ-R COG's strong ties to Horn-Cattell's theory of intelligence

provide a reliable and valid format for exploring these potential differences (Ysseldyke, 1991).

Further studies examining the change in the Phi matrix pattern in relation to age and learning style would be beneficial. These same patterns found in the Phi matrix, then, may be linked to the discussion of differential learning styles for ethnic minorities (Dunn, 1988; Dunn & Dunn, 1978; Dunn, Gemake, Jalali, & Zenhausern, 1990; Dunn,R., Griggs, S.A., & Price, G., 1993). Furthermore, the use of the Phi matrix in CFA analysis may facilitate examining the differential development of cognitive processes and the relationship of these processes to learning.

If it becomes possible to ascertain that differential acculturation results in differential development of cognitive processes, a number of areas of research will come to the forefront. A series of questions will then need to be asked. To wit, how does the development of these processes influence a child's ability to perform successfully within a public educational setting? Can these abilities be taught within a regular school setting? Does understanding these processes lead to more effective curriculum development and learning?

Differential learning styles of individuals, as well as members of groups, have been explored in attempts to link learning style with effective teaching methods (Clark &

Halford, 1983; Ellis & Kimmmel, 1992; Dunn et al., 1990, Dunn, Griggs, & Price, 1993). Furthermore, research, particularly into relationships between growth patterns of cognitive factors and ethnicity, may provide a crucial insight into differential learning styles (Kleinfeld & Nelson, 1991).

An essential issue to address here is the relationship between learning style and instructional method (Taylor, 1991). The link between the identification of a learning style (Kleinfeld & Nelson, 1991) and the type of instruction necessary for optimal learning has been tenuous (Reschly, 1991). Reschly (1991) is critical of standardized testing because of the ambiguous link that is formed between strengths and deficits, identified by the tests, and specific classroom instruction or remediation. He has more specifically been concerned about the WJ-R's links to effective instructional methodology (Reschly, 1991).

Initially, differential growth patterns of cognitive processes, if any, must be identified for particular minority groups. Once identified, they may provide insight into how acculturation affects specific abilities, which in turn leads to differential learning patterns. Defining the relationship between particular patterns of abilities (processes) and specific instructional methods could facilitate the identification of those methods that lead to more effective teaching.

The WJ-R COG provides researchers with a strong, theoretically based test that can be used to explore the relationship between differential cognitive patterns, styles of learning, and instructional methodology. The WJ-R COG must, ultimately, be linked to instructional methods that facilitate cognitive development and determine appropriate methods for remediation.

Many school districts within the state of Arizona are making special education placement decisions based solely on the discrepancy between a student's ability and her/his academic scores. From the author's practical experience, the WJ-R Achievement portion of the battery is frequently used by schools to provide the standardized academic scores that are compared to the student's IQ.

Initially, special education teachers had hoped to use the WJ-R COG for determining the ability level of the student and the existence of processing deficits. Currently, however, the WJ-R COG is not used as the measure of ability. This may be due in part to the perception of special education teachers that the standardized score derived from the WJ-R COG is consistently lower than the IQ score provided by the WISC-III. Low standardized scores make it more difficult to show the discrepancy between ability and academic achievement that is required for placement in a special education program.

Another possible reason for the infrequent use of the WJ-R COG battery is that many school districts no longer require that a processing deficit be identified for placement in a specific learning disabilities program. Only a discrepancy between IQ and academic achievement is required.

If a link is established between the pattern of processing abilities identified by the WJ-R COG and methods of instruction, this portion of the battery would be very useful to both special education and regular education teachers. As the assessment process presently exists, use of the WJ-R COG is unlikely to increase.

There are aspects of this study that might bear modification when developing a design for future studies. Increasing the size of the Hispanic sample to allow for separation by age and limiting both the ethnic diversity of groups within the non-Hispanic sample and the cultural diversity within the Hispanic sample are two such aspects.

The insufficient size of the Hispanic sample made it difficult to control for the confounding aspect of age. A future study, planned so that the size of both the non-Hispanic and Hispanic groups would allow for separation by age, may be able to control for this potentially confounding variable. In addition, age might be more specifically defined than by grade levels.

The current study was only moderately successful in controlling for ethnicity. The non-Hispanic sample contained other ethnic minorities that were not defined as Hispanic. In order to control for the effects of ethnicity, other minority populations should not be included in the non-Hispanic group. A more specific criterion, limiting ethnic diversity, might have provided more precise results.

Significant cultural diversity within the Hispanic population (Reynolds & Brown, 1984) might also be a confounding factor when evaluating for bias. The WJ-R norming samples, however, do not take into account the cultural diversity of the Hispanic sample.

Controlling for the above aspects would not only provide a more rigorous study of pattern similarities; it would also lend additional support to the position that the WJ-R is not biased in relation to the Hispanic population. An additional result might be a clearer representation of the developmental growth curves and their relationship to age and/or cultural differences.

APPENDIX A

HUMAN SUBJECTS APPROVAL LETTER

Human Subjects Committee

THE UNIVERSITY OF
ARIZONA.
HEALTH SCIENCES CENTER

1622 E. Mabel St.
Tucson, Arizona 85724
(602) 626-6271

July 18, 1994

Carla E. Hinton, M.Ed.
c/o Shitala Mishra, Ph.D.
Department of Educational Psychology
Education Building
Main Campus

RE: COGNITIVE PERFORMANCE PATTERN UNDERLYING WJ-R TEST PERFORMANCE OF HISPANIC CHILDREN

Dear Ms. Hinton:

We have received documents concerning your above cited project. It is our understanding that this project involves use of existing data which will have no identifiers. Regulations published by the U.S. Department of Health and Human Services [45 CFR Part 46.101(b)(4)] exempt this type of research from review by our Committee.

Thank you for informing us of your work. If you have any questions concerning the above, please contact this office.

Sincerely yours,

William F. Denny, M.D.
Chairman
Human Subjects Committee

WFD:js

cc: Departmental/College Review Committee

APPENDIX B

TEN CHI-SQUARE DIFFERENCE TESTS FOR THE TEN RANDOM SAMPLES FOR THE NON-HISPANIC/HISPANIC ANALYSIS

Chi-Square Difference Tests for the Ten Random Samples for the Non-Hispanic/Hispanic Analysis.

Model	df	x	df	x	p
Sample #1					
1	161	364.03			
2	133	280.63	28	83.40	< .001
3	119	249.61	14	31.02	> .005
4	112	239.68	21	40.95	> .005
#2					
1	161	325.69			
2	133	264.11	28	61.58	<.001
3	119	233.29	14	30.82	>.005
4	112	229.39	21	34.72	>.005
#3					
1	161	353.44			
2	133	277.67	28	75.77	<.001
3	119	259.49	14	18.18	>.100
4	112	250.02	21	27.65	>.100
#4					
1	161	334.71			
2	133	266.10	28	68.61	<.001
3	119	241.35	14	24.75	>.025
4	112	228.95	21	37.15	>.010
#5					
1	161	363.48			
2	133	299.98	28	63.50	<.001
3	119	273.49	14	26.49	>.010
4	112	266.38	21	33.60	>.025

#6

1	161	363.46			
2	133	299.98	28	63.48	<.001
3	119	273.49	14	26.49	>.010
4	112	266.38	21	33.49	>.025

#7

1	161	337.23			
2	133	285.24	28	60.80	<.001
3	119	260.28	14	24.96	>.025
4	112	243.62	21	41.62	>.005

#8

1	161	362.93			
2	133	293.12	28	69.81	<.001
3	119	265.31	14	27.81	>.010
4	112	261.02	21	32.10	>.050

#9

1	161	325.69			
2	133	264.11	28	61.58	<.001
3	119	233.29	14	30.82	>.005
4	112	229.39	21	34.72	>.005

#10

	161	342.32			
1	133	278.46	28	63.86	<.001
2	119	246.83	14	31.63	>.005
3	112	240.64	21	37.82	>.010
4					

Alpha level of .005 is used to determine significance. This level was adjusted to allow for alpha slippage due to multiple analysis performed.

APPENDIX C

INDEXES PRESENTED FOR PREFERRED MODEL 2 AND MODEL 1

Indexes presented are for the preferred model 2 and model 1.

Model	GFI	PNFI	SRMR
Sample #1			
1	.92	.81	.091
2	.93	.69	.045
Sample #2			
1	.92	.82	.077
2	.94	.69	.047
Sample #3			
1	.90	.81	.082
2	.93	.69	.052
Sample #4			
1	.91	.82	.085
2	.94	.69	.047
Sample #5			
1	.90	.82	.076
2	.91	.68	.042
Sample #6			
1	.91	.82	.077
2	.93	.68	.045

Model	GFI	PNFI	SRMR
Sample #7			
1	.90	.82	.063
2	.92	.68	.044
Sample #8			
1	.92	.82	.069
2	.94	.69	.043
Sample #9			
1	.91	.82	.073
2	.93	.69	.043
Sample #10			
1	.90	.82	.077
2	.92	.68	.038

REFERENCES

Anastasi, A. (1985). Some emerging trends in psychological measurement: A fifty year perspective. Applied Psychological Measurement, 9, 121-138.

Balderjahn, I. (1988). A note to Bollen's alternative fit measure. Psychometrika, 53, 283-285.

Barnett, D. W. (1983). Nondiscriminatory Multifactored Assessment, New York: Human Sciences Press.

Bentler, P. M. (1990). Comparative fit indexes in structural models. Psychological Bulletin, 107, 238-246.

Bentler, P. M., & Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, 88, 588-606.

Berliner, D. C. (1988). Meta-comments: A discussion of critiques of L. M. Dunn's monograph bilingual hispanic children on the U.S. mainland. Hispanic Journal of Behavioral Sciences, 10, 273-299.

Bernal, E. M. (1975). A response to "educational uses of tests with disadvantaged subjects". American Psychologist, 30, 93-95.

Bersoff, D. N. (1981). Testing and the law. American Psychologist, 36, 1047-1056.

Bersoff, D. N. (1983). Regarding psychologists' testing: The legal regulation of psychological assessment. In C. J. Scheirer & B. L. Hammonds (Eds.), Psychology and the Law (pp. 37-88), Washington, DC: American Psychological Association.

Bersoff, D. N. (1984). Social and legal influences on test development and usage. In B.S. Plake (Ed.), Social and technical issues in testing: Implications for test construction and usage (pp. 87-109), Hillsdale, NJ: Erlbaum.

Bollen, K. A. (1990). Overall fit in covariance structure models: Two types of sample size effects, Psychological Bulletin, 107, 256-259.

Born, M., Bleichrodt, N., & Van Der Flier, H. (1987). Journal of Cross-Cultural Psychology, 18, 283-314.

Bracken, B. A. (1987). Limitations of preschool instruments and standards for minimal levels of technical adequacy. Journal of Psychoeducational Assessment, 5, 313-326.

Byrne, B. M., Shavelson, R. J., & Muthen, B. (1989). Testing for the equivalence of factor covariance and mean structures: The issue of partial measurement invariance. Psychological Bulletin, 105, 456-466.

Cattell, R. B. (1963). Theory of fluid and crystallized intelligence: A critical experiment. Journal of Educational Psychology, 54, 1-22.

Clark, L. A., & Halford, G. S. (1983). Does cognitive style account for cultural differences in scholastic achievement? Journal of Cross-Cultural Psychology, 14, 279-296.

Cleary, T. H., Humphreys, L. G., Kendrick, S. A., & Wesman, A. (1975). Educational uses of tests with disadvantaged students. American Psychologist, 30, 15-41.

Costenbader, V. K., & Perry, C. (1990). Test review of the Woodcock-Johnson psychoeducational battery-revised, Journal of Psychoeducational Assessment, 8, 180-184.

Cummings, J. (1984). Bilingual special education: Issues in assessment and pedagogy. San Diego: College Hill.

Cummings, J. (1985). Review of the Woodcock-Johnson Psycho-Educational Battery. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp.1759-1762). Lincoln, NE: University of Nebraska Press.

Cummings, J., & Moscato, E. (1984a). Research on the Woodcock-Johnson Psycho-Educational Battery: Implications for practice and future investigation. School Psychology Review, 13, 33-40.

Cummings, J., & Moscato, E. (1984b). Reply to Thompson and Brassard. School Psychology Review, 13, 45-48.

Daniel, L. G. (1989). Comparisons of exploratory and confirmatory factor analysis. Paper presented at the annual meeting of the Mid-South Educational Research Association, Little Rock, AR.

Drasgow, F. (1984). Scrutinizing psychological tests: Measurement equivalence and equivalent relations with external variables are the central issues. Psychological Bulletin, 95, 134-135.

Drasgow, F. (1987). Study of the measurement bias of two standardized psychological tests. Journal of Applied Psychology, 72, 19-29.

Dunn, R., Gemake, J., Jalali, F., & Zenhausern, R. (1990) Cross-cultural differences in learning styles of elementary-age students from four ethnic backgrounds. Journal of Multicultural Counseling and Development, 18, 68-92.

Dunn, R., Griggs, S. A., & Price, G. E. (1993). Learning styles of Mexican American and Anglo-American elementary school students. Journal of Multicultural Counseling and Development, 21, 237-247.

Elliot, R. (1988). Tests, abilities, race, and conflict. Intelligence, 12, 333-350.

Figueroa, R. A. (1985). Assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), Best practices in school psychology (pp. 113-123). Kent, OH: National Association of School Psychologists.

Flaugher, R. L. (1978). The many definitions of bias. American Psychologist, 33, 671-679.

Garwood, S. G., & Fewella, R. R., & Neisworth, J. T. (1988). Public Law 94-142: You can get there from here! Topics In Early Childhood Special Education, 8, 1-11.

Gorsuch, R. L. (1983). Factor Analysis. NJ: Erlbaum.

Gustafsson, J. (1984). A unifying model for the structure of intellectual abilities. Intelligence, 8, 179-203.

Hagerty, G. J., & Abramson, M. (1987). Impediments to implementing national policy change for mildly handicapped students. Exceptional Children, 53, 315-323.

Hawley, W. D. (1988). "Missing pieces of the educational reform agenda: Or, why the first and second waves may miss the boat." Educational Administration Quarterly, 416-437.

Hilard, E. R. (1989). The early years of intelligence measurement. In R. L. Linn (Ed.), Intelligence: Measurement, theory, and public policy (pp. 7-28). Chicago: University of Illinois Press.

Hollander, P. A. (1978). Legal handbook for educators. Colorado: Westview Press.

Hollander, P. A. (1982). Legal context of educational testing. In A. K. Wigdor & W. R. Garner (Eds.), Ability testing: Uses, consequences, and controversies (pp. 195-231). Washington, DC: National Academy Press.

Horn, J. L. (1968). Organization of abilities and the development of intelligence. Psychological Review, 75, 242-259.

Horn, J.L. (1978). The nature and development of intellectual abilities. In R. T. Osborne, C. E. Noble, & N. Weyl (Eds.), Human variation: The biopsychology of age, race, and sex (pp. 107-136). New York: Academic Press.

Horn, J. L. (1979). Trends in the measurement of intelligence. Intelligence, 3, 229-240.

Horn, J. L. (1989). Models of intelligence In R. L. Linn (Ed.), Intelligence measurement, theory, and public policy (pp. 29-73). Chicago: University of Illinois Press.

Horn, J. L., & Cattell, R. B. (1967). Age differences in fluid and crystallized intelligence. Acta Psychologica, 26, 107-129.

Humphreys, L. G. (1985). Race differences and the Spearman hypothesis. Intelligence, 9, 275-283.

Jackson, G. D. (1975). On the report of the ad hoc committee on educational uses of tests with disadvantaged students. American Psychologist, 30, 88-93.

James, L. R., Mulaik, S. A., & Brett, J. M. (1982). Causal Analysis: Assumptions, models, and data. London: Sage.

Joreskog, K. G., & Sorbom, D. (1989). LISREL VII:A guide to the program and applications (2nd ed.). Chicago, SPSS.

Joreskog, K. G., & Sorbom, D. (1993a). Lisrel 8: Structural equation modeling with the simplis command language, Scientific Software International.

Joreskog, K. G., & Sorbom, D. (1993b). Lisrel 8 User's Reference Guide, Scientific Software International.

Katzenmeyer, W. G., & Stenner, A.J. (1977). Estimation of the invariance of factor structures across race and sex with implications for hypothesis testing. Educational and Psychological Measurement, 37, 111-119.

Kaufman, A. (1985). Review of the Woodcock-Johnson Psycho-Educational Battery. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 1762-1765). Lincoln, NE: University of Nebraska Press.

Kaufman, A., & Kaufman, N. (1983). The Kaufman Assessment Battery for Children. Circle Pines, MN: American Guidance Service.

Keith, T. Z. (1987). Assessment Research: An assessment and recommended interventions. School Psychology Review, 16, 2776-289.

Kim, J., & Mueller, C. W. (1978a). Introduction to Factor Analysis: What is it and how to do it, Newbury Park, CA: Sage.

Kim, J., & Mueller, C. (1978b). Factor Analysis: Statistical methods and practical issues. Beverly Hills, CA: Sage.

Kirby, P. C. (1987). Confirmatory factor extraction versus confirmatory factor rotation. Paper presented at the Annual Meeting of the Southwest Educational Research Association, Dallas, TX, January 30, 1987.

Kleinfeld, J., & Nelson, P. (1991). Adapting instruction to Native Americans' learning styles. Journal of Cross-Cultural Psychology, 22(2), 273-282.

Loehlin, J. C. (1992). Latent variable models: An introduction to factor, path, and structural analysis (2nd ed.). New Jersey: Erlbaum.

Long, J.S. (1983a). Confirmatory factor analysis: A preface to LISREL. Series no. 07-033. London: Sage.

Long, J. S. (1983b). Covariance structure models: An introduction to LISREL. Sage University Paper series on Quantitative Application in the Social Sciences, series no.07-034. London: Sage.

Marsh, H.W., Balla, J.R., & McDonald, R.P. (1988). Goodness-of-fit indexes in confirmatory factor analysis: The effect of sample size. Psychological Bulletin, 103, 391-410.

Martin, E. W. (1989). Lessons from implementing P.L. 94-142. In J. J. Gallagher, P. L. Trohanis, & R. M. Clifford (Eds.) Policy implementation and P.L. 99-457: Planning for young children with special needs (pp. 19-32). Baltimore, MD: Paul H. Brookes.

Mather, N., & Roberts, R. (1994). Learning disabilities: A field in danger of extinction? Learning Disabilities Research and Practice, 9, 49-58.

McDonald, R. P., & Marsh, H. W. (1990). Choosing a multivariate model: Noncentrality and goodness of fit. Psychological Bulletin, 107, 247-255.

McGrew, K. S. (1986). Clinical interpretation of the Woodcock-Johnson Tests of Cognitive Ability. Orlando: Grune & Stratton.

McGrew, K.S., Werder, J.K., & Woodcock, R.W. (1990). WJ-R technical manual. Allen, TX: DLM.

Messick, S. (1989). Validity. In R. L. Linn (Ed.), Educational Measurement (pp. 13-103). New York: American Council on Education & Macmillan.

Mishra, S. P. (1981). Factor analysis of the McCarthy Scales for groups of white and mexican-american children. Journal of School Psychology, 19, 178-182.

Mishra, S. P., & Lord, J. (1982). Reliability and predictive validity of the WISC-R with Native-American Navajos. Journal of School Psychology, 20, 150-154.

Mulaik, S. A., James, L. R., Van Alstine, J., Bennett, N., Lind, S., & Stilwell, C. D. (1989). Evaluation of goodness-of-fit indices for structural equation models. Psychological Bulletin, 105, 430-445.

Peter, P. A., & Smith P. O. (1993). WISC-III and WJ-R: Predictor and discriminant validity for students with severe emotional disturbance. Journal of Psychoeducational Assessment: WISC-III Monograph, 114-124.

Prasse, D. P., & Reschly, D. J. (1986). Larry P: A case of segregation, testing, or program efficacy? Exceptional Children, 52, 333-346.

Public Law 94-142, Education for All Handicapped Children Act, 1975.

Public Law 99-457, Education for All Handicapped Children Act Amendments, 1986.

Reschly, D. J. (1978). WISC-R factor structures among Anglos, Blacks, Chicanos, and Native-American Papagos. Journal of Consulting and Clinical Psychology, 46, 417-422.

Reschly, D. J. (1990). Found: Our intelligences: What to they mean? Journal of Psychoeducational Assessment, 8, 259-267.

Reschly, D. J. (1991). Bias in cognitive assessment: Implications for future litigation and professional practices. Diagnostique, 17, 86-90.

Reynolds, C. R. (1980). Differential construct validity of a preschool battery for blacks, whites, males, and females. Journal of School Psychology, 18, 112-125.

Reynolds, C. R. (1982). Methods for detecting construct bias and predictive bias. In A.R. Berk (Ed.), Handbook of methods for detecting test bias (pp. 190-227). Baltimore, MD: The John Hopkins University Press.

Reynolds, C. R., & Brown, R. T. (1984). Bias in mental testing: An introduction to the issues. In C. R. Reynolds & R. T. Brown (Eds.), Perspectives on bias in mental testing (pp. 1-40). New York: Plenum Press.

Salvia, J., & Yesseldyke, J. (1981). Assessment in special and remedial education. Boston: Houghton-Mifflin.

Samuda, R. J. (1975). The psychological testing of American minorities. New York: Harper & Row.

Sandoval, J. (1979). The WISC-R and internal evidence of testing bias with minority groups. Journal of Consulting and Clinical Psychology, 48, 919-927.

Sattler, J. M. (1988). Assessment of Children (3rd ed.). San Diego: Jerome M. Sattler, Publisher.

Serwatka, T., Dove, T., & Hodge, W. (1986). Black students in special education: Issues and implications for community involvement. The Negro Educational Review, 37, 17-26.

Shinn, M. Algozzine, B., Marston, D., & Ysseldyke, J. (1982). A theoretical analysis of the performance of learning disabled students on the Woodcock-Johnson Psycho-Educational Battery. Journal of Learning Disabilities, 15, 221-226.

Taylor, R. L. (1989-1990). Woodcock-Johnson Psycho-Educational Battery-Revised (WJ-R). Diagnostique, 15, 264-276.

Taylor, R. L. (1991). Bias in cognitive assessment: Issues, implications, and future directions. Diagnostique, 17, 3-5.

Taylor, R. L., & Richards, S. B. (1991). Patterns of intellectual differences of Black, Hispanic, and White children. Psychology in the Schools, 28, 5-9.

Taylor, R. L., & Ziegler, E. W. (1987). Comparisons of the first principal factor on the WISC-R across ethnic groups. Educational and Psychological Measurement, 47, 691-695.

Thompson, P., & Brassard, M. (1984a). Validity of the Woodcock-Johnson Tests of Cognitive Ability: A comparison with the WISC-R in LD and normal elementary students. Journal of School Psychology, 22, 201-208.

Thompson, P., & Brassard, M. (1984b). Cummings and Moscato soft on Woodcock-Johnson. School Psychology Review, 13, 41-44.

Thorndike, R. L., Hagen, E. P., & Sattler, J. M. (1986). Stanford-Binet Intelligence Scale-Fourth Edition. Chicago, IL: Riverside.

Valencia, R. R., & Rankin, R. J. (1986). Factor analysis of the K-ABC for groups of Anglo and Mexican-American children. Journal of Educational Measurement, 23, 209-219.

Washington, E. D., & McLoyd, V. C. (1982). The external validity of research involving American minorities. Human Development, 25, 324-339.

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

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

Wilen, D. K., & van Maanen Sweeting, C. (1986). Assessment of limited english proficient Hispanic students. School Psychology Review, 15, 59-75.

Wigdon, A. K., & Garner, W. R., (1982).(Eds.) Ability testing: Uses, consequences, and controversies: Part I Report of the committee. Washington, DC: National Academy Press.

Woodcock, R., & Johnson, M. (1977). Woodcock-Johnson Psycho-Educational Battery. Allen, TX: DLM.

Woodcock, R. (1984a). A response to some questions raised about the Woodcock-Johnson: The mean score discrepancy issue. School Psychology Review, 13, 342-354.

Woodcock, R. (1984b). A response to some questions raised about the Woodcock-Johnson: Efficacy of the aptitude clusters. School Psychology Review, 13, 355-362.

Woodcock, R. W. (1990). Theoretical foundations of the WJ-R measures of cognitive ability. Journal of Psychoeducational Assessment, 8, 231-258.

Woodcock, R. W., & Mather, N. (1989). Woodcock-Johnson Tests of Cognitive Ability. Examiner's Manual. Allen, TX: DLM.

Ysseldyke, J. (1985). Woodcock-Johnson Psycho-Educational Battery. In J. V. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 1762-1764). Lincoln, NE: University of Nebraska Press.

Ysseldyke, J. E. (1990). Goodness of fit of the Woodcock-Johnson Psycho-Educational Battery-Revised to the Horn-Cattell Gf-Gc theory. Journal of Psychoeducational Assessment, 8, 268-275.

Ysseldyke, J. E., Algozzine, B., & Shinn, M. (1981). Validity of the Woodcock-Johnson Psycho-Educational Battery for learning disabled youngsters. Learning Disability Quarterly, 4, 244-249.

Ysseldyke, J., Shinn, M., & Epps, S. (1981). A comparison of the WISC-R with the Woodcock-Johnson Tests of Cognitive Ability. Psychology in the Schools, 18, 15-19.