AN ANALYSIS OF THE RELIABILITY AND VALIDITY OF THE NAGLIERI NONVERBAL ABILITY TEST (NNAT) WITH ENGLISH LANGUAGE LEARNER (ELL) MEXICAN AMERICAN CHILDREN

A Dissertation

by

CARLO ARLAN VILLARREAL

Submitted to the Office of Graduate Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

May 2005

Major Subject: School Psychology

AN ANALYSIS OF THE RELIABILITY AND VALIDITY OF THE NAGLIERI NONVERBAL ABILITY TEST (NNAT) WITH ENGLISH LANGUAGE LEARNER (ELL) MEXICAN AMERICAN CHILDREN

A Dissertation

by

CARLO ARLAN VILLARREAL

Submitted to Texas A&M University
in partial fulfillment of the requirements
for the degree of

DOCTOR OF PHILOSOPHY

Approved as to style and content by:

_____	_____
Salvador Hector Ochoa (Co-Chair of Committee)	Michael J. Ash (Co-Chair of Committee)
_____	_____
Victor Willson (Member)	Rogelio Saenz (Member)

_____

Michael R. Benz
(Head of Department)

May 2005

Major Subject: School Psychology

ABSTRACT

An Analysis of the Reliability and Validity of the Naglieri Nonverbal Ability Test (NNAT) with English Language Learner (ELL) Mexican American Children.

(May 2005)

Carlo Arlan Villarreal, B.A., Baylor University

Co-Chairs of Advisory Committee: Dr. Salvador Hector Ochoa
Dr. Michael J. Ash

The purpose of this study was to investigate the reliability and validity of the results of the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997a) with a sample of English Language Learner (ELL) Mexican American children and to compare the performance on the NNAT of 122 ELL Mexican American children with children from the standardization sample. The rationale for conducting this study was the need to identify culturally sensitive and technically adequate nonverbal measures of ability for the fastest growing minority group within America's public schools today, Mexican American children. The NNAT was administered to participants with parental consent. Statistical analyses of the scores did yield positive evidence of internal consistency for the Nonverbal Ability Index (NAI) total score of the NNAT. However, when individual clusters were analyzed, Pattern Completion, Reasoning by Analogy, and Serial Reasoning did not yield positive evidence of internal consistency. Only Spatial Visualization approached the reliability standard deemed acceptable for tests of cognitive ability. The mean differences of the NNAT scores between two independent groups were also assessed in the present study. Results of the statistical analyses did not yield statistically significant differences across age and grade factors between the scores of the ELL Mexican American sample and the standardization sample. Finally, the proposed factor structure of the NNAT was compared with the factor structure found with the ELL Mexican American sample. Goodness-of-fit test statistics indicate that the proposed four-factor structure does not fit well with the data obtained from this sample of ELL Mexican American students. Furthermore, although the NNAT is considered to be a

unidimensional test of general ability, nine factors were extracted upon analysis, providing evidence that the items on each of the four clusters do not function together as four distinct dimensions with this ELL Mexican American sample. Given that the individual clusters that collectively combine to yield the NAI total score are not based on any particular model of intelligence, interpretation of specific strengths and weaknesses should be discouraged. Finally, the NNAT's overall score should be interpreted with caution and may best be used in conjunction with multidimensional ability and/or intelligence measures.

ACKNOWLEDGMENTS

It would have been impossible to conduct this investigation and write this dissertation without the assistance and support that I obtained from many people. First of all, I would like to thank the members of my dissertation committee: Dr. Salvador Hector Ochoa, Dr. Michael Ash, Dr. Victor Willson, and Dr. Rogelio Saenz. Thank you for guiding me through this process and for providing me with support when it was requested of you. I would especially like to thank Dr. Willson for his assistance with the statistical analyses portion of my dissertation as well as Dr. Ochoa for offering his added encouragement throughout the duration of my doctoral studies. Another Rio Grande Valley native has been dispersed into the professional ranks of psychology, in part, because of you. Mil gracias.

I would like to thank the executive director of elementary education at the school district who arranged the initial meeting with the four elementary school principals that graciously allowed me to conduct the assessments in their schools. I would also like to thank the four elementary school principals for allowing me to conduct my assessments in their schools. In addition, I would like to thank the 3rd and 4th grade teachers and the respective administrative staff of each elementary school who facilitated the testing process during school hours. I especially appreciate the cooperation of the parents and children who participated in the study. Without them, none of this would have been possible.

I would also like to express my sincere appreciation to two friends, Romeo Zuniga and Jonathan Conrad. Your friendship and support during the data collection phase of my research study and the final writing phase of my dissertation have proven invaluable to me. I am indebted to you both.

And last, I would also like to thank members of my family who have offered their support in so many ways throughout my years of educational training.

Neena, although you are my youngest sister, you often displayed your motherly instincts to me during my stressful times by always inquiring whether I was eating

enough and doing all you could to ensure that I was. Although your subtle undertones sometimes communicated to me how crazy you thought I was for pursuing 13 years of post high school education, you always remained steadfast in supporting your Bubba. Thank you.

Grandma, thank you for the constant encouragement and morale boosters that you offered me throughout my educational endeavors. Your spiritual guidance and support, through years of prayer and advice, have helped me to remain focused on what is important in life. And through your 90 years of living, Grams, there is still a lot I stand to learn from you.

Mom, you have provided me with a living example of how perseverance, stamina, and will power can provide one with the necessary means to achieve any goal. Your high expectations for me, beginning with the look you gave me when I received my first non 'A' in high school that instilled in me a desire to want to hide subsequent report cards from you, fostered a drive within me to always want to strive for better. I also want to thank you for encouraging me, at the age of 20, to "pursue a more lucrative career" when I communicated to you my initial aspirations for the theater. It was at this moment that I was eventually catapulted into the field of psychology. And Mom, 13 years later, there are no regrets. I think you had some foresight there.

Dad, although physically you are no longer with us today, your presence in my life is just as real. With each difficult decision that I am confronted with in my life, I will forever ask myself, "What would Dad do?" Thank you for encouraging me to resist the temptation of wanting to rush through life in pursuit of something higher, something better, by teaching me the importance of enjoying life, helping me learn to enjoy the present.

And finally, I cannot end without acknowledging my God. You are and always will be my One constant.

Mom and Dad, this dissertation and doctorate degree are dedicated to you both.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER I

INTRODUCTION

Hispanic Americans are currently comprised of 43.8 million people and represent the fastest growing minority group in the United States today. According to the U. S. Department of International Information Programs (2004), population estimates indicate that Hispanic Americans are projected to account for one in every four Americans by the year 2050, totaling 102.6 million people. With varying acculturation levels, Hispanic Americans often exhibit cultural and linguistic factors that are as diverse as are their experiences within mainstream America. It is estimated that approximately 29 million United States residents age 5 and older speak Spanish at home, constituting a ratio of more than 1 in every 10 United States residents (U. S. Department of State International Information Programs, 2004). For the Hispanic child, these various cultural and linguistic factors often manifest themselves in different language proficiency levels in English and Spanish. As such, some Hispanic children will subsequently be identified as English Language Learners (ELL) upon entering into America's public school system due to their lack of facility, fluency, or linguistic competence in English as a second language relative to a native speaker-listener of the language (Kretschmer, 1991). In a survey prepared by Kindler (2002) for the United States Department of Education, he found that when state-by-state comparisons were made regarding the linguistic diversity of America's public school student enrollment, California (1,512,655) enrolled the largest number of public school students identified as ELL, followed by Puerto Rico (612,121), Texas (601,791), Florida (290,024), New York (266,774), Illinois (140,528), and Arizona (135,503). Among those identified as ELL, 79.2% were Hispanic and spoke Spanish followed by Vietnamese (2%), Hmong (1.6%), Cantonese (1%), and Korean (1%).

This dissertation follows the style and format of *School Psychology Review.*

Unfortunately, children from non-English speaking backgrounds are at a higher risk for inappropriate assessment upon entering into America's public school and/or mental health systems. Insufficiently trained testing personnel with regard to bilingual assessment, a lack of testing personnel proficient in the child's native language, and a lack of instruments with adequate norms to evaluate our Hispanic American children are three of the most often cited reasons for inappropriate assessment of Hispanic American youth (Figueroa & Hernandez, 2000).

Past critics of the use of psychometric tests with culturally and linguistically diverse populations have argued that the normative framework on which most test scores have been based has often assumed a high degree of experiential homogeneity, cultural and linguistic similarity, and equity in learning opportunities among test takers (Heller, Holtzman, & Messick, 1982). Under these conditions, the results of an administered test truly become a measure that solely belongs to the individual and his or her abilities. However, given that the United States is one of the most pluralistic and culturally diverse societies in the world, these same tests would work best in a perfect democracy of monolingual and monocultural citizens (Figueroa & Hernandez, 2000).

Hispanic Americans, in particular, pose a significant challenge to these aforementioned assumptions. Because of the vast within-group differences that exist within the Hispanic American population, coupled with the more studied between-group differences with white middle class America, the assumptions of tests concerning homogeneity may very well be untenable for this population. Hispanic Americans, for one, have varying levels of exposure to and demonstrate different proficiency levels in Spanish and English. Their cultural experiences in the United States are multigenerational and reflect on vast intra-group differences regarding their socioeconomic status, acculturation levels, and linguistic proficiencies. Yet, according to Figeroa & Hernandez (2000), "Hispanic students are tested everyday and are compared to middle class America in the unique reification of democracy and assimilation that tests impose" (p. 1). Crucial decisions concerning grade retention and

promotion as well as determining eligibility for special services and programs are continually made based on the results of a test. Consequently, although one cannot deny that tests can and often do accurately reflect existing differences in performance between students, those who have not had the cultural experiences of middle class America can be placed at a serious disadvantage in taking standardized tests (Figueroa & Hernandez, 2000).

One of the disadvantages for Hispanic American students with regards to assessment practices is reflected in their tendency to be over-diagnosed with learning disabilities due, in part, because assessors are not considering level of English and/or Spanish language proficiency before assessing them with English language-loaded tests. Cummins (1984) stated that one of the most serious problems with the assessment of ELL students who are referred for special education testing is that they frequently are not identified as ELL students prior to the assessment. What the data resulting from these assessments then reflect are not accurate estimates of the child's true ability but rather a lack of proficiency in their second language. Chamberlain & Medeiros-Landurand (1991) suggested that assessing proficiency in both languages is essential in order to determine whether the ELL or bilingual student's academic struggles are due to an inherent disability or whether such difficulties reflect normal second language acquisition. Furthermore, Willig (1986) stated that a true disability must be evident in both languages for a child labeled as a second language learner to be identified as having a learning disability. Therefore, along with the fact that appropriate interpretation by trained assessors of test performance is particularly important for culturally or linguistically different students, there is also a great need for psychometrically sound and appropriate instruments that assess the true abilities of linguistically and culturally diverse student populations (Figueroa & Hernandez, 2000).

One way to help ensure the appropriate psychological and/or psycho-educational assessment of linguistically and culturally diverse students is to utilize an instrument that is both free of linguistic factors and culturally neutral. The testing community has responded to attenuate test bias with this population by suggesting the use of nonverbal

tests of mental ability with these diverse school populations. To this day, the use of nonverbal measures to assess the intellectual abilities of linguistically diverse students is generally regarded as an acceptable practice and is a practice frequently utilized by school psychologists across the country (Clarizio, 1982; Ochoa, Powell, & Robles-Pina, 1996). One of the most recently marketed nonverbal tests of general ability is the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997a). Given the technical information provided in the test manual citing positive results of the NNAT with the standardization sample, which included an ethnically and linguistically diverse sample, investigating the psychometric properties of the NNAT with ELL Mexican American children, the largest of the Hispanic subgroups living in the United States, appears to be both promising and needed. Thus, in particular, there is a need to investigate if the NNAT is a valid and reliable, and therefore appropriate instrument to use with ELL Mexican American children.

Statement of the Problem

Policy regarding the nondiscriminatory assessment of linguistically and culturally diverse students has brought about an increased awareness regarding test bias among testing personnel throughout the United States. The National Association of School Psychologists (1997) standards for the provision of school psychological services recommend that, with regard to non-biased assessment techniques, multifaceted assessment batteries should be used that include a focus of the student's strengths and that the data derived from assessments should be interpreted in the context of the student's sociocultural background and the setting in which he/she is functioning. The 1999 *Standards for Educational and Psychological Testing* volume by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) also recommends that testing practice should be designed to reduce threats to reliability and validity of test score inferences that may arise from language differences, that the test should be administered in the test taker's most proficient language unless proficiency in the less proficient language is part of the assessment, and inferences about test takers'

general proficiency should be based on tests that measure a range of language features, and not a single skill. These and other standards have helped to create a sense of uniformity regarding testing procedure with minority populations. However, the fact is that the gap between policy and what is actually practiced remains, contributing to the continuing disproportionate number of minority students being classified as learning disabled, mentally retarded, and emotionally disturbed.

Mexican Americans comprise one linguistically and culturally diverse group that has historically been misdiagnosed as a result of inappropriate assessment. Along the U.S.-Mexico border at the start of each school year, an influx of recently immigrated Mexican children pose unique challenges to the United States public school system. With such vast differences that may exist within the Mexican American population with regard to levels of education, English and Spanish language proficiency levels, and intellectual abilities, it is paramount that assessment personnel understand and take into account the many cultural and linguistic factors involved in testing diverse populations. In an effort to accurately assess these children's true ability, appropriate nonverbal intelligence and/or ability test measures are warranted. The NNAT is one test that seeks to meet the unique assessment needs testing personnel face with diverse student populations. Because of its unidimensional theoretical orientation, the NNAT is ideal for screening for potential learning difficulties as well as identifying for giftedness. Its' large scale group administered format also make it both time and cost efficient choices for school districts to utilize in their assessment of linguistically and culturally diverse students.

Naglieri & Ronning (2000), in their study that compared performances of white, African American, Hispanic, and Asian children on the NNAT, suggested that researchers should examine differences more closely with other populations such as those with limited English language skills. This recommendation, coupled with the fact that there are no reported validity studies in the technical manual that specifically target special populations, speaks to the need for such studies to be conducted. Therefore, the

purpose of the present study is derived from this need to investigate the reliability and validity of the NNAT with ELL Mexican American children.

Research Questions

1. What is the internal consistency of the four clusters and the Naglieri Ability Index (NAI) total score of the NNAT with ELL Mexican American children? Are the obtained reliability coefficients technically adequate? Are they comparable to those obtained with the standardization sample?
2. How does the performance of the sample of ELL Mexican American children differ from the standardization sample with respect to the four cluster scores and the overall Nonverbal Ability Index (NAI) raw score of the NNAT? Are the differences between the means of the two groups statistically significant?
3. How does Naglieri's proposed four-factor structure for the NNAT compare with the factor structure found with ELL Mexican American children?

The current study was undertaken to address these three questions. Before reporting on the study itself, the relevant literature will be reviewed. Methodology will be described in Chapter III; results will be presented in Chapter IV. Finally, in Chapter V, the implications of these results for the intellectual assessment of Mexican American children will be discussed.

CHAPTER II

LITERATURE REVIEW

This chapter provides a review of research regarding the assessment of intelligence, specifically targeting the following topics: history of intelligence testing with Hispanic populations, the need for bilingual assessment in the United States, the overrepresentation of minorities in special education, and a review of the current nonverbal intelligence and/or ability tests available to testing personnel throughout the United States.

History of Intelligence Testing in the United States with Hispanic Populations

Historically, the testing community has fallen short in adequately addressing the unique linguistic, cultural, and educational needs of the Hispanic American student population. Perhaps the roots of disregard for this minority population can be traced back to popular ideas of the late 19th century. Sir Francis Galton's views on hereditarianism and his racist pronouncements about individual differences between whites and people of color began to enter the American discourse of the intelligence testing movement in the United States and influence scholars, researchers, and test developers alike in their approaches to assessing diverse populations. By the early 20th century, Lewis Terman had popularized the (Alfred) Binet-(Theodore) Simon scale of intelligence in the United States and by 1916 had, according to historical analysis, been translated, culturally appropriated, psychometrically modified, and normed by American psychologists (Gould, 1981). However, the standardization sample of 1,000 children was almost exclusively white (Western European descent) and middle class and was not representative of children of different cultural, social, and linguistic backgrounds. This deliberate exclusion of Mexican American, African American, and other children of color from the standardization sample would be criticized for nearly six decades after the original Standford-Binet intelligence was developed before minority children would be

included in the standardization samples of subsequent revisions to the test (Valencia & Suzuki, 2001).

The 1920's produced a great amount of research on ethnic minorities as psychologists began to recognize the need to better understand the differences between white America and these populations with regards to differential test performance. Unfortunately, many of the studies came under the title of "Race Psychology", and were conducted on Mexican American, African American, and American Indian children, and reflected a naïve use of test scores to support genetic arguments about lower intellectual potential in children of color. Numerous investigations of this period failed to control for key variables such as socioeconomic status and did not fully consider the possibility of language confoundment (i.e., lower intellectual performance of the Mexican American children was due, in part, to limited English proficiency). For example, Garretson's (1928) study (as cited in Valencia & Suzuki, 2001) sought to explain the causes of "retardation" among Mexican American children attending school in a small public school system in Arizona by comparing 197 "American" and 117 "Mexican" children enrolled in 1st through 8th grade. Garretson's (1928) study showed that the median IQ for each of the verbal and nonverbal tests administered was higher for white children at nearly all grade levels. When asked to explain the depressed scores obtained by the Mexican American group, Garretson suggested the underlying cause to be innate, even though socioeconomic status and level of English language proficiency were not controlled for nor had all the results been reported. Specifically, Garretson's (1928) study also showed that Mexican American children actually outperformed their white counterparts in two grade levels and on two different tests. However, it appeared that to have explained the unexpected (i.e., Mexican American outperforming white children) would have proven to be a difficult task when hereditarianism was the zeitgeist. So although this era saw an augmentation of research on diverse populations, most utilized faulty research methods and aimed at supporting the eugenic philosophy regarding race differences of intelligence that permeated American society at that time (Valencia & Suzuki, 2001).

It wasn't until the 1930's that the Mexican American population began to see scholars, researchers, and other professionals of Mexican descent that sought to address some of the complex issues of intelligence testing with minority populations. At the fore of this new wave of minority research was George Sanchez, a Mexican American psychologist, who aimed to investigate the culture and language variables and their potential influence on the minority child's intelligence test scores. As a result, George Sanchez sought to address two fundamental questions that have challenged and continue to this day to perplex test-makers, test-givers, and test-users alike. First, does Spanish usage in the home or as the primary language affect test scores? And second, do any aspects of Hispanic culture in the United States attenuate or change test outcomes? In his 1934 article entitled "Bilingualism and Mental Measures: A Word of Caution" (as cited in Figueroa & Hernandez, 2000), George Sanchez addressed these questions:

> Is the fact that a child makes an inferior score on an intelligence test prima facie evidence that he is dull? Or is it a function of the test to reflect the inferior or different training and development with which the child was furnished by his home, his language, the culture of his people, and by his school? The school has the responsibility of supplying those experiences to the child which will make the experiences sampled by standard measures as common to him as they were to those on whom the norms of the measures were based. When the school has met the language, cultural, disciplinary, and informational lacks of the child and the child has reached a saturation point of his capacity in the assimilation of fundamental experiences and activities, then failure on his part to respond to tests of such experiences and activities may be considered *his* failure.

Padilla & Aranda (1974) summarized additional ground-breaking issues that George Sanchez addressed in his 1934 article that would serve as beacons of reform in test development, individual assessment, and social concerns involving test use with minority populations to this day. According to Padilla & Aranda (1974), Sanchez stated

that: 1) tests are not standardized on the Spanish-speaking population of this country, 2) test items are not representative of the Spanish-speaking culture, 3) the entire nature of intelligence still is a controversial issue, 4) test results from Spanish-speaking individuals continue to be accepted uncritically, 5) revised or translated tests are not necessarily an improvement on test measures, 6) the influence of testing on the educational system is phenomenal, and 7) attitudes and prejudices often determine the use of test results. Interestingly, it is this last issue that provided the roots for the second phase of the testing-with-minority-populations controversy to develop in the United States.

During the 1940's and early 1950's, group intelligence testing in the nation's public schools became a routine practice and one relatively free of controversy. As a result, and partially due to some scholars who were absent from the world of research as they served in the armed forces in World War II and the Korean War, there was an overall decline and inactivity in research and publishing on intelligence. However, by 1954, issues of intelligence testing with minority populations would be thrust into the controversial debates once again as the legality of certain uses of intelligence test results would being questioned. *Brown v. Board of Education* (1954) recognized that educating any "class of children" separately, even if done in equal facilities, was intrinsically unequal because of the stigma attached to segregation and because of the denial of association with children from other classes. However, as Bersoff (1982) noted, though the courts ruled that intelligence and achievement tests were used unconstitutionally against African Americans in the goal of segregation, it never determined that the tests themselves were actually invalid. It wouldn't be until the advent of the civil rights movement during the late 1950's and early 1960's that the issue of deriving valid test results by utilizing valid testing practices with minority populations would be addressed. First, the practice of using group intelligence tests in the curricular assignments of minority students would be abolished (*Hobson v. Hansen,* 1967), schools would be required to test a language minority child in his or her native language (*Diana v. State Board of Education,* 1970), and multifaceted evaluations that included the use of

nonverbal measures if the child's native language was not English, assessment of adaptive behavior, and an interview with the parents in the child's home would be required (*Guadalupe Organization v. Tempe Elementary School District,* 1972). Each of these court cases would eventually have its influence on the evolving set of regulations and policies on the development and use of tests in the United States with diverse populations (Valencia & Suzuki, 2001).

The most important set of regulations and policies on the development and use of tests in the United States are the *Standards for Educational and Psychological Testing* volume by the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME). The first set of these standards entitled *Standards for Educational and Psychological Tests and Manuals* (AERA, 1966) reflected on one overriding goal. This goal made it essential for test manuals to carry information sufficient to allow any qualified user to make sound judgments regarding the usefulness and interpretation of the test. However, very few references touched on linguistic and cultural issues and the few that did included such recommendations as: 1) if the validity of the test is likely to be different for subsamples, then the manual should report the results for each subsample separately and 2) demographic information such as age, sex, socioeconomic level, and minority group membership should be given in the manual (Figueroa & Hernandez, 2000).

The second edition of the *Standards for Educational and Psychological Testing* (AERA, 1974) reflected the results of litigation regarding testing issues with minority populations during that period. For example, it suggested that the overrepresentation of African American and Spanish-speaking children "with limited cultural exposure" in Special Education was caused by test users' lack of knowledge about the limitations of tests when cultural differences existed. Therefore, test users were instructed to know the research literature on tests and testing particularly with respect to the problems associated with testing individuals with "limited or restricted cultural exposure". In response to the limited number of culturally sensitive test measures available at the time,

the 1974 standards urged that efforts to solve educational and psychological problems should not be abandoned simply because of the absence of an appropriate standardized instrument, but rather recommended that linguistically and culturally appropriate alternative ways to assess should be utilized. The most salient weaknesses in these standards, however, were not the recommendations themselves but rather the lack of detailed instruction on how to sufficiently meet those standards (Figueroa & Hernandez, 2000).

In the third edition of the *Standards for Educational and Psychological Testing* (AERA, 1985), an attempt was made to address, more than ever before, the challenges culture, language, disability, and socioeconomic status posed for the testing community. It stated that for a non-native English speaker or for a speaker of some dialects of English, every test becomes, in part, a language or literacy test. As a result, these tests contained an unknown, systematic degree of error that consequently rendered them biased for English language learners. In response to this problem, the issues of translating tests and the use of interpreters were introduced. However, regarding the former, the standards explicitly stated that simply translating the test items would not ensure equality in item difficulty in the other language and would require empirical evidence to deem it a valid and reliable measure. Regarding the use of interpreters, which included those that translated the test into the primary language during the assessment process or who helped administer a test that had already been translated, the issue was never addressed. In addition, the standards were silent on the apparent inability of tests and test users to differentiate among cultural factors, language proficiency levels, and mental/emotional disabilities. However, tests for determining English language proficiency were recognized as necessary for making educational placement decisions with linguistically diverse children (Figueroa & Hernandez, 2000).

A more in-depth analysis of the many concerns impacting the assessment of linguistically and culturally diverse populations appeared in the most recent edition of *Standards for Educational and Psychological Testing* (AERA, 1999). However, recommendations made in this volume appeared to reflect more on testing

accommodations for linguistically and culturally diverse individuals rather than on modifications to the current use of culturally sensitive assessment practices. For example, recommending test accommodations for English language learners such as: 1) using only sections of the test that match the linguistic proficiency of the test-taker, 2) changing the test and response formats, and 3) allowing for more time to take the test does not meet the need of making the test more culturally or linguistically appropriate. Similarly, stating that the tester must assume responsibility of the interpreter, if one is used, without prescribing an empirically validated model for training interpreters ensures that standardization requisites, psychometric properties, and the interpretation of test scores will be negatively affected (Figueroa & Hernandez, 2000).

Because individuals of diverse linguistic backgrounds present such unique challenges to testing personnel, the *Standards for Educational and Psychological Testing* (AERA, 1999) manual devoted a chapter to address some of the unique assessment needs of linguistically diverse individuals. For example, with regards to testing bilingual individuals, the manual discusses the notion of a bilingual continuum that can make establishing language proficiency levels in English and/or Spanish a difficult task. For instance, children whose parents speak Spanish may be able to understand Spanish but express themselves better in English. In addition, the distinction is made between conversational and academic language in that although some bilingual individuals may prefer to use their native language in social situations, they may choose to use English primarily in academic or work-related activities. Similarly, other individuals familiar with two languages may perform more slowly, less efficiently, and at times less accurately on problem-solving tasks that are administered in the less familiar language. Therefore, the provisions of the standards recommend that an understanding of an individual's type and degree of bilingualism be sought by assessors to help ensure proper test usage and urge that because language dominance is not necessarily an indicator of language competence in taking a test, accommodations may be necessary even when administering the test in the more familiar language (Figueroa & Hernandez, 2000).

In addition, the *Standards for Educational and Psychological Testing* (AERA, 1999) offered provisions in the use of interpreters by stating in Standard 9.11 that, ideally, assessments of individuals identified as ELL should be conducted by a professionally trained bilingual examiner. However, in the event that a bilingual examiner is not available, the provisions state that an interpreter may be used. According to the provisions of the standards, the interpreter needs to be fluent in both languages of the test and the examinee's native language as well as have a basic understanding of the process of psychological and educational assessment. Specifically, according to the provisions of the standards, the interpreter should also understand the importance of accurately conveying to the examiner an examinee's actual responses as well as become familiar with details of the test content and administration prior to testing to ensure fluidity of the testing process.

However, regarding norming issues and establishing the reliability and validity of tests for bilingual Hispanic children, Figueroa and Hernandez (2000) state that bilingual children cannot validly be compared against norms for children whose linguistic experience and development is with only one language. Specifically, in accordance with Standard 9.2, they state that bilingual children need norms derived from bilingual samples, controlling for differential levels of linguistic proficiencies. Additionally, they recommend that when attempting to establish the psychometric properties of a test with bilingual Hispanic children, the validity and reliability of the test must be established within different levels of linguistic proficiencies within different Hispanic cultural groups. This underscores the need to address the many intra-group linguistic and cultural differences within the Hispanic population. And although these current standards sought to ameliorate some of the most pressing issues facing assessment practices with bilingual, Hispanic populations, the unfortunate fact is that these standards have outdistanced current test technology and testing practices with individuals of differing linguistic backgrounds, a concern that is still with us today (Figueroa & Hernandez, 2000).

Need for Bilingual Assessment in the United States

Public Law 94-142 (Education for All Handicapped Children Act of 1975) mandated that nondiscriminatory assessment practices be utilized when assessing students from culturally and linguistically diverse backgrounds. Given that the population of students identified as ELL is growing at exponential rates in the United States speaks to the need for fair and accurate assessment practices to be conducted in the child's native language. According to a report compiled by Kindler (2002) for the United States Department of Education, California (n=1,512,655) enrolled the largest number of public school students identified as ELL, followed by Puerto Rico (n=612,121), Texas (n=601,791), Florida (n=290,024), New York (n=266,774), Illinois (n=140,528), and Arizona (n=135,503). Upon closer examination, the states with the highest percentages of ELL students were California (25%), New Mexico (19.9%), Arizona (15.4%), Alaska (15.0%), Texas (14.0%), and Nevada (11.8%). In the same report, they state that the United States reported over 460 languages spoken by ELL students nationwide.

As discussed in Chapter I, among those identified as second language learners, 79.2% were Hispanic and spoke Spanish followed by Vietnamese (2%), Hmong (1.6%), Cantonese (1%), and Korean (1%). Among other language groups with more than 10,000 speakers within America's public schools include Arabic, Armenian, French, Haitian Creole, Hindi, Japanese, Navajo, Porteguese, Serbo-Creole, and Tagalog. However, there are substantial regional variations in linguistic diversity. For instance, in many states such as Alaska, Hawaii, Maine, Minnesota, Montana, South Dakota, North Dakota, and Vermont, Spanish is not the dominant language among the school-aged ELL population. Similarly, while Vietnamese ranked second nationally, many states show other languages rank second: Arabic is second in Illinois, Ohio, West Virginia, and Michigan; Korean in Maryland; Lao in Arkansas; Native American languages in Arizona, New Mexico, Idaho, Montana, Oklahoma, and Utah; Portuguese in Connecticut, Massachusetts, New Jersey, and Rhode Island; Haitian Creole in Delaware,

Florida, and the Virgin Islands; Chinese in New York and Kentucky; and Tagalog in Nevada, Guam, and Palau (Kindler, 2002).

In addition, according to the United States Department of State International Information Programs (2004), the current Hispanic population is estimated to be 43.8 million. Among this population, 40% are said to have been foreign-born, 52% of which are purported to have entered into the United States between 1990 and 2002. In addition, 9.9 million foreign-born individuals are said to have been born in Mexico, followed by Cuba (n=887,000), Dominican Republic (n=654,000), Columbia (n=566,000), and Guatemala (n=511,000). Hence, this diversity of language backgrounds in ELL students across the country has major implications for the assessment practices with linguistically diverse students, however, none more evident than with the Spanish speaking ELL population (Kindler, 2002). These demographic figures further emphasize the need for psychometrically sound and language/culture fair instruments to assess the needs of this ever-growing culturally and linguistically diverse population.

Research indicates that utilizing conventional tests that require the linguistically diverse child to read and respond to a test in English, hence relying on verbal ability, is often inadequate. Cummins (1984) states that these results often do not accurately estimate this child's true potential. As a result, nonverbal ability testing is an emerging testing technology that for some holds promise, particularly given the critical need to assess school ability for a rapidly growing diverse student population (Anastasi & Urbina, 1997).

However, obtaining valid school ability information for linguistically diverse students does present significant challenges for assessors. One common mistake assessors and educators make is failing to determine level of proficiency of the child's first and second language before proceeding with testing. Cummins (1980) has proposed two levels of language proficiency: Basic Interpersonal Communication Skills (BICS) and Cognitive Academic Language Proficiency (CALP). According to Cummins (1980), BICS includes the ability to participate in complex, context-embedded, face-to-

face communication and generally takes two years to acquire. CALP, on the other hand, requires the ability to understand and produce language typical of academic instruction, has a higher cognitive load than BICS, and typically takes about five to seven years to attain. This has great implications when assessing children who have recently immigrated to the United States. Often educators mistake a child's ability to engage in casual conversation as proof that challenging academically related tasks can be completed. However, Cummins (1980) states that in order for a child to be able to master a second language, it is necessary that the child attain CALP in his native language first.

The need for school psychologists who are trained in the area of bilingual assessment and are both aware and recognize the impact socioeconomic status, language proficiency, bilingualism, and culture can have on those tested with instruments standardized on white middle class children cannot be overstated. Without these fair instruments and trained personnel to execute sound and appropriate assessment procedures, ELL children will continue to be misunderstood, misclassified, and overrepresented in special education classrooms.

Overrepresentation of Minorities in Special Education

Section Four of the exclusionary clause safeguard included in Public Law 94-142 states that a child should not be labeled as learning disabled if the "discrepancy between ability and achievement is primarily the result of environment, cultural, of economic disadvantage" (U.S. Department of Education, 1977, p. 65083). Two amendments are included in the Individuals with Disabilities Education Act Amendments of 1997 (Individuals with Disabilities Education Act Amendments, 1997) that have significant implications for those assessing linguistically diverse populations. The two requirements, which are included in section 614 (b)(5) of this law, state: "In making a determination of eligibility under paragraph 4(A), a child shall not be determined to be a child with disability if the determinant factor for such determination is lack of instruction in reading or math or limited English proficiency"(PL 105-17).

However, in a study conducted by the Civil Rights Project at Harvard University (Parrish, 2002), it was found that racial inequity still persists within special education programs throughout the country. For example, Hispanic children are 10% more at risk in California, 22% more at risk in New York, 27% more at risk in New Mexico, 20% more at risk in Arizona, and 23% more at risk in Texas of being identified as having a specific learning disability. In contrast, they are 43% more at risk in Delaware and Pennsylvania, 38% more at risk in Connecticut and Minnesota, 37% more at risk in Virginia, and 36% more at risk in Utah of being identified as having a specific learning disability. In addition, Hispanic children are 98% more at risk in Rhode Island, 72% more at risk in New York, 89% more at risk in Colorado, 68% more at risk in Washington, 66% more at risk in New Mexico, and 65% more at risk in Arizona of being identified as mentally retarded.

One often cited factor contributing to this overrepresentation pertains to the assessors' inability to distinguish between a true handicapping condition and whether the bilingual child's speech and language behavior is simply characteristic of a student learning English as a second language (Maldonado-Colon, 1986). Carrasquillo (1991) states that "available assessment procedures do not provide adequate information to distinguish characteristics of second language learners from those of handicapped students" (p. 19). In addition, there is a shortage in the number of bilingual assessors throughout the country who are available to competently conduct bilingual assessments (Ochoa et al., 1996). And although the testing community has taken measures to ameliorate this issue of overrepresentation (i.e., litigation, utilizing more culture/language appropriate assessment tools), the concern remains.

It appears that the main reason for this continuing overrepresentation of bilingual children in special education classes is biased assessment practices (Carrasquillo, 1991). Jones (1976) believes that bias is involved at three distinct levels: 1) at the content level where the decisions are first made about what items to include in a test, 2) at the level of standardization where decisions are made about the population for whom the test is appropriate, and 3) at the point where efforts are undertaken to determine whether or not

tests accomplish what they have been designed to accomplish. In response to these issues, several court cases made public this problem of bias assessment practices with regards to minority children.

Among the most significant cases relating to the improper classification and placement of ELL children into special education programs and classes are *Arreola v. Board of Education* (1968), *Diana v. State Board of Education* (1970), and *Covarrubias v. San Diego Unified School District* (1971). In the *Arreola* case, the due process rights of the parents and children to have a hearing before placement into classes for the educable mentally retarded (EMR) was established. In the *Diana* case, the defendants agreed to: 1) test in the child's primary language, 2) use nonverbal tests, and 3) collect and use extensive supporting data. In the *Covarrubias* case, the requirement of informed parental consent before EMR placement was established (Baca & Cervantes, 1998). According to Casso (1973), each of the three lawsuits had the plaintiff children re-tested and found that they were not retarded and should never have been placed into EMR classes. In addition, they reiterated to the testing community how damaging misplacement can be for the children tested. However, once these concerns were made public, educators and legislators alike began to take steps toward improving the testing process and the classification procedures for all children, particularly the ELL child. In an excerpt from a manual distributed by the U.S. Office of Education, Office of Bilingual Education, several guidelines of 1974 reflected such growing sentiment:

> A procedure also should be included in terms of a move toward the development of diagnostic prescriptive techniques to be utilized when for reasons of language differences or deficiencies, non-adaptive behavior, or extreme cultural differences, a child cannot be evaluated by the instrumentation of tests. Such procedures should ensure that no assessment will be attempted when a child is unable to respond to the tasks or behavior required by a test because of linguistic or cultural differences unless culturally and linguistically appropriate measures are administered by qualified persons. In those cases in which appropriate

> measures and/or qualified persons are not available, diagnostic-prescriptive educational programs should be used until the child has acquired sufficient familiarity with the language and culture of the school for more formal assessment (p. 37).

Even though the problem of disproportionate numbers of minority children in special education programs was being addressed through litigation sparked by the civil rights movement, the concern persisted and continues to plague today's educational system. Artiles & Trent (1994) state that the existence of biased systemic procedure, particularly faulty referral and assessment practices, dramatic changes in the socioeconomic and demographic composition of the United States, and the tendency for uninformed educators to equate disability with cultural differences all contribute to the continuing problem today. What Artiles & Trent (1994) recommend is that we begin to examine the problem of overrepresentation from a multivariate perspective and that the difficulties exhibited by some culturally and linguistically diverse students should be explained beyond the traditional within-child deficit model. Deno (1970) further explained the need to rethink the dependency on the medical model because he stated that the emphasis on 'defect' residing in the child tends to focus attention away from the external variables which educators may be in a position to do something about.

It is imperative that school psychologists examine the possibility that the learning difficulties of minority students might be pedagogically-induced (Cummins, 1986). Specifically, they need to examine for the possible within-system deficits in addition to the within-child deficits that assessment tools have identified. Some of the within-system deficits that could impede an ELL child's academic performance include: 1) type, quality, or lack of a bilingual program (Cziko, 1990; Ramirez, 1992), 2) the teacher's knowledge, skills, and sensitivity about accommodating for cultural and linguistic factors (Garcia & Ortiz, 1988), and 3) whether the teacher implements a reciprocal interaction or transmission oriented approach to teaching (Cummins, 1984).

Best practices in the assessment of ELL and/or bilingual students would include that school psychologists assess a child's proficiency in his or her native language as

well as in English in order to determine what language or languages testing should proceed in (Figueroa, 1990). In addition, to using data obtained from standardized testing, it also stresses the need for school psychologists to conduct observations at different settings (school and home) as well as to obtain additional information from parents to better understand how the ELL child's two worlds differ. Chamberlain & Medeiros-Landurand (1991) also suggest that assessors need to assess ELL students in their first language and in English so that a true disability may better be distinguished from limited English proficiency. Finally, it is recommended that school psychologists: 1) utilize more than one instrument to assess intelligence, 2) include nonverbal measures as part of their battery to assess intelligence and 3) should conduct their own language proficiency evaluations rather than relying on existing language proficiency data (Ochoa et al., 1996). The following will include a review of some of the most commonly used nonverbal measures of general intelligence and/or ability.

Review of Nonverbal Intelligence/Ability Tests

Among the most common methods of assessing ELL students is through the use of nonverbal measures (Ochoa et al., 1996), of which there are two basic types. The comprehensive nonverbal test assesses multiple facets of one's intelligence (e.g., memory, reasoning, attention), are often used to make important educational placement decisions, and as a result are often referred to as "high stakes" tests. The two nonverbal multidimensional tests currently available are the Leiter Performance Scale-Revised (Leiter-R; Roid & Miller, 1997) and the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998).

Leiter International Performance Scale. The Leiter International Performance Scale (LIPS; Leiter, 1979) is completely nonverbal in test instruction and response format and is another nonverbal intelligence test that was designed to assess the intellectual functioning of individuals who were non-English speaking, individuals who had hearing impairments, and individuals with communication disorders. In its most recent revision, the Leiter Performance Scale-Revised (Leiter-R, Roid & Miller, 1997) is based on the hierarchical *g* model of intellectual functioning proposed by Gustafson

(1984) and Carroll (1993), which emphasizes visualization (*Gv*) and fluid reasoning (*Gf*).

Universal Nonverbal Intelligence Test. The Universal Nonverbal Intelligence Test (UNIT; Bracken & McCallum, 1998) is one of only two current intelligence tests that is completely nonverbal in instruction and response format. Since language is not a factor in one's performance on this test, the UNIT is not only adequate for assessing ELL and culturally diverse individuals, but is also ideal for students with severe language and speech related disabilities. The authors also state that the UNIT assesses multiple facets of intelligence as well as other higher order cognitive processes. Specifically, it contains six subtests designed to assess functioning according to a two-tier model of intelligence (memory and reasoning), which utilizes two organizational strategies (symbolic and nonsymbolic organization). Theoretically, the UNIT is intended to provide a measure of *g* with two factor-based scales (Bracken & McCallum, 1998).

In contrast, a unidimensional test has a narrower breadth but is often more suitable for screening applications and group administration, hence allowing for more efficient scoring and administration. Among the most widely used unidimensional nonverbal tests for children are the Raven's Progressive Matrices (RPM; Raven, Court, & Raven, 1986), the Test of Nonverbal Intelligence-Third Edition (TONI-3; Brown, Sherbenou, & Johnson, 1997), and the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997a). McCallum, Bracken, & Wasserman (2001) state that each of these measures may be used in varying degrees to assess general cognitive and intellectual ability, to screen students potentially eligible for special services, and to more fairly assess students with limited English proficiency or with diverse cultural and educational backgrounds. Additionally, they can serve to screen students who would be disadvantaged by traditional language-loaded assessments, to assess what students *can* do despite whatever language, motor, or color vision limitations they may have, and to help identify intellectually gifted students.

Raven's Progressive Matrices. Raven's Progressive Matrices (RPM; Raven, Court & Raven, 1986) was developed at the beginning of World War III and is a nonverbal measure of Spearman's *g* factor which is comprised of abstract/figural problems. Including the original Standard Progressive Matrices (Raven, 1938), the Coloured Progressive Matrices (Raven, 1947), and the revised Progressive Matrices and Vocabulary Scales (Raven, Court, & Raven, 1986), the Raven's Matrices later served as the model for subtests using matrix analogies for future intelligence tests (i.e., Kaufman Adolescent and Adult Intelligence Test [KAIT; Kaufman & Kaufman, 1993], Wechsler Adult Intelligence Scale-Third Edition [WAIS-3; Wechsler, 1997]).

Test of Nonverbal Intelligence. The TONI (TONI; Brown et al., 1982) is yet another individually or group administered test designed to be a language-free measure of cognitive ability that can be used with linguistically handicapped or deprived individuals. In its most recent revision, the TONI-3 (TONI-3; Brown, Sherbenou, & Johnsen, 1997) uses performance measures instead of paper and pencil tasks, it requires problem solving instead of the recall of specific factual information, uses novel problems to decrease the impact of prior exposure, and utilizes pantomimed instructions to the examinees. However, because only abstract reasoning and problem solving are measured, its measured cognitive abilities are considered too narrow by most critics.

Naglieri Nonverbal Ability Test. The NNAT (NNAT; Naglieri, 1997a) is the latest expansion and revision of the Matrix Analogies Test (MAT) and is based upon the MAT-Expanded Form (MAT-EF; Naglieri, 1985a) and the MAT-Short Form (MAT-SF; Naglieri, 1985b) and is considered to be appropriate for use with children of culturally, linguistically, and socio-economically diverse background. Normed on approximately 89,000 children, the NNAT is designed to provide an assessment of school ability for students in grades K-12 and to predict future school achievement by using items that contain only shapes and designs and requires students to rely on reasoning skills rather than verbal ability to answer the items. Administered in either an individual or group format, the 38 items on the NNAT are grouped into four clusters: Pattern Completion (PC), Reasoning by Analogy (RA), Serial Reasoning (SR), and Spatial Visualization

(SV). Roughly based on Spearman's hierarchical *g* model of intellectual functioning, the NNAT yields a general ability score (Nonverbal Ability Index) which has a mean of 100 and a standard deviation of 15 (NNAT; Naglieri, 1997c).

Nonverbal assessment has been defined in numerous ways but typically refers to instruments with reduced emphasis on language on the part of the examiner and examinee (Anastasi, 1988). Harris, Reynolds, & Koegel (1996) state that because of this decreased emphasis on language, the impact of culturally based linguistic differences can be reduced. However, simply because the test is nonverbal does not guarantee that it is fair. Reynolds & Kaiser (1990) recommend that statistical analyses that systematically examine the performance of the tests when used with non-majority groups are necessary. Specifically, instruments need to be evaluated in terms of typical psychometric criteria, as well as on the basis of their usefulness with non-majority groups (Athanasiou, 2000). Although some preliminary studies conducted with the NNAT show adequate reliability and validity indices with different populations, additional research is needed.

To date, there are few studies examining the reliability and validity properties of the NNAT with culturally and linguistically diverse populations. One such study of the NNAT, which based its internal consistency reliability coefficients on the Kuder-Richardson Formula #20, showed evidence of high total test internal score consistency, with reliability coefficients ranging from 0.83 to 0.93 by grade (median across grades was 0.87) and 0.81 to 0.88 by age across the seven levels (McCallum et al, 2001). Additional investigations of test fairness with the NNAT examined differences between three demographically matched samples of white and African American students, white and Hispanic students, and white and Asian students. Selected from the larger standardization sample and matched according to geographic region, socioeconomic status, ethnicity, and type of school setting (public or private), effect size differences were reported to be small between groups, according to Cohen's (1988) suggestion to interpret *d* ratios less than 0.5 as small. The white and African American samples' effect size differences were small (*d*-ratio=0.25), the white and Hispanic samples' effect size

differences were also small (*d*-ratio=0.17), and the white and Asian samples' effect size differences were negligible (*d*-ratio=0.02). These results show that the NNAT shows small group mean score differences when selected racial and ethnic groups are compared (Naglieri &Ronning, 2000).

Using the same standardization sample, Naglieri & Ronning (2000) also reported correlations for the same racial and ethnic groups between NNAT performance and performance on the reading and math subtests of the Stanford Achievement Test. The minority groups typically had NNAT reading and math correlations that were at least as high as those found for the white samples. The differences between the white and minority group correlations were nonsignificant for the three samples, with the only exception of the white and African American comparisons in reading (0.48 vs. 0.62). Naglieri & Ronning (2000) indicate that these two correlations were significantly different ($z=6.8$, $p<.01$), suggesting that the NNAT was a better predictor of reading for the African American than the white children. Overall, however, the correlations for kindergarten-12th grade white vs. African American, white vs. Hispanic, and white vs. Asian samples varied from a low of 0.46 to a high of 0.68 (Mdn=0.68*)*, indicating a moderate relationship between the NNAT and achievement in reading and math for the various groups. Specifically, the median correlations between the NNAT and reading and math were 0.52 and 0.63 across the three matched samples respectively.

Summary

There is a great need for psychometrically sound and appropriate instruments to assess the intelligence of school-aged children within the United States. Since Mexican American children comprise the largest minority group within the United States and present with varying levels and/or degrees of cultural factors that can influence assessment practices with this population (i.e., English language proficiency levels, acculturation levels, cognitive abilities, etc…), there is also a salient need to utilize culturally fair psychometric instruments and practices for this population. Legal cases and federal legislation have mandated that assessment professionals address the needs of America's culturally and linguistically diverse school-aged population. Among the most

salient concerns providing a major impetus for litigation regarding the assessment practices of minority populations has been the disproportionate number of minority children enrolled in special education classes. Frequently, linguistically diverse children have been identified as having a learning disability or as having a speech and language impairment due to their limited English proficiency (Ortiz & Polyzoi, 1986). Consequently, erroneous referrals and assessment practices have unfortunately led to the misdiagnosis of many of our culturally and linguistically diverse student population.

Researchers in the area of bilingual and multicultural assessment have delineated guidelines regarding best practices in the assessment of culturally and linguistically diverse populations to help attenuate faulty assessment practices and erroneous referrals. Among them include assessing in both child's native and English language, formally determining the child's level of language proficiency before proceeding with subsequent testing, and using both informal and formal language measures (Figueroa, 1990; Ochoa et al., 1996). Ochoa et al. (1996) also provided several reasons why language proficiency assessments need to be conducted with ELL students. First, federal guidelines mandate such action be taken within America's school system. Second, language proficiency assessment can help determine whether a child is in the appropriate educational environment (i.e., bilingual vs. English-only setting). Third, language proficiency data can help the assessor determine which language testing should proceed in (Figueroa, 1990). Finally, language proficiency assessment data can shed light on whether an ELL child's academic problems stem from the second-language-learning process or whether such difficulties are reflective of a genuine disability (Chamberlain & Medeiros-Landurand, 1991). In addition, many researchers also recommend that nonverbal IQ scores be included in the assessment battery when testing ELL students (Holtzman & Wilkinson, 1991; Ochoa et al., 1996).

The aforementioned demographic statistics and research data collectively demonstrate that a great need exists for psychometrically sound and appropriate instruments to assess the intelligence of culturally and linguistically diverse children within the United States. Studies are specifically needed to examine the psychometric

usefulness of existing intelligence and/or ability tests for use with America's diverse populations. Given the results of previous research conducted on the NNAT with minority populations, further research with linguistically diverse populations is warranted.

Therefore, the purpose of this study was to examine the internal consistency, reliability, and validity of the NNAT for a group of ELL Mexican American children. The methodology to meet these goals will be described in Chapter III. The results will be presented in Chapter IV, with implications for research and practice discussed in Chapter V.

CHAPTER III

METHODS

This chapter will discuss the methods used in conducting the reliability and validity study of the NNAT with ELL Mexican American children. The first section includes a description of the participants and the criteria used to select them. The second section explains the research procedures used in the study. The third section describes the NNAT and provides a summary of current studies examining its reliability and validity properties. Finally, the fourth section discusses the design used in the investigation. It describes the qualitative and quantitative methods used in gathering and analyzing the data as well as a description of the instrument used.

Participants

The participants of this study were 122 Mexican American children from grades 3 (n=91) and 4 (n=31), ages 8 to 10 with a mean age of 9 years 7 months and a standard deviation of 7 months. With respect to gender, 62 participants were male and 60 were female. This group was selected because these are the grade levels in which more referrals are made for special education placement. The participants in this study were children who had not been referred for special education evaluation, had not been retained, and who were not receiving special education services.

This was a sample of convenience selected from four public elementary schools within a rural school district located in the southwest region of the United States. During the 2003-2004 academic school year, the school district was comprised of approximately 7,500 students, 93% of which were of Mexican American descent, 85% that were identified as economically disadvantaged, and 28% that were identified as ELL.

As ELL's, each student had been previously identified as needing bilingual services, thus meeting the state criteria to be enrolled in the school district's bilingual education program. Each participant in this study was currently enrolled in a bilingual

program. Students were taken from a total of 16 different bilingual education classrooms, with the number of subjects ranging from 2-15 ELL students in each classroom who had returned signed parental consent forms.

Procedures

This dissertation was approved by the Texas A&M University Institutional Review Board (IRB). Prior to embarking on the data collection phase of this study, an approval letter was received from the school district's superintendent allowing for the principal investigator to proceed with testing. To obtain the sample for the study, the principal investigator contacted the school district's executive director of elementary education to explain the details of the study and discuss the logistics of successfully executing the research study. Once the four elementary schools with the largest ELL student population had been identified, a conjoint meeting with the four respective principals of each school was held.

During this meeting, each principal identified bilingual education teachers of non-special education ELL students and provided their pertinent contact information. Once teachers were identified, meetings were scheduled during their conference periods to discuss the procedures and details of the study, review both English and Spanish versions of consent forms, assent forms, and information letters for each child's parent(s), and a comprehensive list of students was obtained that met the criteria discussed in the aforementioned participant section of this chapter. After the teachers selected the students and identified each parents' language preference, a letter in Spanish or English was sent home to the parents explaining the research study, along with an informed consent form for them to sign and return to the school. These consent forms were sent to the home of these students the week testing was scheduled to begin. One hundred and thirty letters were returned and one hundred twenty eight parents gave consent for their child's participation in the study. However, two of those students were never tested due to their absence from school on the days of testing, four students were excluded from the final analysis due to exceeding the age limit, leaving a total of one hundred twenty two students tested for this dissertation study.

After obtaining parental permission, students were scheduled for testing according to times recommended by their respective teachers. Once dates and times from all teachers were obtained, a schedule for testing was developed and provided to each principal and teacher. All participants in the study signed an assent form prior to being tested; 15 minutes were allowed for testing instructions and a time limit of 30 minutes was imposed for testing according to the instructions included in the NNAT's technical manual (Naglieri, 1997b). A group-administered format was followed according to standardized assessment procedures. Level D of the NNAT, which includes identical test content for both 3rd and 4th grade students, was administered to all participants. All classrooms with less than 8 students were combined with larger classrooms and tested in their respective classrooms and/or in an adjacent room to the school library. The testing environment was quiet and free from distractions. After all testing was completed, a thank you letter was mailed to the principals of each school as well as to the school district's executive director of elementary education.

Instrument

Naglieri Nonverbal Ability Test. The Naglieri Nonverbal Ability Test (NNAT: Naglieri,1997a) was used in this study and is designed to be a brief, culture-fair, group or individually administered nonverbal measure of ability that does not require the child to read, write, or speak (Naglieri, 1997c). The test takes approximately 45 minutes to administer (15 minutes for verbal instruction, 30 minutes for test administration) and is designed as a nonverbal measure of general ability, or "g", comprising of progressive matrix items that use shapes and geometric designs interrelated through spatial logic or logical organization. All of the NNAT items have the same basic requirement: the child must examine the relationships among the parts of the matrix and determine which response is the correct one on the basis of only the information provided in the matrix. The NNAT items are organized into seven levels, each containing 38 dichotomously scored items for students in grades K-12th. Each level contains a carefully selected set of items that are most appropriate for children at the grade or grades for which that level is intended (Naglieri & Ronning, 2000). Specifically, level A is for kindergarten, level B

for 1st grade, level C for 2nd grade, level D for 3rd and 4th graders, level E for 5th and 6th graders, level F for 7th, 8th, and 9th graders, and level G for 10th, 11th, and 12th graders. Level D of the NNAT, which is composed of identical test content for both 3rd and 4th grade, was administered to all participants of this study (Naglieri, 1997c).

The NNAT is a revision and extension of the Matrix Analogies Test-Short Form and Expanded Form (Naglieri, 1985a, 1985b), which uses the same four matrices-based item types (i.e., Pattern Completion, Reasoning by Analogy, Serial Reasoning, and Spatial Reasoning) as well as brief verbal directions. Pattern Completion is comprised of 6 items, Reasoning by Analogy is comprised of 10 items, Serial Reasoning is comprised of 8 items, and Spatial Visualization is comprised of 14 test items respectively. In the technical manual, Naglieri (1997b) contends that each of the four clusters, to a certain extent, reflects general ability in a somewhat unique way. Although the four clusters yield separate raw scores which, when converted to scaled scores together comprise the Nonverbal Ability Index score (NAI), the technical manual emphasizes that the NAI score should be used to interpret general ability, is the most reliable predictor of a student's academic success, and is the best indicator of overall general ability (Naglieri, 1997b). The NNAT presents matrices using yellow, blue, and white figural stimuli to minimize effects of impaired color vision on test performance. The NNAT was also designed to assess performance that is not dependent upon stores of acquired knowledge. Basically, factual knowledge, vocabulary, mathematics, and reading skills are not prerequisites for solving NNAT items (Naglieri, 1997b).

The NNAT standardization sample included approximately 22,600 students in the Fall sample and approximately 67,000 students in the Spring sample. The race/ethnic composition of the Fall standardization sample included 14.1% (n=3,187) that were African American, 10.5% (n=2,373) that were Hispanic, 69.4% (n=15,684) that were white, 2.9% (n=655) that were Asian, and 1.4% (n=316) that were American Indian. In addition, the Fall standardization sample included approximately 1.2% (n=271) who were identified as ELL students. The race/ethnic composition of the Spring standardization sample included 11.8% (n=7,906) that were African American,

11.3% (n=7,571) that were Hispanic, 71% (n=47,570) that were white, 3.1% (n=2,077) that were Asian, and 1.3% (n=871) that were American Indian. Upon closer examination of the Spring standardization sample, approximately 1.7% (n=1,139) were children who were identified as ELL students (Naglieri, 1997c).

With respect to internal consistency reliability, the reliability coefficient for the Naglieri Ability Index (NAI) total score was 0.80 for grade 3 and 0.83 for grade 4 (Naglieri, 1997c). In addition, the internal consistency coefficients for the NAI total score according to different age groups were 0.83 for 8-year-olds, 0.85 for 9-year-olds, and 0.87 for 10-year-olds respectively (Naglieri, 1997c).

The Pattern Completion reliability coefficient for the standardization sample was 0.44 for grade 3 and 0.48 for grade 4. The Reasoning by Analogy reliability coefficient for the standardization sample was 0.30 for grade 3 and 0.41 for grade 4. The Serial Reasoning reliability coefficient for the standardization sample was 0.70 for grade 3 and 0.68 for grade 4. Finally, the Spatial Visualization reliability coefficient for the standardization sample was 0.71 for grade 3 and 0.78 for grade 4 respectively (Naglieri, 1997c).

The standard error of measurement (SEM) for the NAI total score was 2.9 for grade 3 and 2.8 for grade 4. The SEM for Pattern Completion for the standardization sample was 0.90 for grade 3 and 0.80 for grade 4. The SEM for Reasoning by Analogy for the standardization sample was 1.5 for both grades 3 and 4. The SEM for Serial Reasoning was 1.2 for grade 3 and 1.1 for grade 4. Finally, the SEM for Spatial Visualization was 1.7 for both grades 3 and 4 respectively (Naglieri, 1997c).

The NNAT's predictive validity has been demonstrated in various studies reported in the NNAT's technical manual (NNAT, 1997b). For example, correlations between the NNAT and the Stanford Achievement Test-Ninth Edition (SAT-9) were fairly consistent across K-12th grade levels as well as across A-G test levels used. Specifically, the correlations between the NNAT and the SAT-9 for 3rd graders (n=2,357) in Total Reading (r=0.53), Total Mathematics (r=0.61), Language (r=0.58), and Complete Battery (r=0.62) were comparable to those found with the 4th grade

standardization sample (n=2,054) in Total Reading (r=0.59), Total Mathematics (r=0.68), Language (r=0.62), Thinking Skills (r=0.65), and Complete Battery (r=0.66). Overall, Naglieri (1997b) reports that a median correlation of 0.61 between the NNAT and the SAT-9 was found across K-12th grade for 21,476 children included in the study.

In a similar study examining the correlations between the NNAT and the APRENDA2, a Spanish achievement test, the correlations for 3rd graders (n=1,364) in Total Reading (r=0.34), Total Mathematics (r=0.51), Language (r=0.43), and Basic Battery (r=0.49) reflected similarly deflated correlations for 4th graders (n=877) in Total Reading (r=0.32), Total Mathematics (r=0.56), Language (r=0.41), Thinking Skills (r=0.47), and Basic Battery (r=0.48) when compared to the results obtained in the aforementioned study (Naglieri, 1997c).

Finally, a general component of construct validity was examined and reported in the technical manual (NNAT, 1997b). Specifically, the consistency of the NNAT in assessing the same complex of skills across the 7 different levels of the NNAT was examined. The following results were yielded: level A (n=1,202) had a correlation coefficient between adjacent levels of 0.80, level B (n=1,057) had a correlation coefficient between adjacent levels of 0.82, level C (n=1,160) had a correlation coefficient between adjacent levels of 0.81, level D (n=1,021) had a correlation coefficient between adjacent levels of 0.82, level E (n=580) had a correlation coefficient between adjacent levels of 0.81, and level F (n=456) had a correlation coefficient between adjacent levels of 0.81 respectively.

However, it should be noted that although a total of approximately 1,410 children that had been identified as ELL students were included in the standardization sample, no specific data were reported in the technical manual with regards to the reliability and validity of the NNAT with an ELL Mexican American sample.

Design

Reliability. Reliability involves the consistency of scores obtained by the same person when retested with the same instrument on different occasions, with a different set of equivalent items, or under other variable testing conditions (Anastasi, 1988).

Essentially, the reliability of a test is the degree to which the test and its scores reflect true or nonerror variance. The two basic methodological considerations in assessing reliability are time and content. Specifically, the various types of reliability include: test-retest, parallel or alternate forms, scorer or rater reliability, split-half or internal consistency, and Kuder-Richardson 20 or internal consistency (Sattler, 2001).

In an effort to determine how well the NNAT measures a single unidimensional latent construct**,** internal consistency was examined from a single administration of the test by computing Cronbach's alpha, an extension of both the split-half estimates and the Kuder Richardson formula, on the 38 total items as well as on each of the NNAT's four respective clusters. The Kuder Richardson formula is equal to Cronbach's alpha when items are dichotomously scored. Cronbach's alpha is a function of the number of test items *and* the average inter-correlation among the items. Intuitively speaking, if the inter-item correlations are high, then there is evidence that the items are measuring the same underlying construct. However, if the reliability coefficients do not yield technically adequate values, the data might reflect that the NNAT might not purely be a unidimensional test and require a more extensive factor analysis to determine which items load highest on which dimensions and how many distinct factors can statistically be extracted to comprise different dimensions (Smith, 1998).

In addition to using Cronbach's alpha to obtain reliability coefficients, a procedure developed by Feldt (1980) to test for reliabilities between two independent groups was used. V. Willson (personal communication, September, 21, 2004) also provided a formula (appendix A) to obtain pooled reliability estimates from the 3rd and 4th grade students on the NNAT NAI total score as well as the four individual clusters for both the ELL sample and the standardization sample since only separate reliability coefficients for the 3rd and 4th grade student standardization sample were reported in the technical manual. *F*-statistics were subsequently computed to determine whether there were statistically significant differences in the pooled reliability coefficients between the ELL sample and the standardization sample.

Because only two groups were being compared in this study, *t*-tests for independent groups were conducted to calculate the differences in means between the standardization sample and the sample included in this research study. By utilizing the mean, standard deviation, and standard error of measurement to compute the confidence intervals, the derived *t* statistic determines whether there are statistically significant differences between the two independent groups. Specifically, *t* tests were computed to compare the performance of a sample of ELL Mexican American children on the NAI total raw score as well as the cluster scores of the NNAT with the performance of those included in the standardization sample.

Validity. The *Standards for Educational and Psychological Testing* (AERA, 1999) define validity as the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests. Therefore, the process of validation is not to generally determine whether a test is valid, but rather whether it is valid for a particular interpretation in a specific application. There are three types of validity, one of which comprises two subtypes: predictive, concurrent, content, criterion-related, and construct validity. This study sought to examine the construct validity of the NNAT with an ELL Mexican American sample. The American Educational Research Association (1966) stated that construct validity is evaluated by investigating what qualities a test measures, that is, by determining the degree to which explanatory concepts or constructs account for performance on a test.

In this study, we investigated the extent to which the NNAT's proposed factor structure fit the factor structure yielded from the data with the ELL Mexican American sample. Because the NNAT is roughly based on Spearman's hierarchical *g* model of intellectual functioning and, as such, is purported to measure general ability, the NNAT is considered to be a unidimensional test. However, although the technical manual does not specify how, it does state that the NNAT also yields separate raw scores for the four clusters that individually measure general ability in slightly different ways. Therefore, a confirmatory factor analysis was conducted to determine how well the items on each of the four different clusters on the NNAT (i.e., Pattern Completion, Reasoning by

Analogy, Serial Reasoning, and Spatial Visualization) function together as separate constructs. In addition, because there is currently no research examining the factor structure of the NNAT with an ELL Mexican American sample, an exploratory factor analysis using Pearson Product Moment correlations was conducted to determine the construct validity of the NNAT. The Pearson Product Moment correlation coefficient can be used to predict and/or estimate the degree of the relationship between two variables and is considered to be the best estimate when the relationship is linear (Anastasi & Urbina, 1997).

Summary

This research study investigated the internal consistency reliability of the four clusters as well as the total Nonverbal Ability Index (NAI) score of the NNAT with ELL Mexican American children. In addition, using *t* tests for independent groups, it explored whether there were statistically significant differences between the standardization sample and the ELL Mexican American sample with regards to their performance on the total NAI score as well as on each of the individual clusters. Finally, it investigated how the proposed four-factor structure of the NNAT compared with that yielded by the ELL Mexican American sample.

Research Questions

1. What is the internal consistency of the four clusters and the Naglieri Ability Index (NAI) total score of the NNAT with ELL Mexican American children? Are the obtained reliability coefficients technically adequate? Are they comparable to those obtained with the standardization sample? This question will be addressed by computing Cronbach's alpha and using Feldt's (1980) formula for computing reliability coefficients for two independent groups. In addition, V. Willson (personal communication, September 21, 2004) provided a formula so that the pooled reliability estimates for 3rd and 4th grade students in both the ELL sample and the standardization sample could be computed and compared to determine whether statistically significant differences existed.

2. How does the performance of the sample of ELL Mexican American children differ from the standardization sample with respect to the four cluster scores and the overall Nonverbal Ability Index (NAI) raw score of the NNAT? Are the differences between the means of the two groups statistically significant? This question will be addressed by computing *t*-tests for independent groups.
3. How does Naglieri's proposed factor structure for the NNAT compare with the factor structure found with ELL Mexican American children? This question will be addressed by computing a confirmatory factor analysis and exploratory factor analysis of Pearson Product Moment correlations.

CHAPTER IV

DATA ANALYSIS AND RESULTS

This chapter includes the results of the data analysis and is organized according to the three research questions. Chapter V will provide a discussion about the results.

1. **What is the internal consistency of the four clusters and the Naglieri Ability Index (NAI) total score of the NNAT with ELL Mexican American children? Are the obtained reliability coefficients technically adequate? Are they comparable to those obtained with the standardization sample?**

Table 1 presents the reliability coefficients for the NNAT's NAI total score and for the four cluster scores for the ELL Mexican American sample and the standardization sample. Because the Kuder Richardson formula is equal to Cronbach's alpha when items are dichotomously scored and NNAT items are all scored dichotomously, Cronbach's alpha was used to calculate a coefficient of internal consistency. The reliability criterion deemed as acceptable for tests of cognitive ability is 0.80 (Sattler, 1988). The results of the reliability analyses determined that the NAI total score for both 3rd and 4th grade students exceeded the acceptable level with reliability coefficients of 0.83 each. However, only the Spatial Visualization cluster scores for both 3rd and 4th grade students approached the acceptable level with reliability coefficients of 0.74 and 0.78 respectively. The reliability coefficients of the three other clusters (Pattern Completion 0.50, Reasoning by Analogy 0.44, and Serial Reasoning 0.69) for the 3rd grade students were all below the cited acceptable standard. Similarly, the reliability coefficients of the same three clusters (Pattern Completion 0.52, Reasoning by Analogy 0.53, and Serial Reasoning 0.48) for the 4th grade students were all below the cited acceptable standard.

Table 1 also presents the pooled reliability estimates of the NNAT's NAI total score and the four cluster scores for both the 3rd and 4th grade ELL Mexican American

Table 1

Internal Reliability Coefficients, Pooled Reliability Estimates, and F-values for the Performance on the NNAT's NAI Total Score and Cluster Scores: ELL Mexican American Sample vs. Standardization Sample

					Pooled Reliability Estimates		F-statistic
	3rd Grade ELL Sample (n=91)	3rd Grade Standardization Sample (n=5428)	4th Grade ELL Sample (n=31)	4th Grade Standardization Sample (n=5180)	ELL Sample (n=122) 3rd and 4th	Standardization Sample (n=10,608) 3rd and 4th	df=121, 10,607
NAI Total Score (38 items)	0.83	0.80	0.83	0.83	0.828221	0.812488	1.09150
Pattern Completion (6 items)	0.50	0.44	0.52	0.48	0.505949	0.458871	1.09529
Reasoning by Analogy (10 items)	0.44	0.30	0.53	0.41	0.467812	0.365283	1.19265
Serial Reasoning (8 items)	0.69	0.70	0.48	0.68	0.671295	0.689726	0.94393
Spatial Visualization (14 items)	0.74	0.71	0.78	0.78	0.738161	0.749770	0.95566

sample and the standardization sample. When F-statistics were computed, no statistically significant differences in pooled reliability estimates were found.

2. **How does the performance of the sample of ELL Mexican American children differ from the standardization sample with respect to the four cluster scores and the overall Nonverbal Ability Index (NAI) raw score of the NNAT? Are the differences between the means of the two groups statistically significant?**

Table 2 presents the means, standard deviations, and standard errors of measurement for the 3rd grade ELL Mexican American sample and 3rd grade standardization sample. The significance level for the *t*-tests was set at .05. Since the analyses for both 3rd and 4th grade students were not testing for directionality, a two-tailed t-test was conducted.

Table 3 presents the mean differences between the two groups, the *t*-statistics, and *p*-values for the 3rd grade ELL Mexican American sample and 3rd grade standardization sample. The obtained *p*-values indicate that there were no statistically significant differences between the ELL Mexican American 3rd grade sample and the standardization sample on the NAI total raw score as well as on the 4 cluster scores. The *p*-values for the NAI total score (0.4878) as well as the 4 cluster scores (Pattern Completion 0.7038, Reasoning by Analogy 0.7023, Serial Reasoning 0.5796, and Spatial Visualization 0.0814) were all above the significance level.

Table 4 presents the means, standard deviations, and standard errors of measurement while Table 5 presents the mean differences between the two groups, the t-statistics, and p-values for the 4th grade ELL Mexican American sample and 4th grade standardization sample. The obtained p-values indicate that there were no statistically significant differences between the ELL Mexican American 4th grade sample and the 4th grade standardization sample on the NAI total raw score as well as on the 4 cluster scores. The p-values for the NAI total score (0.7932) as well as the 4 cluster scores (Pattern Completion 0.0525, Reasoning by Analogy 0.7324, Serial Reasoning 0.3163, and Spatial Visualization 0.2846) were all above the significance level, indicating no

Table 2

Means, Standard Deviations, and Standard Errors of Measurement: ELL Mexican American Sample vs. Standardization Sample- 3rd Grade

	3rd Grade ELL Sample n= 91			3rd Grade Standardization Sample n=5428		
	Mean	SD	SEM	Mean	SD	SEM
NAI Total Raw Score	20.34	6.23	0.65	20.80	6.50	2.90
Pattern Completion	4.85	1.24	0.13	4.90	1.20	0.90
Reasoning by Analogy	4.77	1.72	0.18	4.70	1.90	1.50
Serial Reasoning	5.62	2.04	0.21	5.50	2.10	1.20
Spatial Visualization	5.12	3.13	0.33	5.70	3.20	1.70

Table 3

Mean Differences, *t*-statistics, and *p*-values: ELL Mexican American Sample vs. Standardization Sample- 3rd Grade

	Mean Difference	*t*-statistic	*p*-value
NAI Total Raw Score	0.46	0.46	0.4878
Pattern Completion	0.05	0.38	0.7038
Reasoning by Analogy	-0.07	0.38	0.7023
Serial Reasoning	-0.12	0.55	0.5795
Spatial Visualization	0.58	1.74	0.0814

Table 4

Means, Standard Deviations, and Standard Errors of Measurement: ELL Mexican American Sample vs. Standardization Sample- 4th Grade

	4th Grade ELL Sample n=31			4th Grade Standardization Sample n=5180		
	Mean	SD	SEM	Mean	SD	SEM
NAI Total Raw Score	23.03	5.98	1.07	23.30	6.80	2.80
Pattern Completion	5.42	0.96	0.17	5.10	1.10	0.80
Reasoning by Analogy	5.19	1.87	0.34	5.30	2.00	1.50
Serial Reasoning	6.26	1.51	0.27	6.00	1.90	1.10
Spatial Visualization	6.16	3.48	0.62	6.80	3.60	1.70

Table 5

Mean Differences, *t*-statistics, and *p*-values: ELL Mexican American Sample vs. Standardization Sample- 4th Grade

	Mean Difference	*t*-statistic	*p*-value
NAI Total Raw Score	0.27	0.26	0.7932
Pattern Completion	-0.32	1.94	0.0525
Reasoning by Analogy	0.11	0.34	0.7324
Serial Reasoning	-0.26	1.00	0.3163
Spatial Visualization	0.64	1.07	0.2850

statistically significant differences were found between the two independent groups.

3. How does Naglieri's proposed factor structure for the NNAT compare with the factor structure found with ELL Mexican American children?

Prior to conducting the confirmatory factor analysis, communality estimates were calculated to determine the proportion of common variance on each test item that could be explained by the proposed factor structure. Based on the preliminary analysis of the communality values of each test item, items 1, 2, and 33 did not appear to contribute to the test variance that could be explained by the four-factor structure. Therefore, items 1, 2 and 33 were eliminated from both confirmatory and exploratory factor analyses. Figure 1 displays the hypothesized 35-item four factor measurement model for the NNAT, including pattern coefficients for each item-cluster pairing as well as error variances for each item.

Table 6 displays the goodness-of-fit statistics from the confirmatory factor analysis that were yielded from the default model with the ELL Mexican American sample, the saturated model, which is one in which the number of estimated parameters equals the number of data points, and the independence model, which is one of complete independence of all variables in the model (i.e., all correlations among variables are zero) and is the most restricted (Byrne, 2000). Specifically, the first set of fit statistics shows 111 parameters, a chi-square value of 688.568, 554 degrees of freedom, and a probability value of less than .0001 ($p<.0001$). What these data suggest is that the fit of the data obtained from the ELL Mexican American sample to the model that the NNAT follows a four-factor structure model is not adequate.

Table 7 depicts incremental or comparative indexes of fit which are based on a comparison of the hypothesized model against some standard, typically the independence or null model noted earlier (Hu & Bentler, 1995). According to Hu & Bentler (1999), indices of fit range from zero to 1.00, with values approaching 0.95 showing evidence of good fit. The normed fit index (NFI) reflects a value of 0.450 while the relative fit index (RFI; Bollen, 1986) reflects a value of 0.409. Bollen (1989) also developed the incremental fit index (IFI) to take into account degrees of freedom.

Figure 1

Hypothesized 35-Item Four Factor Measurement Model for the NNAT

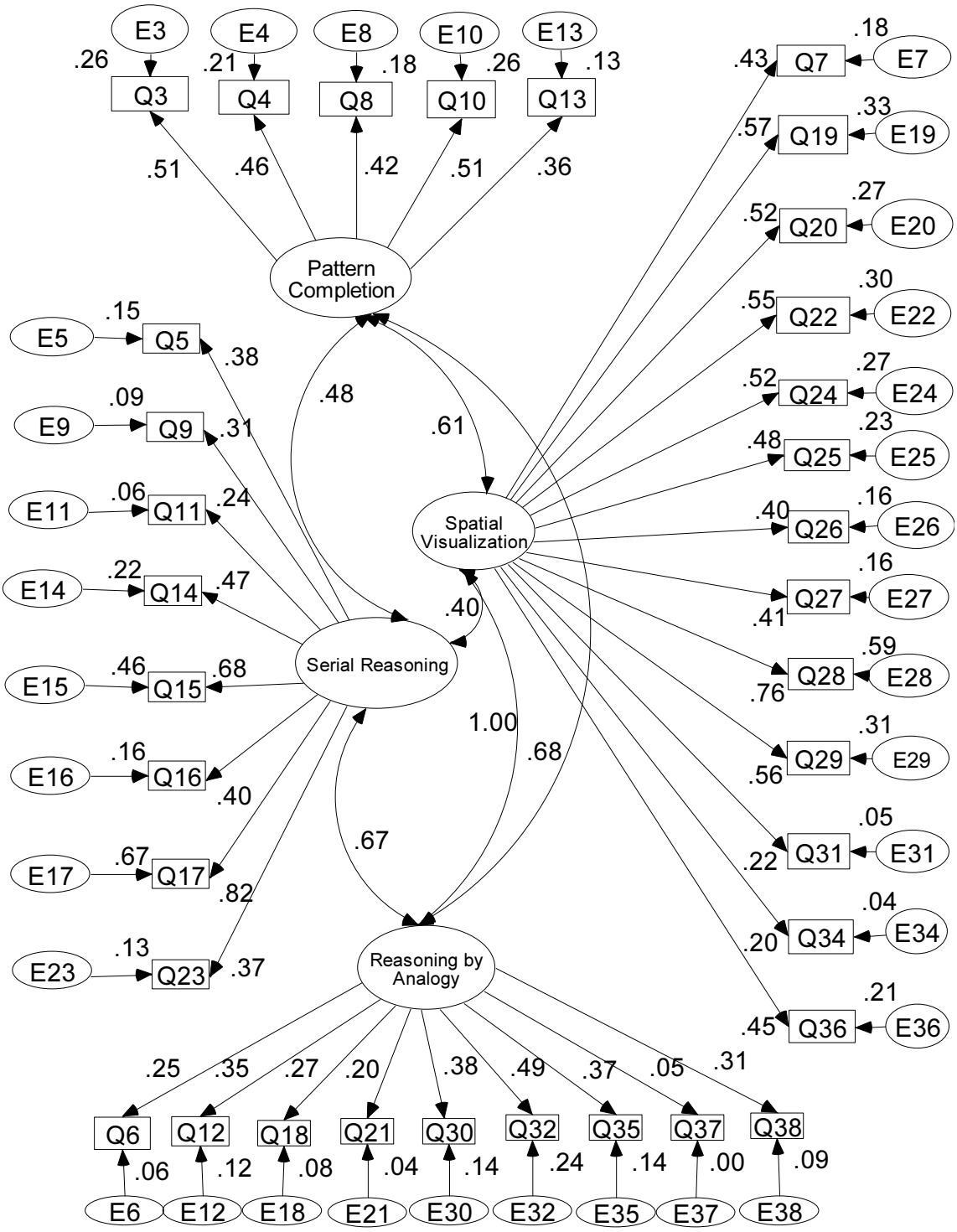

Table 6

Number of Parameters, Chi-Square Statistics, Degrees of Freedom, and Probability Values

Model	Number of Parameters	Chi-Square	Degrees of Freedom	P-value	Chi-Square/ Degrees of Freedom
Default model	111	688.568	554	.000	1.243
Saturated model	665	.000	0	***	****
Independence model	70	1251.234	595	.000	2.013

Table 7

Normed Fit Index, Relative Fit Index, Incremental Fit Index, Tucker-Lewis Index, and Comparative Fit Index Values

Model	Normed Fit Index	Relative Fit Index	Incremental Fit Index	Tucker-Lewis Index	Comparative Fit Index
Default model	0.450	0.409	0.807	0.780	0.795
Saturated model	1.000	***	1.000	***	1.000
Independence model	.000	.000	.000	.000	.000

Table 7 shows an IFI value of 0.807. The Tucker-Lewis index (TLI; Tucker & Lewis, 1973) yielded a value of 0.780, while the comparative fit index (CFI; Bentler, 1990), which takes into account sample size, yielded a value of 0.795. In summary, Table 7 offers additional data that indicates a lack of fit between the model found with the data from the ELL Mexican American sample and the NNAT's proposed four-factor structure.

Table 8 displays the root mean square error of approximation (RMSEA) which takes into account the error of approximation in the population and asks the question, "How well would the model, with unknown but optimally chosen parameter values, fit

Table 8

Root Mean Square Error of Approximation, Lower and Higher Confidence Intervals, and *p*-values for the Closeness of Fit

Model	Root Mean Square Error of Approximation	Lower Confidence Interval	Higher Confidence Interval	P-value for Test of Close Fit
Default model	0.045	0.033	0.055	0.781
Independence model	0.095	0.088	0.103	0.000

the population covariance matrix if it were available" (Browne & Cudeck, 1993, pp. 137-138). This discrepancy, as measured by RMSEA, is expressed in degrees of freedom, thus making the index sensitive to the number of estimated parameters. Table 8 shows a RMSEA value for the hypothesized model of 0.045, with the 90% confidence interval ranging from 0.033 to 0.055, and the *p*-value for the test of closeness of fit equal to 0.781. Hu & Bentler (1999) indicate that a cutoff RMSEA value close to 0.06 indicates good fit and is needed before one can conclude that there is a relatively good fit between the hypothesized model and the observed data. However, although Browne & Cudeck (1993) state that a probability value associated with test of close fit greater than 0.50 indicates that the hypothesized model fits the observed data well, sample size significantly influences size of confidence interval, thereby affecting the precision with which one can determine accurately the degree of fit in the population. Furthermore, given the complexity of this model, as evidenced by the large number of estimated parameters (n=111), and the small sample size (n=122), a very large sample size would be required in order to obtain a reasonably narrow confidence interval (Browne & Cudeck, 1993).

Table 9 displays the parameter estimates, standard errors of measurement, critical ratio values, and *p*-values for the default model on each of the 35 items of the NNAT included in the analyses while Table 10 displays correlation estimates between each of the four clusters on the NNAT. The correlation estimates ranged from 0.480 to 0.680,

Table 9

Parameter Estimates, Standard Errors of Measurement, Critical Ratio Values, and *p*-values for the Default Model with ELL Mexican American Sample

Variable	Cluster	Parameter Estimates	Standard Error	Critical Ratio	p-value	Label	Variable	Cluster	Parameter Estimates	Standard Error	Critical Ratio	p-value	Label
Q3	Pattern Completion	1.000					Q18	Reasoning by Analogy	1.292	0.568	2.276	.023	Par_19
Q13	Pattern Completion	0.814	0.296	2.746	.006	par_1	Q12	Reasoning by Analogy	1.687	0.652	2.588	.010	Par_20
Q7	Spatial Visualization	1.000					Q23	Serial Reasoning	1.000				
Q19	Spatial Visualization	1.403	0.355	3.956	***	Par_2	Q17	Serial Reasoning	2.061	0.569	3.620	***	Par_21
Q20	Spatial Visualization	1.363	0.362	3.765	***	Par_3	Q5	Serial Reasoning	.631	0.226	2.793	.005	Par_22
Q22	Spatial Visualization	1.342	0.346	3.875	***	Par_4	Q14	Serial Reasoning	1.239	0.402	3.080	.002	Par_23
Q24	Spatial Visualization	1.352	0.357	3.781	***	Par_5	Q21	Reasoning by Analogy	0.937	0.502	1.868	.062	Par_24
Q25	Spatial Visualization	1.221	0.337	3.625	***	Par_6	Q6	Reasoning by Analogy	0.859	0.401	2.144	.032	Par_25
Q26	Spatial Visualization	1.050	0.322	3.265	.001	Par_7	Q35	Reasoning by Analogy	1.442	0.542	2.659	.008	Par_26
Q27	Spatial Visualization	1.068	0.326	3.279	.001	Par_8	Q37	Reasoning by Analogy	0.245	0.426	0.574	.566	Par_27
Q28	Spatial Visualization	1.948	0.442	4.412	***	Par_9	Q15	Serial Reasoning	1.713	0.487	3.515	***	Par_30
Q29	Spatial Visualization	1.369	0.350	3.912	***	Par_10	Q16	Serial Reasoning	1.010	0.353	2.861	.004	Par_31
Q31	Spatial Visualization	0.563	0.276	2.039	.041	Par_11	Q30	Reasoning by Analogy	1.716	0.637	2.694	.007	Par_32
Q36	Spatial Visualization	0.936	0.267	3.514	***	Par_12	Q32	Reasoning by Analogy	2.148	0.725	2.962	.003	Par_33
Q8	Pattern Completion	0.940	0.304	3.087	.002	Par_13	Q11	Serial Reasoning	0.582	0.284	2.051	.040	Par_35
Q10	Pattern Completion	1.174	0.342	3.436	***	Par_14	Q9	Serial Reasoning	0.699	0.285	2.454	.014	Par_36
Q38	Reasoning by Analogy	1.000					Q4	Pattern Completion	0.818	0.252	3.248	.001	Par_37
Q34	Spatial Visualization	0.436	0.230	1.899	.058	Par_15							

Table 10

Correlation Estimates Between Each of the Four Clusters of the NNAT for ELL Mexican American Sample

			Correlation Estimates
Pattern Completion	<--->	Spatial Visualization	0.614
Pattern Completion	<--->	Reasoning by Analogy	0.680
Serial Reasoning	<--->	Reasoning by Analogy	0.670
Spatial Visualization	<--->	Reasoning by Analogy	1.002
Serial Reasoning	<--->	Spatial Visualization	0.404
Serial Reasoning	<--->	Pattern Completion	0.480

with the correlation estimate between the Spatial Visualization and Reasoning by Analogy clusters (1.002) showing the presence of a perfect reliability of items, offering additional data to support a lack of fit between the hypothesized model and observed data.

Given the poor reported goodness-of-fit statistics, there is evidence that the hypothesized four-factor structure model of the NNAT does not fit well with this sample of ELL Mexican American children. Therefore, an exploratory factor analysis (EFA) was conducted to investigate possible factor structure.

Upon examining the eigenvalues of the 35 test items displayed in Table 11, it was determined that due to drops between components six and seven as well as nine and ten, two separate EFA's were conducted to determine which model was most representative of this sample of ELL Mexican American children. Utilizing the Principal Axis method, six factors were initially extracted that accounted for approximately 43% of the total test variance. Specifically, using a minimum loading criterion of 0.40, eight items were found to load on factor one, four items on factor two, six items on factor three, four items on factor four, four items on factor five, and five items on factor six. It was also determined that four items did not yield values <0.40,

Table 11

Total Variance Explained by Factors

	Initial Eigenvalues		
Component	Total	% of Variance	Cumulative %
1	6.041	17.261	17.261
2	2.437	6.964	24.225
3	1.865	5.328	29.553
4	1.690	4.828	34.381
5	1.564	4.469	38.849
6	1.505	4.301	43.150
7	1.405	4.015	47.166
8	1.292	3.692	50.858
9	1.272	3.634	54.492
10	1.168	3.336	57.828
11	1.129	3.227	61.055
12	1.049	2.998	64.052
13	0.987	2.819	66.872
14	0.969	2.770	69.642
15	0.877	2.507	72.148
16	0.813	2.323	74.472
17	0.796	2.276	76.748
18	0.745	2.130	78.877
19	0.719	2.053	80.930
20	0.650	1.858	82.788
21	0.623	1.781	84.570
22	0.589	1.682	86.251
23	0.562	1.605	87.857
24	0.520	1.487	89.344
25	0.472	1.350	90.694
26	0.457	1.305	91.998
27	0.442	1.262	93.260
28	0.396	1.130	94.390
29	0.379	1.083	95.474
30	0.314	0.896	96.370
31	0.292	0.833	97.203
32	0.283	0.810	98.013
33	0.263	0.753	98.765
34	0.223	0.637	99.403
35	0.209	0.597	100.000

therefore, they did not load on any of the six factors.

In contrast, after using the same method to extract nine factors that accounted for approximately 54% of the total test variance, Table 12 shows that four items were found to load on factor one, six items on factor two, five items on factor three, four items on factor four, two items on factor five, three items on factor six, three items on factor seven, four items on factor eight, and three items on factor nine. In addition, items 32, 34, and 35 were found to have cross-loaded on two factors. It was also determined that items 6, 10, 13, and 24 did not yield values <0.40, therefore they did not load on any of the nine factors. Furthermore, no patterns were evident between items on a particular cluster and their loadings on one or more of the nine extracted factors. Of the nine factors extracted, one factor had items loading from all four clusters (e.g., Pattern Completion, Reasoning by Analogy, Serial Reasoning, and Spatial Visualization) of the NNAT, two factors had items loading from three clusters of the NNAT, five factors had items loading from two clusters of the NNAT, and only one factor had items loading from the same cluster.

Summary

Overall, the results on the NAI total score for the 3rd and 4th grade ELL Mexican American sample demonstrated adequate internal consistency. In addition, only the results of the 3rd and 4th grade ELL Mexican American sample on the Spatial Visualization cluster approached an adequate level of internal consistency reliability. Results on the Pattern Completion, Reasoning by Analogy, and Serial Reasoning clusters by both the 3rd and 4th grade ELL Mexican American sample were all below the acceptable level of 0.80. In addition, when pooled reliability estimates for the 3rd and 4th grade ELL Mexican American sample were compared with those derived for the 3rd and 4th grade standardization sample, no statistically significant differences were found.

When performance of the ELL Mexican American sample was compared with the standardization sample with respect to the NAI total score and the four cluster scores, no statistically significant differences were found. However, only the performance of the ELL Mexican American sample on the Pattern Completion cluster approached

Table 12

Pattern Coefficients for Nine Extracted Factors

	1	2	3	4	5	6	7	8	9
Q3	.359	-.052	-.122	-.134	**.543**	-.037	.147	.074	-.084
Q4	-.213	-.082	-.052	.123	**.857**	.146	.036	-.044	.051
Q5	.078	-.108	-.146	**.715**	-.161	.179	-.049	.235	-.015
Q6	.275	.107	-.289	.312	.069	.124	.273	-.129	.380
Q7	**.410**	.137	-.006	-.224	.293	.285	.103	.032	.031
Q8	.106	.102	-.001	**.638**	.213	-.236	-.110	.008	-.068
Q9	-.096	.103	-.134	.061	.147	**.782**	.035	-.104	-.051
Q10	-.010	.178	.109	.324	.358	.166	-.221	-.108	.394
Q11	.258	.020	.096	-.158	.113	.266	.091	.029	**.569**
Q12	.229	**.626**	-.107	-.038	-.319	.226	-.148	.085	.013
Q13	.301	.221	.046	-.295	.232	.090	-.289	.108	.133
Q14	**.700**	.187	.116	-.024	-.069	-.169	-.104	-.153	.171
Q15	**.849**	-.056	-.052	.092	-.120	-.064	.101	.043	.131
Q16	.319	-.095	-.216	.226	.126	.179	-.036	-.095	**-.480**
Q17	**.592**	-.153	.038	.242	.016	.304	.111	-.089	-.172
Q18	.132	-.113	.308	-.048	-.059	**.502**	-.039	.016	.122
Q19	-.045	**.442**	.033	.064	.262	-.267	.260	.097	-.042
Q20	.022	**.858**	-.243	-.016	.015	-.018	-.034	.180	.087
Q21	.142	-.392	**.482**	.125	.092	-.201	.160	.256	.276
Q22	.101	**.400**	.345	.001	-.169	-.067	.398	-.269	.060
Q23	.333	-.044	-.024	.150	.171	-.239	-.063	**.539**	-.060
Q24	.024	.238	.206	.002	.299	-.045	-.044	.228	-.118
Q25	-.172	.108	.242	**.410**	.060	.187	.097	.042	-.033
Q26	-.002	.227	.160	.112	-.342	.196	.093	**.465**	.129
Q27	.071	-.001	**.730**	.092	.024	.027	-.191	-.178	-.076
Q28	-.124	**.644**	.101	.096	.085	.076	.097	.148	-.170
Q29	.086	-.021	.221	-.027	.106	.166	**.589**	.094	-.052
Q30	.280	.059	-.080	-.055	-.093	-.007	**.570**	.208	-.075
Q31	-.055	-.187	**.775**	-.142	-.117	.031	.037	-.076	.030
Q32	.046	.087	**.470**	**.442**	-.025	-.071	-.133	.091	-.014
Q34	-.347	-.053	-.047	-.010	.202	**.443**	-.027	**.553**	.162
Q35	-.115	**.462**	-.240	.092	-.076	-.149	.107	**.710**	.014
Q36	-.086	.033	**.404**	-.054	.019	.098	.327	.057	-.108
Q37	.181	.041	.257	.110	-.167	.087	**-.644**	.100	-.233
Q38	-.124	.127	.288	-.099	.130	.149	.038	-.115	**-.541**

Note. Values in bold indicate item factor loadings <0.40

statistical significance when compared to the standardization sample.

Finally, goodness-of-fit statistics yielded from the CFA indicated a lack of fit between the NNAT's hypothesized four-factor measurement model and the observed data. Because no additional items were lost when the EFA was conducted with nine factors and there were a minimum of two items loading on each of the nine extracted components, a nine factor structure model was determined to be most representative of this sample of ELL Mexican American children.

CHAPTER V

DISCUSSION AND CONCLUSIONS

The purpose of this study was to investigate the reliability and validity of the results of the Naglieri Nonverbal Ability Test (NNAT; Naglieri, 1997a) with a sample of 122 English Language Learner (ELL) Mexican American children. Hispanic Americans currently represent the fastest growing minority group in the United States while Mexican Americans represent the largest of the Hispanic American subgroups. Due to their vast diverse cultural and linguistic factors, Mexican American children often present unique and challenging issues to assessment personnel throughout the United States which make accurate psychoeducational and/or psychological evaluations more difficult to achieve.

There are various reasons why Mexican American children are frequently being subjected to inappropriate assessment practices. Among them is a lack of statistically-sound, culturally-sensitive psychometric instruments to evaluate this population as well as a lack of trained bilingual assessors proficient in the child's native language to conduct sound bilingual psychoeducational and/or psychological evaluations. The rationale for conducting this study was the lack of adequate psychometric and psychoeducational instruments to evaluate Mexican American children and youth and the need to identify more appropriate assessment measures with a diverse student population.

Mexican Americans are one group of culturally and linguistically diverse individuals that historically have been misdiagnosed and overrepresented in special education classrooms. Although they have been involved in a number of precedent setting cases regarding discriminatory assessment practices, the lack of adequate intelligence and/or ability measures available to assessment personnel that meet the standards set forth by experts in the field of psychology still exists today (AERA, 1999). The NNAT was developed to provide testing personnel throughout the country with a

culturally sensitive assessment tool for our growing culturally and linguistically diverse school-aged population. Although a culturally and linguistically diverse group was included in the NNAT's standardization sample, no specific data regarding the reliability and validity of the NNAT with an ELL Mexican American sample were provided. Therefore, the present study was designed to specifically investigate the reliability and validity properties of the results of the NNAT with ELL Mexican American children.

Reliability and Validity of the NNAT

Research question 1 addressed the internal consistency of the NNAT with ELL Mexican American children. In this study, the internal consistency coefficients were calculated using Cronbach's alpha and using Feldt's (1980) formula for computing reliability coefficients for two independent groups. The Cronbach's alpha correlations did yield positive evidence for the internal consistency for both the 3rd and 4th grade ELL Mexican American sample on the Naglieri Ability Index (NAI) total score. However, the Cronbach's alpha correlations did not yield positive evidence for the internal consistency of three of the four cluster scores. Only Spatial Visualization, for both 3rd and 4th grade ELL samples, approached the reliability standard deemed acceptable for tests of cognitive ability.

These results do not differ from the results obtained by Naglieri (1997b) for the 3rd and 4th grade standardization sample in that only the reliability coefficients for the NAI total score yielded positive evidence for the internal consistency. Similarly, only Spatial Visualization, for both the 3rd and 4th grade standardization sample, approached the reliability standard deemed acceptable for tests of cognitive ability.

In addition, V. Willson (personal communication, September 21, 2004) provided a formula so that pooled reliability estimates for 3rd and 4th grade students in the ELL Mexican American sample and the standardization sample could be computed and compared to determine whether statistically significant differences existed. F-values did not indicate that statistically significant differences existed between the pooled reliability estimates of the 3rd and 4th grade ELL Mexican American sample and the 3rd and 4th grade standardization sample.

Research question 2 addressed whether the performance of the sample of ELL Mexican American children with respect to the four cluster scores and the overall NAI score of the NNAT was significantly different from the performance of the standardization sample. The mean NAI total score difference between the 3rd grade ELL sample and the 3rd grade standardization sample was 0.46 while the mean differences for the cluster scores were 0.05 (Pattern Completion), -0.07 (Reasoning by Analogy), -0.12 Serial Reasoning, and 0.58 (Spatial Visualization) respectively. There were no statistically significant differences found between the performance of the 3rd grade ELL sample and the 3rd grade standardization sample on the NAI total score or on each of the four cluster scores.

The mean NAI total score difference between the 4th grade ELL sample and the 4th grade standardization sample was 0.27 while the mean differences for the cluster scores were –0.32 (Pattern Completion), 0.11 (Reasoning by Analogy), -0.26 (Serial Reasoning), and 0.64 (Spatial Visualization) respectively. Although there were no statistically significant differences in performance between the 4th grade ELL sample and the 4th grade standardization sample, differences in performance on Pattern Completion did approach statistical significance, yielding a *t*-statistic of 1.94 and a *p*-value of 0.0525.

Research question 3 addressed how well the hypothesized four-factor structure model of the NNAT fit with the observed data obtained from the ELL Mexican American sample by conducting a confirmatory factor analysis. Given that the probability of obtaining the derived chi-square value that would be equal to or greater than 688.568 was less than .0001 as well as the fact that all of the comparative indexes of fit were well below a value of .95, the hypothesized four-factor structure model of the NNAT did not fit well with the data obtained from the ELL Mexican American sample.

In an effort to investigate possible factor structure of the NNAT with the ELL Mexican American sample, an exploratory factor analysis was also conducted. Using a minimum loading criterion of 0.40, a nine factor model (which accounted for approximately 54% of the total test variance) was found to be most representative of the

ELL Mexican American sample. Specifically, four items were found to load on factor one, six items on factor two, five items on factor three, four items on factor four, two items on factor five, three items on factor six, three items on factor seven, four items on factor eight, and three items on factor nine.

Limitations of the Study

This study has a number of limitations that can affect the generalizability of the results. First, participants in this study were not randomly selected from the Mexican American population; the sample was one of convenience obtained from four public elementary schools in a rural Southwest region of the United States. The small sample size of 122 children that only included 3rd and 4th grade students represents another limitation to the study. In addition, no data on parental level of education was collected. Therefore, generalizations derived from the results of this study should be limited to students with similar characteristics to those included in this study.

Implications for Future Research

Because the NNAT is considered a test of general cognitive ability and is most often used as a screener within the public schools to determine a child's need for additional testing and subsequent educational placement (eligibility for gifted and talented programs, Special Education services), its utility as a predictor of future school success should be further evaluated by examining its concurrent validity with school achievement measures most often used by public school assessment personnel throughout the United States. In addition, comparison studies with other nonverbal intelligence and/or ability measures would provide more information regarding the construct validity properties of the NNAT with similar measures.

As stated earlier, Hispanic Americans represent a culturally and linguistically diverse group with vast intraindividual differences with regard to acculturation levels, English/Spanish language proficiency levels, socioeconomic status, etc…As such, the reliability and validity properties of the NNAT should also be examined using individuals of different Hispanic American subgroups. Examining these statistical properties with other ELL identified individuals such as Vietnamese, Hmong,

Cantonese, and Korean would shed additional light on the culturally and linguistically sensitive aspects of the NNAT with foreign language speaking individuals. Finally, because levels of language proficiency in both English and Spanish were not controlled for, future research on the reliability and validity properties of the NNAT with linguistically diverse populations may seek to control for the influences of different levels of language development on performance on the NNAT by grouping subjects according to their levels of language proficiency in both English and Spanish.

The results of this study indicate problems with the internal consistency of the four clusters with a sample of ELL Mexican American children. Only the NAI total score yielded adequate internal consistency reliability coefficients. In addition, the hypothesized four-factor structure of the NNAT did not fit well with the observed data obtained with this ELL Mexican American sample. An Exploratory Factor Analysis conducted with this sample indicated that a nine-factor model of the NNAT appeared to be most representative of this sample of 3rd and 4th grade ELL Mexican American sample. In addition, when factor loadings were examined for patterns, no patterns between items and each of the nine factors were yielded. These data, coupled with the fact that the theoretical foundations upon which the four clusters were based are relatively unknown, stresses that conclusions made concerning individual strengths and weaknesses based on the performance on individual clusters should be discouraged. Given that the NNAT offers a relatively narrow scope of general cognitive ability speaks to the need to use this measure in conjunction with other multidimensional nonverbal measures of cognitive ability so as to provide a more comprehensive profile of a culturally and/or linguistically diverse child's cognitive abilities.

REFERENCES

American Educational Research Association (AERA), American Psychological Association, The National Council on Measurement in Education. (1966). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association, The National Council on Measurement in Education. (1974). *Standards for educational and psychological tests and manuals.* Washington, DC: American Psychological Association.

American Educational Research Association (AERA), American Psychological Association, The National Council on Measurement in Education. (1985). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

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

Anastasi, A. (1988). *Psychological testing* (6th ed.). New York: Macmillan.

Anastasi, A., & Urbina, S. (1997). *Psychological testing* (7th ed.). Englewood Cliffs, NJ: Prentice Hall.

Arreola v. Santa Ana Board of Education. No. 160-577 (Orange County, CA 1968).

Artiles, A. J., & Trent, S. C. (1994). Overrepresentation of minority students in special education: A continuing debate. *The Journal of Special Education, 27,* 410-437.

Athanasiou, M.S. (2000). Current nonverbal assessment instruments: A comparison of psychometric integrity and test fairness. *Journal of Psychoeducational Assessment, 18,* 211-229.

Baca, L. M., & Cervantes, H. T. (1998). *The bilingual special education interface.* Columbus, OH: Prentice-Hall, Inc.

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

Bersoff, D. N. (1982). The legal regulation of psychology. In C. R. Reynolds & T. B. Gutkin (Eds.), *The handbook of school psychology* (pp. 1043-1074). New York: John Wiley.

Bollen, K. A. (1986). Sample size and Bentler and Bonnett's nonnormed fit index. *Psychometrika, 51,* 375-377.

Bollen, K. A. (1989). A new incremental fit index for general structural equation models. *Sociological Methods & Research, 17,* 303-316.

Bracken, B. A., & McCallum, R. S. (1998). *Universal Nonverbal Intelligence Test.* Itasca, IL: The Riverside Publishing Company.

Brown, L., Sherbenou, R.J., & Johnsen, S.K. (1982). *Test of Nonverbal Intelligence: A language free measure of cognitive ability.* Austin, TX: Pro-Ed.

Brown, L., Sherbenou, R.J., & Johnsen, S.K. (1997). *Test of Nonverbal Intelligence: A language free measure of cognitive ability.* (3rd ed.) Austin, TX: Pro-Ed.

Brown v. Board of Education of Topeka, 347 U.S. 483 at 494 (1954).

Browne, M. W., & Cudeck, R. (1993). Alternative ways of assessing model fit. In K.A. Bollen & J.S. Long (Eds.), *Testing structural equation models* (pp. 136-162). Newbury Park, CA: Sage.

Byrne, B. M. (2000). *Structural equation modeling with AMOS.* Mahwah, NJ: Lawrence Erlbaum Associates.

Carrasquillo, A. L. (1991). *Hispanic children and youth in the United States: A resource guide.* New York: Garland Publishing.

Carroll, J.B. (1993). *Human cognitive abilities: A survey of factor-analytic studies.* New York: Cambridge University Press.

Casso, H. (1973). *A descriptive study of three legal challenges for placing Mexican American and other linguistically and culturally different children into educably mentally retarded classes.* Unpublished doctoral dissertation, University of Massachusetts, Amherst.

Chamberlain, C., & Medeiros-Landurand, P. (1991). Practical considerations for the assessment of LEP students with special needs. In E.V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students (*pp. 111-156). Austin, TX: Pro-Ed.

Clarizio, H. F. (1982). Intellectual assessment of Hispanic children. *Psychology in the Schools, 19,* 61-71.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). San Diego, CA: Academic Press.

Covarrubias v. San Diego Unified School District. No. 70394-T (S.D. Cal. 1971).

Cummins, J. (1980). The entry and exit fallacy in bilingual education. *NABE Journal, 4*(3), 25-59.

Cummins, J. (1984). *Bilingualism and special education: Issues in assessment and pedagogy.* San Diego, CA: College-Hill Press.

Cummins, J. (1986). Psychological assessment of minority students: Out of context, out of focus, out of control? *Journal of Reading, Writing, and Learning Disabilities International, 2(*1), 9-19.

Cziko, G. A. (1990). The evaluation of bilingual education: From necessity and probability to possibility. *Educational Researcher, 21,* 10-15.

Deno, E. (1970). Special education for the mildly retarded: Is much of it justifiable? *Exceptional Children, 23,* 229-237.

Diana v. State Board of Education, Civil Action No. C-70-37 (N. D. Cal.) (1970).

Education for All Handicapped Children Act (1975). P. L. 94-142, 20 U.S.C., 1400-1485, 34 CFR-300.

Feldt, L.S. (1980). A test of the hypothesis that Cronbach's alpha reliability coefficient is the same for two tests administered to the same sample. *Psychometrika, 45,* 99-105.

Figueroa, R. A. (1990). Best practices in the assessment of bilingual children. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology-II (*pp. 93-106). Washington, DC: National Association of School Psychologists.

Figueroa, R. A., & Hernandez, S. (2000). *Testing Hispanic students in the United States: Technical and policy issues.* Washington, DC: President's Advisory Commission on Educational Excellence for Hispanic Americans-Commission Assessment Committee.

Garcia, S. B., & Ortiz, A. A. (1988). *Preventing inappropriate referrals of language minority students to special education.* (Report No. 300860069). Silver Spring, MD: National Clearinghouse for Bilingual Education. (ERIC Document Reproduction Service No. ED309591)

Garretson, O.K. (1928). Study of the cause of mental retardation among Mexican American children. *Journal of Educational Psychology, 19,* 31-40.

Gould. S. J. (1981). *The mismeasure of man.* New York: Norton.

Guadalupe Organization v. Tempe Elementary School District, No. 3, Civ. No. 71-435 (D. AZ) (1972).

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

Harris, A. M., Reynolds, M. A., & Koegel, H. M. (1996). Nonverbal assessment: Multicultural perspectives. In L. A. Suzuki, P.J. Meller, & J.G. Ponterotto, (Eds.), *Handbook of multicultural assessment: Clinical, psychological, and educational applications* (pp.223-252). San Francisco: Jossey-Bass.

Heller, K.A., Holtzman, W. H., & Messick, S. (1982). *Placing children in special education: A strategy for equity.* Washington, DC: National Academic Press.

Hobson v. Hansen, 269 F. Supp. 401 (D. C. 1967) aff'd sub nom., Smuck v. Hobson, 408 F. 2d 175 (D.C.Cir.)(1969).

Holtzman, W.H. Jr., & Wilkinson, C. Y. (1991). Assessment of cognitive ability. In E.V. Hamayan & J.S. Damico (Eds.), *Limiting bias in the assessment of bilingual students.* (pp. 247-280). Austin, TX: Pro-Ed.

Hu, L., & Bentler, P.M. (1995). Evaluating model fit. In R. Hoyle (Ed.), *Structural equation modeling: Issues, concepts, and applications* (pp. 76-99). Newbury Park, CA: Sage.

Hu, L., & Bentler, P.M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: conventional criteria versus new alternatives. *Structural Equation Modeling: A Multidisciplinary Journal, 6*(1), 1-55.

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

Jones, R. L. (1976). *Mainstreaming and the minority child.* Minneapolis: Council for Exceptional Children.

Kaufman, A. S., & Kaufman, N. L. (1993). *Kaufman Adolescent and Adult Intelligence Test.* Circle Pines, MN: American Guidance Service.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services 2000-2001 summary report.* Retrieved August 28, 2004, from George Washington University Web site: http://www.ncela.gwu.edu/

Kretschmer, R. E. (1991). Exceptionality and the limited-English-proficient student: Historical and practical contexts. In E. V. Hamayan & J. S. Damico (Eds.), *Limiting bias in the assessment of bilingual students (*pp. 1-38). Austin, TX: Pro-Ed.

Leiter, R. G. (1979). *Leiter International Performance Scale instruction manual.* Chicago: Stoelting.

Maldonado-Colon, E. (1986). The communication disordered Hispanic child. *Monograph of the BUENO Center for Multicultural Education, 1*(4), 59-67.

McCallum, R. S., Bracken, B. A., & Wasserman, J. D. (2001). *Essentials of nonverbal assessment.* New York: John Wiley & Sons.

Naglieri, J. A. (1985a). *Matrix Analogies Test-Expanded Form*. San Antonio, TX: Psychological Corporation.

Naglieri, J. A. (1985b). *Matrix Analogies Test-Short Form*. San Antonio, TX: Psychological Corporation.

Naglieri, J. A. (1997a). *Naglieri Nonverbal Ability Test.* San Antonio, TX: Harcourt Brace.

Naglieri, J. A. (1997b). *Naglieri Nonverbal Ability Test: Multilevel technical manual.* San Antonio, TX: Harcourt Brace.

Naglieri, J. A. (1997c). *Naglieri Nonverbal Ability Test: Multilevel norms booklet.* San Antonio, TX: Harcourt Brace.

Naglieri, J.A., & Ronning, M. E. (2000). Comparison of white, African American, Hispanic, and Asian Children on the Naglieri Nonverbal Ability Test. *Psychological Assessment, 12,* 3, 328-334.

National Association of School Psychologists (NASP). (1997). *Professional conduct manual for school psychologists containing the principles for professional ethnics and the standards for the provision of school psychological services.* Bethesda, MD: Author.

Ochoa, S. H., Powell, M. P., & Robles-Pina, R. (1996). School psychologists' assessment practices with bilingual and limited English proficient students. *Journal of Psychoeducational Assessment, 14,* 250-275.

Ortiz, A., & Polyzoi, E. (1986). *Characteristics of limited English proficient Hispanic students in programs for the learning disabled: Implications for policy, practice, and research. Part I. Report summary.* (Eric Reproduction No. ED 267578).

Padilla. A.M., & Aranda, P. (1974). *Latino mental health: Bibliography and abstracts.* Rockville, MD: Alcohol, Drug Abuse, and Mental Health Administration.

Parrish, T. (2002). Disparities in the identification, funding, and provision of special education. In D. J. Losen & G. Orfield (Eds.), *Racial inequity in special education* (pp. 15-38). Cambridge, MA: Harvard University Press.

Ramirez, J. D. (1992). Executive summary of volumes I and II of the final report: Longitudinal study of structured English immersion strategy, early-exit and late-exit transitional bilingual education programs for language-minority children. *Bilingual Research Journal, 16,* 1-62.

Raven, J.C. (1938). *Raven's Standard Progressive Matrices.* London: H.K. Lewis.

Raven, J.C. (1947). *Raven's Coloured Progressive Matrices.* London: H.K. Lewis.

Raven, J.C., Court, J.H., & Raven, J. (1986). *Manual for Raven Progressive Matrices and Vocabulary Scales.* London: H.K. Lewis.

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In C. R. Reynolds and T. B. Gutkin (Eds.), *Handbook of school psychology* (2nd ed., p. 519). New York: Wiley.

Roid, G.H., & Miller, L. J. (1997). *Leiter International Performance Scale-Revised: Examiner's manual.* Wood Dale, IL: Stoelting.

Sattler, J. M. (1988). *Assessment of children.* (3rd ed.), San Diego, CA: J.M. Sattler.

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

Smith, M. (1998). *What does Cronbach's alpha mean?* Retrieved August 31, 2004, from the University of Southern California, Academic Technologies Services Web site: http//www.ats.ucla.edu/stat/spss/

Tucker, L. R., & Lewis, C. (1973). A reliability coefficient for the maximum likelihood factor analysis. *Psychometrika, 38*, 1-10.

U.S. Department of Education. (1977). Assistance to states for education of handicapped children: Procedures for evaluating specific learning disabilities. *Federal Register, 42,* 65083.

U.S. Department of State International Information Programs. (2004, July 15). *Census Bureau facts for Hispanic heritage month.* Retrieved September 1, 2004, from http://www.usinfo.state.gov/usa/diversity/

U.S. Office of Education, Office of Bilingual Education (1974). Manual for application for grants under bilingual education. In A. H. Leibowitz (Ed.), *The bilingual education act: A legislative analysis* (p. 37). Rosslyn, VA: National Clearinghouse for Bilingual Education.

Valencia, R. R, & Suzuki, L. A. (2001). *Intelligence testing and minority students: foundations, performance factors, and assessment issues.* Thousand Oaks, CA: Sage Publications, Inc.

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

Willig, A. (1986). Special education and the culturally and linguistically different child: An overview of issues and challenges. *Reading, Writing, and Learning Disabilities, 2,* 161-173.

APPENDIX A

Combined True Score Variance

$$S_{TX}^2 = \frac{(n_3 - 1)S_{T3}^2 + (n_4 - 1)S_{T4}^2}{n_3 + n_4 - 2}$$

True Score Variance-Grade 3

$$S_{T3}^2 = r_3\,(S_{X3}^2)$$

Observed Score Variance-Grade 3

$$S_{X3}^2 = \frac{SEM_3^2}{1 - r_3}$$

True Score Variance-Grade 4

$$S_{T4}^2 = r_4 (S_{X4}^2)$$

Observed Score Variance-Grade 4

$$S_{X4}^2 = \frac{SEM_4^2}{1 - r_4}$$

Grand Means-Grades 3 and 4

$$\overline{X_{..}} = \frac{n_3 \overline{X_3} + n_4 \overline{X}_4}{n_3 + n_4}$$

Variance of Means-Grades 3 and 4

$$S_{\overline{X}}^2 = \frac{n_3(\overline{X_3} - \overline{X_{..}})^2 + n_4(\overline{X_4} - \overline{X_{..}})^2}{n_3 + n_4}$$

Total Variance

$$S_{\tilde{X}}^2 = \frac{(n_3 - 1)S_{x3}^2 + (n_4 - 1)S_{x4}^2}{n_3 + n_4 - 2} + \frac{n_3(\overline{X}_3 - \overline{X_{..}})^2 + n_4(\overline{X_4} - \overline{X_{..}})^2}{n_3 + n_4}$$

Pooled Reliability Estimate

$$\hat{P}_{3+4} = \frac{S_{TX}^2 + S_{\overline{X}}^2}{S_{\tilde{X}}^2 + S_{\overline{X}}^2}$$

VITA

Carlo Arlan Villarreal
402 Ebony Lane
Port Isabel, TX 78578

Education

1998 to 2005	Ph.D., School Psychology Program (APA Accredited), Texas A&M University, Major: School Psychology Emphasis: Disabled and At-Risk Hispanic Children and Youth Co-chairs: Salvador Hector Ochoa, Ph.D. & Michael J. Ash, Ph.D.
1991 to 1993	B.A., Psychology Department, Baylor University, Major: Psychology Emphasis: Clinical Child Psychology

Work Experience

July, 2003 to July, 2004	Pre-doctoral Psychology Intern Childrens Hospital Los Angeles University of Southern California/University Affiliated Program