

ENHANCING THE VALIDITY OF INTELLECTUAL ASSESSMENT: THE  
EFFICACY OF THE UNIVERSAL NONVERBAL INTELLIGENCE TEST  
FOR THE HMONG POPULATION

A Seminar Paper Presented to  
the Graduate Faculty of the  
College of Liberal Studies  
University of Wisconsin-La Crosse

In Partial Fulfillment  
of the Requirements for the Degree  
Master of Science in Education-School Psychology

By  
Christopher Keiler

May 2000

W<sup>c</sup>  
00  
.K4  
c.2.

UNIVERSITY OF WISCONSIN-LA CROSSE  
COLLEGE OF LIBERAL STUDIES

SEMINAR PAPER APPROVAL

Christopher D. Keiler

Student

I recommend acceptance of this seminar paper in partial fulfillment of requirements for the degree of Master of Science in Education-School Psychology.

Milton J. Dehn  
Seminar Paper Advisor

4/24/00  
Date

This seminar paper is approved.

John E. Mazern  
Dean, College of Liberal Studies

25 April 2000  
Date

David Tymeson  
Director, University of Graduate Studies

5-2-00  
Date

## Abstract

Keiler, C.D. Enhancing the Validity of Intellectual Assessment: The Efficacy of the Universal Nonverbal Intelligence Test for the Hmong Population. Master of Science in Education-School Psychology, May 2000, 55 pp.

This paper evaluates the Universal Nonverbal Intelligence Test (UNIT), both in comparison to other contemporary nonverbal scales of intelligence and as a potential means of valid and fair assessment for the Hmong population. Issues surrounding the testing of minority populations, including potential sources of bias in intelligence testing, current practices of intellectual assessment for limited English proficient (LEP) populations, and the efficacy of nonverbal measures of intelligence for Hmong and LEP populations, are explored throughout this paper. The author concludes that the UNIT has remarkable developmental procedures and psychometric properties. In addition, the author suggests that the UNIT functions as an appropriate instrument for evaluating the intellectual skills of Hmong children. A final section includes recommendations for practitioners and future research.

## TABLE OF CONTENTS

Chapter	Page
I. INTRODUCTION.....	1
Key Terminology.....	2
Potential Sources of Bias.....	4
The Rising LEP Population.....	7
Characteristics of the Hmong Population.....	8
Problems Associated with Traditional, Verbally-Loaded Tests.....	12
Relevant Legislation and Litigation.....	14
Using Interpreters and Translated Tests.....	17
A Historical Analysis of Nonverbal Tests.....	20
The UNIT's Normative Data: The Nonrepresentation of the Hmong.....	24
Summary.....	25
Significance of the Problem.....	26
Research Questions.....	26
II. REVIEW OF RELATED RESEARCH.....	27
Validity of IQ Tests for ESL Students.....	28
Evaluation of Nonverbal Scales of Intelligence.....	31
General Information.....	31
Theoretical Underpinnings.....	34
Standardization Procedures.....	35
Psychometric Properties.....	37
Summary of Nonverbal Scales of Intelligence.....	39
IQ Tests for the Hmong Population.....	40
Summary.....	42
III. DISCUSSION.....	44
Intellectual Assessment for the LEP Population.....	44
The Universal Nonverbal Intelligence Test.....	46
The Utility of the UNIT for the Hmong Population.....	47
Implications for School Psychologists.....	49
Gaps in the Literature and Recommendations for Future Research.....	50
REFERENCES.....	52

## CHAPTER I

### Introduction

Assessment of minority children has been of great interest to psychometricians for decades. The appropriateness of conventional IQ tests for various minority populations (e.g., African American, Hispanic, and Asian) has been discussed and debated in the professional literature, bringing about little consensus. Researchers have suggested mean differences in composite IQ scores of up to 15 points between African Americans and Caucasians (Sattler, 1988). Such drastic differences have been used as evidence to support the notion of bias in intelligence testing.

The following chapter will begin with a brief evaluation of various sources of cultural bias in cognitive assessment. The focus will then shift to an evaluation of the Hmong culture with an emphasis on characteristics that may influence performance on intelligence tests. Next, the chapter will discuss concerns of traditional, verbally-loaded intelligence tests followed by a synopsis of legislation, litigation, law, and professional standards that have resulted from such concerns. Next, the chapter will evaluate the efficacy of interpreters and translated tests in the assessment process of English as a second language (ESL) children. A final preliminary section will discuss the historical development of nonverbal intelligence tests, focusing on their utility as valid measures for ESL populations.

This paper will focus strictly on intelligence test-based assessment of intellectual functioning for ethnic minorities who are presumed to have limited English proficiency (LEP). The intent of the paper is to provide a comprehensive analysis of the utility of such instruments. The importance and significance of supplemental assessment procedures, including classroom and environmental observations, thorough interviews, and language proficiency assessment, will not be discussed. The reader should refer to the following articles for the best practices in holistic assessment of children with limited English skills: Esquivel (1985), Figueroa (1990), Figueroa, Sandoval, & Merino (1984), and Lopez (1997).

### Key Terminology

For the purposes of this paper, several key terms need to be defined to help clarify the intended meaning for the reader. These key terms include:

- 1) Hmong– individuals of Southeast Asian descent (primarily from Laos) whose genetic composition consists of at least 50% Hmong.
- 2) Minority– an individual whose ethnicity does not represent that of the cultural majority. For instance, any ethnic group not predominantly European/Anglo-American would be considered an ethnic minority in the United States.
- 3) English-as-a-Second-Language (ESL)-- Any individual who has learned the English language in a secondary manner. These individuals have had prior experience with a foreign/alternative language independent of the English language. This definition is

intended to constitute a variety of individuals, including both bilingual and monolingual people and is independent of their fluency in the English language.

4) Limited English Proficiency (LEP)– An inability to fluently communicate using the English language. There is tremendous variance in the methods used to measure and classify LEP individuals. For the purposes of this paper, individuals who demonstrate deficiencies in basic interpersonal communicative skills (BICS) and cognitive academic language proficiency (CALPS) as identified by Cummins (1984) will be classified as LEP.

5) Fairness– This term represents the notion that a select group of individuals, such as the Hmong population, will score at a level that is equivalent to that of the overall population on a standardized instrument. Further, it is implied that no biases are created in the content or the administration of the test that will influence the scores for the targeted group.

6) Bias– This term refers to systematic differences in intellectual assessment scores between various groups due to reasons apart from actual intellectual functioning levels. Bias results from the diversity of the population and is assumed to stem from characteristics of test content or administration.

7) Intelligence Test– A scale or instrument used to evaluate the intellectual/cognitive functioning of an individual.

8) Performance Scale– This term refers to a portion of an intelligence test that measures the individual's mental reasoning and processing, typically in a visually-based manner.

9) Nonverbal Intelligence Test– This term refers to intelligence tests that typically do not require receptive or expressive verbal abilities on the part of the examinee. Such tests rely on gestures and pantomimes during administration and frequently have visually-based tasks.

#### Potential Sources of Bias

Several realms of bias in intellectual assessment have been identified by researchers as having potentially detrimental impacts on assessment validity for minority populations. Reynolds and Kaiser (1990) have categorized these arguments into six areas of concern. First, it has been suggested that intelligence tests incorporate culturally-inappropriate content, gearing test materials toward that of the middle class Caucasian population. This position primarily focuses on bias in the individual test items. In response to such concerns, statistical analyses of various measures have revealed that very few test items are culturally biased (Sattler, 1988). In addition, such items are frequently identified during standardization procedures and removed (Reynolds & Kaiser, 1990). Identification and removal of such items has led to minimal changes in mean scores (Reynolds & Kaiser, 1990; Sattler, 1988).

Second, several researchers have pointed out the underrepresentation of minority populations in standardization samples of intelligence tests. For instance, of the 2,200 children composing the normative group of the WISC-R, only 330 minority children were included. Furthermore, 305 of these children were African American (Cummins, 1984). Current psychological instruments, such as the WISC-III, have standardization samples stratified in a manner that adequately represent the population of several minority groups. However, despite better standardization procedures, such instruments have failed to reduce bias as evident by the discrepancy in mean scores between various populations (Roid & Miller, 1997). Therefore, bias must result from factors apart from poor standardization procedures.

Third, it has been suggested that current IQ tests have inequitable social consequences for minorities. According to this perspective, discrimination in the educational and social systems influences performance on standardized tests which, in turn, leads to further discrimination in educational expectations and placement. Litigation including *Diana vs. State Board of Education*, *Guadalupe vs. Tempe*, and *Larry P vs Riles* concluded, in part, that minority populations were overrepresented in special education programs based on the results of IQ tests (Esquivel, 1985; Olmedo, 1981). This approach focuses more on the social consequences of intelligence tests and less on the underlying reason for such biases.

Fourth, it has been argued that IQ tests measure attributes of the individual that may be of less importance to members of minority populations. Schiele (1991) has

contended that the epistemology of African Americans, including spirituality, affect, rhythm, and holistic thinking, is ignored in modern day cognitive testing. In addition, Mercer (1979, as cited in Reynolds & Kaiser, 1990) proposed that IQ tests primarily measure ethnocentrism in minority populations, providing no more information than how much an individual has adapted to the dominant culture.

Fifth, it has been asserted that tests of intellectual aptitude have less predictive validity for minority groups than for middle-class Caucasian individuals. According to this assertion, it is hypothesized that the scores obtained from IQ tests do not adequately predict future academic achievement and performance for minority populations. In a meta-analysis, Reynolds and Kaiser (1990) contend that minimal evidence exists to support the notion of predictive validity bias. They suggest that evidence of predictive validity bias is reportedly infrequent and more often than not tends to favor low-SES and minority populations over the Caucasian population.

Finally, it has been suggested that language biases during test administration have a detrimental impact on performance. In general, the professional research claims that minorities with limited English skills may be hindered on verbally loaded tests. Performance scaled scores have been consistently higher than verbal scaled scores for minority groups (Figuroa, 1990). In addition, ESL students score, on average, 12.5 points higher on performance (nonverbal) scales than on verbal scales of intelligence (Cummins, 1984). Therefore, bias may result from a lack of proficiency with the English language. Such deficiencies may make it difficult for students to understand

the content of the assessment (Cummins, 1984) and may also hinder rapport with an English-speaking examiner (Reynolds & Kaiser, 1990).

In summary, many of the proposed sources of bias in intelligence testing have minimal supporting evidence. In addition, greater amounts of evidence have accumulated to refute such claims. Of the preceding arguments, the most feasible explanation is the proposed language bias. It is important to recognize that many minority groups, such as the Hmong population, have a significant proportion of people who have not become proficient in the English language. The use of verbally-loaded intelligence tests with LEP individuals would undoubtedly have detrimental effects on their perceived abilities. The potential language bias of intelligence tests for the Hmong population will be the focus for the remainder of this paper.

### The Rising LEP Population

The Asian population has been rapidly increasing since the 1970's. Ima and Rumbaut (1995) reported that Asian immigrants composed less than 1% of the population in the 1970's, 1.5% in the 1980's, and were projected to reach 4% in 2000. According to 1990 United States Census data, there were 7.2 million Asians in the United States; 90,082 of these Asian's were Hmong. Wisconsin Census statistics reported 30,992 Hmong in 1994. With the large Hmong population in the United States, it is becoming increasingly important to recognize their educational needs.

According to the 1990 census, 31.8 million Americans spoke a language other than English at home (Lopez, 1997). In addition, an estimated 3 to 5 million students in public schools did not identify English as their primary language (Ima & Rumbaut, 1995). With the rising immigrant population in the United States, the number of children who do not identify English as their primary language will most likely increase. Some of these children will be referred for a psychoeducational assessment. Valid assessment of the LEP population is of utmost importance to professionals in the field of school psychology. Assessment instruments, such as intelligence tests, may have a detrimental impact on the educational experiences and outcomes of the individual. Accurate assessment will result in appropriate educational placement and programming. Therefore, it is becoming more and more important to identify an instrument that appropriately and validly assesses the intellectual functioning of LEP individuals as well as individuals of Hmong descent.

### Characteristics of the Hmong Population

Several unique characteristics of the Hmong population need to be considered in the process of assessment. Such characteristics are divided into two separate sections for the purposes of this paper. The first section will discuss historical circumstances and cultural characteristics that have had an impact on the English literacy level of Hmong immigrants. The second section will discuss Hmong values that have a direct influence on the instrument selected for cognitive assessment.

Ima and Rumbaut (1995) concluded that Southeast Asian individuals are more likely to be LEP than other linguistic groups. In addition, 66% of South East Asians maintained LEP status beyond their junior year of high school as compared to 43% for the Hispanic population (Ima & Rumbaut, 1995). It is evident that individuals of Southeast Asian decent have difficulties acquiring the English language. Many of these difficulties may be due to their recent arrival in the United States, primarily in the 1970's and 1980's (Irwin & Madden, 1986). However, it is important to look at some additional characteristics of Hmong culture that may hinder the acquisition of the English language.

A historical evaluation of Hmong culture provides relevant insight into the low level of literacy among Hmong immigrants. According to Morrow (1989), 81% of Hmong who immigrated from Laos had no formal education upon entering the United States. Such a large percentage reflects the lack of both accessibility and relevance of education for those living in Laos. It is also important to note that prior to the 1950's, Hmong children learned exclusively from family members and villagers and did not attend schools. Consequently, high illiteracy rates exist amongst today's Hmong grandparents and parents (Westermeyer, Bouafuely-Kersey, & Her, 1997). A 1985 survey suggested that approximately 54% of the Hmong population residing in the United States were literate in Hmong and 31% were literate in English (Morrow, 1989). Such numbers are not surprising, considering that the Hmong written language was first developed in the late 1950's (Lewis, Vang, & Cheng, 1989).

Several characteristics of the Hmong language may make learning the English language a difficult process. First, there are several phonological differences between the Hmong language and the English language. Hmong may experience difficulties with the ending/final consonants of words, particularly in the case of consonant clusters (Lewis et. al., 1989). Second, the Hmong language, unlike English, is noninflectional. Therefore, Hmong attempting to learn the English language may struggle in the process of forming tenses and affixes (e.g., related, unrelated, relation, relative). In addition, they may incidentally string several verbs together, use adjectives after nouns, and use infinitives and gerunds incorrectly (Lewis et. al., 1989). After considering these potential problems, it is not difficult to imagine the struggle that Hmong children may encounter while learning the English language. The process may become quite tedious and time consuming. Consequently, a large portion of the Hmong population may have limited English skills. It is important to assess a Hmong child's English proficiency to determine if he/she should 1) be tested in a manner that does not rely on verbally-based cognitive tests and 2) be provided with ESL instruction to assist in his/her educational experiences. It is also important to recognize that the child's Hmong language proficiency may be declining during the learning process of the English language, termed subtractive bilingualism (Dao, 1991). Therefore, it can be assumed that many Hmong children may not have adequate skills in either the Hmong or the English language.

A few noteworthy characteristics of the Hmong population should be considered in the process of cognitive assessment to ensure valid results. First, time is viewed as "elastic" in the Hmong culture (Morrow, 1989). Therefore, examiners should be careful when evaluating Hmong children with timed tests. The child may not recognize or value the importance of quickly and efficiently completing the task at hand. It is recommended that intelligence tests that reduce the importance of speed be used with the Hmong population. Second, Hmong children are unlikely to ask for help when they are confused or do not understand a specific task (Williams, 1987). Consequently, it is the responsibility of the examiner to ensure that the child understands the directions of each subtest. Since the child will be unlikely to verbalize uncertainty, the examiner must pay careful attention to nonverbal signs of confusion. This is especially important for tests that incorporate teaching items at the beginning of subtests. Third, it has been reported that Hmong children may not understand paper and pencil tasks (Williams, 1987). Intelligence tests that incorporate such procedures should be used sparingly and additional directions may be necessary if the standardization procedures allow for them. Fourth, people born in Laos most likely have arbitrarily assigned birthdates which may be several years different than the actual date (Lewis, Vang, & Cheng, 1989; Liu & Li, 1994). This is an important point to consider upon interpreting the results of an intelligence test. Parental contact should be made to confirm the accuracy of the child's birth date. This point should also be considered when extreme scores are obtained on intelligence tests.

In summary, the history of Hmong education and literacy levels when combined with the variability between the structure of the Hmong and English languages provides some insight into the difficulties that many Hmong children experience when attempting to learn the English language. In addition, several factors need to be considered in the selection of cognitive assessment instruments for the Hmong population. An appropriate intelligence test for Hmong children with limited English skills should include primarily performance tasks (i.e., a decreased emphasis on verbal tasks) that do not impose time limits or provide bonus points for speed. In addition, an appropriate test should have clear directions, teaching items, and the possibility for corrective feedback and elaborated directions on teaching items.

#### Problems Associated with Traditional, Verbally-Loaded Tests

Research conducted as early as 1910 has indicated language bias in intelligence testing. The Army Alpha test, which was developed to assist in the cognitive screening of soldiers, was thought to be biased for individuals of low literacy. Consequently, a nonverbal counterpart, the Army Beta, was developed (Thorndike, 1997). In the early 1930's, David Wechsler, a pioneer in the development of intellectual assessment, recognized the problems associated with the verbally-loaded Binet scales and developed the Wechsler-Bellevue, focusing on a combination of both verbal and performance/nonverbal tasks (Thorndike, 1997). Although the focus over time has shifted to an emphasis on a balance of nonverbal/performance and verbal

items, the verbal bias in intelligence testing remains. As previously noted, individuals with limited English skills score consistently higher on nonverbal subtests than verbal subtests on modern day scales (Cummins, 1984; Figueroa, 1990).

Currently, the most frequently used tests for LEP children is an adaptation or a translation of the Wechsler scales (Nuttall, 1987). Several factors most likely influence psychologists to rely on these instruments in the assessment of LEP children. First, it is important to recognize the popularity and familiarity of the Wechsler scales. Most school psychologists have received training in administering the Wechsler scales and, consequently, continue to use them in practice. Second, many school psychologists likely have a lack of knowledge of appropriate tests or methods for assessing the LEP population. Esquivel (1985) suggests that a limited number of school psychologists receive adequate training to work with the LEP population. Therefore, the problem may be related to a lack of appropriate training or knowledge. Finally, school psychologists are most likely drawn in by the availability of translated and renormed Wechsler scales. For instance, the WISC-R has been adapted to meet the needs of children from Mexico, Spain, and Hong Kong (Esquivel, 1985).

Unfortunately, the Wechsler scales have not been effective in eliminating language biases for LEP children. Esquivel (1985) reported that LEP children tend to have lower WISC-R scores in Information, Similarities, and Vocabulary and higher scores in Arithmetic, Digit Span, and performance tasks. It can be argued that a profile

of high performance scaled scores and low verbal scaled scores suggests the possibility of language barriers in testing.

### Relevant Legislation and Litigation

A significant amount of legislation and litigation has identified the concerns with traditional methods of assessing and evaluating minority populations who have limited English skills. The following section provides a brief synopsis of major decisions that have influenced the practices used in intellectual assessment. These cases have led to rules and regulations that have promoted more ethical and valid evaluations of LEP students' cognitive functioning.

Language barriers in intelligence assessment were a primary concern addressed in the 1968 class action suit of *Diana vs. the California State Board of Education*. Nine Mexican-American children with limited English skills were misdiagnosed as mentally retarded based on test scores from the Stanford-Binet and the Wechsler Intelligence Scale for Children. It was argued that 1) the tests were inappropriately administered in English despite the fact that the children were primarily Spanish speaking and 2) the tests administered did not adequately represent the cultural background of the Mexican migrants (Esquivel, 1985). After being retested in their native language, the children experienced an average gain of 15 standard score points; seven of the nine children no longer qualified for the mild mentally retarded criteria (Olmedo, 1981).

The court ruled that the Spanish-speaking children had been inappropriately misrepresented and misplaced in special education programs. Consequently, several provisions were created to alleviate language barriers in testing (Esquivel, 1985). First, the court declared that children must be tested in both their native language and in English. Second, it was determined that the nonverbal sections of intellectual tests are an adequate means to assess the abilities of LEP children. Finally, school districts in California were required to retest children previously diagnosed as mentally retarded using nonverbal sections of tests. *Diana vs. the California State Board of Education* was the initial spark in litigation that identified language as a potential barrier in intelligence testing and special education placement. Several similar cases followed *Diana*, spreading the results to various states and across various ethnic groups.

The 1974 decision in *Lau vs. Nichols* had ramifications on the assessment process of LEP children. Twelve Chinese-American students filed a class action suit against the San Francisco Board of Education on the grounds that Chinese students were not receiving an appropriate education to account for their LEP status. It was argued that the students required special English language classes (English as a second language) with bilingual instructors. The Supreme Court ruled that both educational instruction and evaluation should be provided in the child's native language in order to make education both a meaningful and comprehensible experience (Wang, 1995). This decision furthered the notion that non-English speaking children should be assessed using their primary language.

Public Law 94-142, The Education for All Handicapped Children Act, brought the previously described judicial decisions to the legislative level. It mandated that all psychological tests and supplemental materials be nondiscriminatory in nature and be presented in a manner that reflects the child's native language (Figueroa, Sandoval, & Merino, 1984). This position was reconfirmed in Public Law 99-457 (Education of the Handicapped Amendments) and extended in the Individuals with Disabilities Education Act (IDEA) of 1990. IDEA required that all assessment procedures be 1) nondiscriminatory in nature and 2) psychometrically valid for the intended use (Lopez, 1997).

Professional standards have also been established to ensure valid assessment. The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council on Measurement in Education) contend that assessment of LEP individuals with translated tests is not necessarily psychometrically sound; using such an instrument requires that the test's reliability and validity be identified (Figueroa, 1990; Lopez, 1997). The National Association of School Psychologists further supports this position in the Standards for the Profession of School Psychological Services (1995) claiming that assessment should be conducted with the child's dominant spoken language or communication system and interpreted based on empirical evidence.

In summary, litigation, legislation, and professional standards have created guidelines to assist in appropriate cognitive assessment of LEP children.

Unfortunately, the guidelines provided give little information on specifically what techniques, strategies, and instruments should be used in assessment but, rather, focus on what should not be done. Combined with the fact that school psychologists have traditionally received little training in working with ESL populations, many are helplessly left in the dark. Esquivel (1985) has suggested that providing school psychologists with in-service training, focusing on nonbiased assessment, may assist in greater sensitivity and awareness of the issues surrounding the ESL population. However, the problem may actually be larger than a lack of appropriate training. Another contributing problem resides in the fact that little research has been conducted on the efficacy of both nonverbal tests and translated tests for ESL populations. Therefore, the leaders in the field of ESL assessment may be standing right alongside the practitioners in the dark.

#### Using Interpreters and Translated Tests

As previously stated, legislation, litigation, and professional standards support the use of interpreters and translated tests for the cognitive assessment of ESL children. As a result, these strategies have been readily practiced throughout the nation. The efficacy of such measures will be further evaluated in this section.

Theoretically, testing minority children in their native or primary language via interpreters should eliminate language barriers, reducing the potential for bias. However, a careful analysis of the interpretation process provides insight into several

areas of concern. First, a limited supply of bilingual school psychologists results in an increasing need for third-party interpreters. Aside from qualified school psychologists, interpreters who are familiar with psychometrics and the importance of standardization procedures in cognitive assessment are few and far between. A survey of 21 school agencies confirmed that such a shortage of qualified interpreters exists (Nutall, 1987). Therefore, it is often the case that under qualified interpreters are chosen to assist in cognitive assessment.

Unfortunately, interpreters who do not receive training in psychometrics are presumed to incidentally produce indirect translations and inappropriate reinforcement in testing situations (Lopez, 1995). Indirect translations of test items may produce confusion on the part of the examinee, resulting in an inability to answer the question. Indirect translations of item responses may influence how the particular item is scored. Inappropriate reinforcement may cue the child as to whether or not the response was correct or incorrect. Such feedback may have detrimental impacts on the child's overall test performance.

Second, translations may alter the technical properties of tests because of inequivalencies in languages (Olmedo, 1981). For instance, some words may not have a direct translation or an equivalent level of difficulty. This produces complications that are not easily alleviated. Eliminating such difficulties would require one to re-norm the instrument on the desired population and perform a Rasch item analysis to ensure that the progression of difficulty has not changed between the items.

Several researchers have translated and re-normed IQ tests in hopes of better representing the characteristics of minority/ESL populations. Researchers have developed over 100 Spanish translations (Figuroa, 1990; Olmedo, 1981) including instruments such as the WISC-R (Escala de Inteligencia para Nivel Escolar Wechsler), the K-ABC (Kaufman Bateria de Evaluacion Intelectual), the PPVT (Test de Vocabulario en Imagenes Peabody), and the WJ-COG (Prueba de Habilidad Cognoscitiva of the Bateria Woodcock Psico-Educativa en Espanol). Such tests commonly have inherent flaws in form-equivalency and in standardization procedures. First, the psychometric properties do not necessarily translate as a result of the inequivalence of languages (Olmedo, 1981). As previously noted, Rasch item analysis would have to be performed to evaluate for changes in difficulty. Second, the normative data developed for such instruments tend to be limited in size and representativeness (Olmedo, 1981) and come from foreign nations (Figuroa, 1990). This is problematic because the characteristics of the standardization sample may not reflect that of the population of interest. Because of such poor standardization procedures, the newly-developed tests may have less validity for the population of interest than the original versions. A third and final criticism of translated tests is that the predictive validity of such instruments is relatively unknown (Figuroa, 1990). Therefore, it is difficult to determine if these instruments can accurately predict future performance in the school setting and assist in the appropriate placement of referred children.

In summary, using interpreters and translated versions of tests does not appear to be a viable method in the assessment process of ESL children. A lack of appropriate training for interpreters, inequivalence of various languages, and poor standardization procedures have reduced the validity of such procedures. It may be worthwhile for researchers and practitioners to seek other means of assessment. For instance, nonverbal intelligence tests may help to eliminate the barriers that language has created in the valid assessment of the ESL population.

#### A Historical Analysis of Nonverbal Tests

Evidently, verbally-loaded intelligence tests present several challenges when assessing minority or ESL populations. As previously noted, legislation, litigation, and professional standards have stated that administering such tests in the English language has a detrimental impact for ESL students and, consequently, state that administration should occur in the child's native or primary language. However, several factors make such practices difficult, including a lack of bilingual school psychologists, a lack of trained interpreters, and the possibility for misinterpreted questions and responses (Lopez, 1995; Nuttall, 1987; Olmedo, 1981). It is also important to recognize that the child may not only lack proficiency in English but also in his/her native language. In such instances, translating tests would not function as an appropriate strategy for eliminating language barriers in assessment.

A historical and once-popular methodology for reducing language bias in IQ testing involved administering only the Performance scale of the Wechsler tests. Such practices enable the examiner to remove the bulk of the verbal component, minimizing the possibility for language-based bias. Some research has supported this practice, claiming that minority groups, such as the Hmong, achieve performance scale standard scores that do not deviate significantly from that of the total, normative population (Irwin & Madden, 1986). Although performance scales may appear to be a valid measure of intelligence for children with limited English skills, it is important to recognize a severe limitation. According to Cummins (1984), performance scales provide useful assessment information but do not provide a comprehensive analysis of academic potential. Therefore, relying solely on a performance scale to determine a child's intellectual functioning may not be the most psychometrically valid or appropriate alternative.

Several nonverbal IQ tests have been developed to assist in the comprehensive assessment of children with limited language skills. In this paper, the Leiter International Performance Scale (LIPS) and the Test of Nonverbal Intelligence (TONI) will be evaluated from a historical perspective. In addition, a newly-developed instrument, the Universal Nonverbal Intelligence Test (UNIT) will be compared and contrasted to the most recent updates of the LIPS (Leiter-R) and the TONI (TONI-3).

The TONI (Brown, Sherbenou, & Dollan, 1982) can best be described as a quick, nonverbal intellectual screener for individuals between the ages of 5 and 85.

Administration requires gestures and hand movements as opposed to traditional verbal directions. The TONI includes two separate forms (Form A and B), consisting of 50 items each and requires approximately 15 minutes for administration. In the TONI, examinees are required to recognize relationships among a sequence of black and white abstract designs and identify the answer that completes or fits the relationship from a series of four to six options. All test items are untimed.

Research on the TONI has identified several weaknesses and flaws. Martin, Blair, and Bledsoe (1990) have contested that the two forms may not be psychometrically equivalent based on a weak correlation of 0.56. In addition, the TONI has test-retest reliability coefficients of .89 for form A and .83 for form B; neither meet the suggested .90 criteria required for diagnostic purposes (McGhee & Lieberman, 1990). It is also important to note that the TONI has been criticized for a lack of data supporting the validity and reliability of the instrument for various ethnic groups (Clark, 1985). Finally, the TONI has been criticized by researchers as being rather limited in scope, only measuring a small portion of cognitive functioning. For instance, it has been hypothesized that the TONI is a test that measures primarily visual organization, visual comprehension, and the ability to attend to visual detail (Kowall, Watson, & Madak, 1990).

The LIPS (Leiter, 1979) is another noteworthy nonverbal scale of intelligence for children from 2 to 18 years of age. On the LIPS, children are required to arrange a series of blocks according to given rules including matching/pairing, analogies, and

perceptual patterns. Similar to the TONI, responses on the LIPS are not constrained by time limits. It has been argued that the LIPS is a culture-free measure of intelligence despite a lack of solid evidence (Sattler, 1988). In addition, Lewis and Lorentz (1994) noted that Latino children scored significantly higher on the LIPS than on the Wechsler scales, suggesting that the LIPS may reduce language-based biases.

Unfortunately, the LIPS has received a fair amount of criticism. According to Sattler (1988), the LIPS has several limitations including outdated pictures/materials and norms, uneven item difficulty progression, and an inadequate standardization sample. In addition, it has been argued that the LIPS measures primarily perceptual organization and the ability to make discriminations (Sattler, 1988).

The nonverbal test market has exploded in the late 1990's with the introduction of the TONI-3 and the Leiter-R in 1997 and the UNIT in 1998. These three tests provide great promise in nonverbal assessment of intellectual functioning. There are several commonalities among these tests including nonverbal content and administration as well as representative and well-stratified normative populations. However, it is important to note the unique properties of each test. The TONI-3 remains a screening instrument, requiring only 15 minutes for administration. In addition, the size of the normative sample of the TONI-3 is superior to other nonverbal scales with a total of 3,451 participants. The Leiter-R is a comprehensive instrument that consists of 20 subtests measuring several factors including fluid reasoning, fundamental visualization, attention, recognition memory, associative memory, and

memory span. In addition, the Leiter-R offers social-emotional rating scales for a more comprehensive and holistic evaluation. The UNIT spans a smaller age range than the other two tests (ages 5 to 17). It consists of six subtests measuring four factors including memory, reasoning, symbolic processing, and nonsymbolic processing. These three measures will be further evaluated in the next chapter.

In summary, early versions of the LIPS and the TONI had been routinely criticized. Among these concerns are outdated norms, weak theoretical foundations, and unidimensional focuses (McCallum & Bracken, 1997). Recently developed nonverbal measures of intelligence have addressed many of these concerns and provide a great deal of optimism for language-free, nonbiased assessment.

#### The UNIT's Normative Data: The Nonrepresentation of the Hmong

The UNIT is a new cognitive assessment device that assesses intelligence in both a completely nonverbal and a comprehensive manner. The standardization sample selected for the UNIT adequately represents that of various minority groups, including the ESL population. Unfortunately, the Hmong population was not included in the standardization sample. Therefore, it could be argued that the developed norms may not be appropriate for the Hmong culture.

### Summary

The rising LEP population in the United States has brought about the need for appropriate, nonbiased psychological instruments. It is important to recognize that many Hmong children enrolled in public schools may have limited English and/or Hmong language skills. Consequently, traditional verbally-loaded IQ tests may not adequately and validly assess the intellectual abilities of Hmong students. Legislation, litigation, and professional standards have supported the use of interpreters or translated tests to account for such language barriers. Unfortunately, there are several practical and methodological problems that result from such practices. Therefore, alternate forms of assessment must be considered. For example, language biases in intellectual assessment may be alleviated by using nonverbal scales of intelligence. Historically, nonverbal measures of intelligence have been hindered by poor psychometric properties. However, the new generation of nonverbal tests appear to alleviate such concerns.

The UNIT, a recently developed nonverbal IQ test, has been designed to evaluate comprehensive intellectual skills in an entirely nonverbal manner. The UNIT was standardized with a rather diverse population representing various ethnicities and disabilities as well as various levels of parent education and socio-economic status. Chapter 2 of this paper will further evaluate the UNIT as a measure of intellectual functioning for the Hmong population.

### Significance of the Problem

To date, there is no published research that evaluates the efficacy of a comprehensive intelligence instrument for the Hmong population. This “gap” in the research is of considerable importance with the rapidly growing Hmong population throughout the United States. It is recommended that a set of normative data be created for the Hmong population with a culturally and linguistically appropriate instrument such as the UNIT. Such data would prove useful in the accurate assessment and placement decisions of referred Hmong children. In addition, the data will allow practitioners to compare the individual abilities of a Hmong child to his/her local peers.

### Research Questions

The following research questions will be answered through a literature review and subsequent research:

- 1) How does the Universal Nonverbal Intelligence Test (UNIT) compare to other modern nonverbal scales of intelligence (e g , the Test of Nonverbal Intelligence-3 and the Leiter International Performance Scale-Revised)?
- 2) Can the Universal Nonverbal Intelligence Test (UNIT) function as a valid and fair measure of intellectual/cognitive functioning for Hmong children?
- 3) Do significant mean differences exist between the scores of Hmong children receiving ESL services and those not receiving ESL services?

## CHAPTER II

### Review of Related Research

The American population is becoming more and more ethnically and culturally heterogeneous, resulting in a need for dynamic psychometric instruments. Unfortunately, many traditional and modern intelligence tests may not adequately account for such a dynamic population. Sattler (1988) suggested that mean differences of up to 15 points exist between the composite IQ scores of African Americans and Caucasians. Therefore, it could be argued that biases are created in either the content or the administration procedures of intelligence testing. Several realms of bias have been proposed, including culturally-inappropriate content, culturally-inappropriate criteria, nonrepresentative standardization samples, and language-based biases (Reynolds & Kaiser, 1990). Of these concerns, language-based biases appear to have the most empirical and theoretical support. Ethnic populations with a large proportion of limited English proficient (LEP) individuals, such as the Hmong population, may be hindered by such language-based biases on language-loaded tests. A potential means of reducing or eliminating language-based bias lies in the utilization of nonverbal measures of intelligence.

The following chapter intends to evaluate the validity of a recently-developed nonverbal scale of intelligence, the Universal Nonverbal Intelligence Test

(UNIT), for the Hmong population. The following research questions will be answered through a literature review and subsequent research:

- 1) How does the Universal Nonverbal Intelligence Test (UNIT) compare to other modern nonverbal scales of intelligence (e.g., the Test of Nonverbal Intelligence-3 and the Leiter International Performance Scale-Revised)?
- 2) Can the Universal Nonverbal Intelligence Test function as a valid and fair measure of intellectual/cognitive functioning for Hmong children?
- 3) Do significant mean differences exist between the scores of Hmong children receiving ESL services and those not receiving ESL services?

The following chapter will begin with an evaluation of the validity of various measures of intellectual functioning for English as a second language (ESL) and LEP populations. Next, three nonverbal measures of intelligence, the TONI-3, the Leiter-R, and the UNIT, will be evaluated according to their theoretical underpinnings, standardization procedures, and psychometric properties. The final section of this chapter will focus on the utility of intelligence tests for the Hmong population.

#### Validity of IQ Tests for ESL Students

It has been proposed that ESL children, particularly those who are LEP, may be significantly hindered on verbally-loaded intelligence tests (Cummins, 1984; Esquivel, 1985; Figueroa et al., 1984; Lopez, 1997). Such language-based biases have been demonstrated in the psychological research. Cummins (1984) conducted an analysis of

Wechsler Intelligence Scale for Children-Revised (WISC-R) and Wechsler Preschool and Primary Scale of Intelligence (WPPSI) scores, including 264 sets of Performance IQ scores and 234 sets of Verbal IQ scores, from referred kindergarten and first grade ESL students. A careful evaluation of the data revealed that ESL students performed significantly lower on the Verbal IQ scale (mean = 77.9) than on the Performance IQ scale (mean = 89.1). A breakdown of the Verbal IQ scale revealed that ESL students performed poorest on the Information subtest with a median scaled score of 4.9, followed by Vocabulary and Similarities with medians of 6.0 and 6.5, respectively. The ESL students highest Verbal scale subtest scores included Arithmetic and Digit Span with medians of 7.4 and 7.6, respectively. Median scores on Performance scale subtests did not differ significantly, ranging from a median of 8.1 to 9.0. The data supports the notion that the WISC-R and the WPPSI, two of the most popular measures of intellectual functioning, may have embedded language-based bias for the ESL population.

Current nonverbal scales of intelligence, such as the Test of Nonverbal Intelligence-3 (TONI-3), the Leiter International Performance Scale-Revised (Leiter-R), and the UNIT have been evaluated to determine if they are fair instruments for ESL and LEP populations. Ninety ESL children were included in the standardization sample of the TONI-3. The mean standard score of this particular population was 93, deviating 7 points or approximately one-half of a standard

deviation, from the mean of the entire standardization population (Brown, Sherbenou & Johnsen, 1997). Consequently, the TONI-3 may be more appropriate than traditional, verbally-loaded tests for assessing LEP populations.

During the developmental process of the Leiter-R, mean scores for two groups of ESL students were compared to the mean scores of the standardization population. The first group, composed of 73 ESL children whose dominant language was Spanish, had mean scores ranging from 7.1 to 10.9 on the individual subtests of the Visualization and Reasoning (VR) Battery. The second group, composed of 26 Asian ESL children, had mean scores ranging from 7.0 to 11.0 on the subtests of the VR Battery. According to the manual, these scores were consistent with individuals who have similar levels of parental education. Overall, for both ESL groups, the mean scores were within one-third of a standard deviation of the normative sample mean (Roid, & Miller, 1997).

The efficacy and fairness of the UNIT for the ESL population was evaluated in a study of 78 individuals. The mean Extended Battery scaled score for the ESL population was 93.3. This score falls 6.7 points and approximately one-half of a standard deviation from the population mean. It is also important to note that 90% of the individuals composing this group had a parental education level equivalent to high school or below. This variable may account for some of the minimal variation in the average ESL score as compared to that of the overall population (McCallum & Bracken, 1998).

In summary, traditional, verbally-loaded scales of intelligence, such as the WISC-R and the WPPSI, may underestimate the cognitive abilities of ESL students due to language-based biases. Nonverbal scales of intelligence appear to alleviate such concerns to some degree, as evident by fairness data developed on the TONI-3, the Leiter-R, and the UNIT. School psychologists should consider using nonverbal scales of intelligence in the cognitive assessment of both ESL and LEP children. Such practices will assist in valid and appropriate placement decisions.

#### Evaluation of Nonverbal Scales of Intelligence

The following section will present basic information on nonverbal measures of intelligence including the TONI-3, the Leiter-R, and the UNIT. An analysis of the theoretical foundation, standardization procedures, and psychometric properties will be presented for each instrument. It is important to keep in mind that traditional nonverbal measures of intelligence have been plagued with outdated norms, weak theoretical foundations, and unidimensional focuses (McCallum & Bracken, 1997). Current nonverbal tests of intelligence, including those discussed in this section, should address these concerns.

#### General Information.

The TONI-3 (Brown et al., 1997) can best be described as an intellectual screener for individuals between that ages of 6 years 0 months and 89 years

11 months. The test includes two separate forms, Form A and Form B, each consisting of 45 items arranged in order of increasing difficulty. The entire battery can be completed in approximately 15 minutes. Items on the TONI-3 require the examinee to use complex reasoning strategies to identify relationships among sequences of abstract designs. The examinee must identify the correct response for each item from a series of four to six options. The TONI-3 was developed to comply with Jensen's (1980) guidelines for language-free and culture-reduced assessment. These guidelines suggest 1) the inclusion of performance measures, 2) pantomimed instructions, 3) the inclusion of practice items, 4) unlimited response time, 5) abstract content (to reduce cultural biases), and 6) the inclusion of reasoning and problem solving. The TONI-3 is entirely nonverbal, requiring no verbal directions or responses. In addition, it was developed to be motor-reduced for individuals with physical limitations.

The Leiter-R (Roid & Miller, 1997) is a comprehensive, nonverbal test of intellectual functioning for individuals between the ages of 2 years 0 months and 20 years 11 months. The entire battery consists of 20 subtests with 10 measuring visualization, reasoning, and spatial ability (Visual and Reasoning battery) and 10 measuring nonverbal attention and memory (Attention and Memory battery). Depending on the age of the examinee, between 10 and 17 of these subtests are administered in a comprehensive evaluation. Three types of responses are required from the examinee, including placement of response cards, sequencing or placing

response shapes, and identifying response pictures, all of which are done in an entirely nonverbal and motor-reduced manner. The Leiter-R yields growth scores in addition to the traditionally-used norm-referenced scores. The inclusion of growth scores enables the examiner to assess small gains or improvements in cognitive functioning over time. Additional social-emotional rating scales can be filled out by the examiner, the examinee, the parent(s), and the teacher(s). Such rating scales may be particularly relevant for minority children by helping to provide an evaluation of functioning across various contexts.

The UNIT (McCallum & Bracken, 1998) is an instrument of comprehensive, general intellectual functioning for children and adolescents between the ages of 5 years 0 months and 17 years 11 months. The comprehensive battery of the UNIT includes six subtests measuring both broad reasoning and memory functioning with materials that are either symbolic (i.e., visual stimuli that lend themselves to verbal mediation) or non-symbolic (i.e., abstract, nonmeaningful visual stimuli). Bracken and McCallum (1997) contend that the symbolic materials used in the UNIT may provide insight into an individual's verbal abilities. However, such claims are entirely theoretical and have no research support to date. Administration of the comprehensive battery is estimated at 45 minutes. Three types of responses are required from the examinee including sequencing and placing of response chips, pointing to response pictures or objects, and completing mazes with a pencil. Similar to the TONI-3 and the Leiter-R, the UNIT was developed to be a language-free, motor-reduced

measure of intelligence. However, the fine motor skills required for the maze activities contradicts the motor-reduced claim. The UNIT was developed to 1) be entirely nonverbal, 2) assess several facets of intelligence, 3) measure higher-order processing, and 4) have materials that are interesting and motivating to children (McCallum & Bracken, 1997).

#### Theoretical Underpinnings.

The TONI-3 was developed with little theoretical foundation. The authors contend that the TONI-3 was designed to be a strong measure of Spearman's *g*, or general intelligence, with an emphasis on cognitive agility, adaptability, originality, and flexibility. Factor analysis of the test's 45 items supports the notion of a general factor, accounting for 67% of the variance in Form A and 68% of the variance in Form B (Brown et al., 1997).

The Leiter-R was developed based on the hierarchical model of intelligence developed by Carroll (1993), in which a series of sub-skills or abilities underlie an encompassing general intelligence factor (*g*). Skills measured in the Leiter-R, although varying by age, include fluid reasoning (*Gf*), broad visualization (*Gv*), attention, recognition memory, memory span, and general memory. Factor analysis supports the independence of these skills in the Leiter-R (Roid & Miller, 1997).

The UNIT was developed on a variation of Jensen's (1980) parsimonious model of intelligence, claiming that a general *g*-factor of intelligence exists with two

subordinate factors of broad reasoning and memory. In addition, Bracken and McCallum (1997) theorized that the inclusion of both symbolic and nonsymbolic reasoning and memory allows for a more thorough assessment of cognitive functioning. Consequently, a two-tiered model of intelligence, including broad reasoning and memory with additional symbolic and nonsymbolic dimensions was implemented in the UNIT. Factor analysis supports the theoretical structure of the UNIT. Goodness of fit indicator values (GFI) support the notion of a single factor (GFI = .987), the memory and reasoning breakdown (GFI = .991), and the symbolic and non-symbolic differentiation (GFI = .989) in the UNIT (McCallum & Bracken, 1998). Supplementary factor analyses support the notion of a single factor with loadings ranging from .55 to .78. Loadings from a two-factor confirmatory factor analysis of the UNIT support the notion of the memory and reasoning dichotomy (Testerman-Reed & McCallum, 1995).

#### Standardization Procedures.

The TONI-3 was standardized and normed in 1995 and 1996 on a population of 3,451 people. The normative population was representative of the United States population according to geographic region, gender, race, urban/rural residence, ethnicity, disability, socio-economic status, and education or parental education level. Ethnic groups represented in the normative data include Caucasians, African Americans, Hispanics, Asians, and Native Americans. For the purposes of developing

normative tables, the data were broken down into age intervals ranging from 6 months to 10 years (Brown et al., 1997).

The Leiter-R was standardized on a population of 1,719 people. It was stratified across geographic region, gender, race, ethnicity, and parent education. The Leiter-R did not stratify the normative data for socio-economic status or for disability. In addition, only 763 people were included in the standardization procedures of the Attention and Memory (AM) battery. Such standardization concerns put the Leiter-R at a distinct disadvantage compared to that of the TONI-3 and the UNIT. On a more positive note, the normative data of the Leiter-R was broken down into age intervals of 2 to 6 months, making it a highly sensitive instrument (Roid & Miller, 1997).

The UNIT was standardized on a sample of 2,100 people. It was stratified to match the characteristics of the United States population according to geographic region, gender, race, ethnicity, parent education, disability/special education services, and principle language. Various ethnic groups were represented in the standardization sample including Caucasian, African American, Asian/Pacific Islanders, and Native Americans. Unfortunately, the Hmong population was not represented in this process. In addition, the UNIT has a unique advantage in that ESL children composed 2% of the normative population. Since 175 people composed each age group of the normative sample, the UNIT was developed to be a highly sensitive measure of intellectual functioning with normative data being broken down in four month age intervals (McCallum & Bracken, 1998).

### Psychometric Properties.

Reliability measures of internal consistency for the TONI-3 were evaluated using coefficient alpha and alternate form reliability. The coefficient alpha level, which provides information about how well individual items correlate with others, was identified as .93. The alternate form reliability test, which evaluated the equivalence of Form A and Form B, produced a correlation of .84. In addition to tests of internal consistency, time sampling tests identified test-retest scores of .91 and .92, respectively, for Form A and Form B, supporting the notion that the TONI-3 is not hindered by time sampling errors. Lastly, inter-rater reliability coefficients of .99 were obtained on the TONI-3, suggesting that the scale provides little possibility for examiner scoring error (Brown et. al., 1997). These data support the notion that the TONI-3 is a reliable measure of intellectual functioning.

Several statistical analyses were used to evaluate the validity of the TONI-3. Criterion validity was established by evaluating the correlations between the TONI-3 and the Comprehensive Test of Nonverbal Intelligence (Form A = .76, Form B = .74), the Wechsler Intelligence Scale for Children-III (Form A = .63, Form B = .63), and the Wechsler Adult Intelligence Scale-Revised (Form A = .73, Form B = .71). Construct validity was established by showing that raw scores were correlated with age progression and that TONI-3 standard scores correlate with school achievement as identified by the Woodcock Johnson Test of Achievement-Revised (Brown et al., 1997).

Reliability for the Leiter-R composite score is supported through coefficient alpha correlations ranging from .91 to .93. Coefficient alpha correlations between .75 and .91 support the reliability of the factors, or subscales, composing the Leiter-R. A standard error of measurement score of 4.24 was identified for the comprehensive battery. Overall, the data presented supports the notion that the Leiter-R is a reliable measure of intelligence (Roid & Miller, 1997).

Validity for the Leiter-R was primarily established through criterion and construct validity. Criterion validity was established by comparing scores on the Leiter-R to that of the original Leiter International Performance Scale ( $r = .85$ ) and the Wechsler Intelligence Scale for Children-III ( $r = .87$ ). Construct validity was established by comparing Leiter-R scores to academic achievement as measured by the Wechsler Individual Achievement Test Reading ( $r = .67$ ) and Math ( $r = .78$ ) scores and the Woodcock Johnson Test of Educational Achievement-Revised Broad Reading ( $r = .82$ ) and Broad Math ( $r = .82$ ) scores. Overall, the validity studies in the Leiter-R provide ample support (Roid & Miller, 1997).

The UNIT's reliability was established through several means. An internal consistency reliability coefficient of .88 was identified for the Extended Battery. Individual subtest coefficients ranged from .64 to .91. In addition, a test-retest coefficient of .81 was identified for the Extended Battery. Lastly, the standard error of measurement for the Extended Battery was identified as 4.10 (McCallum & Bracken, 1998). Overall, the UNIT appears to be a highly reliable measure of intelligence.

Validity support for the UNIT was established through criterion and construct-referenced correlational data. Criterion-referenced validity was established by comparing the UNIT to the Wechsler Intelligence Scale for Children-III for learning disabled ( $r = .71$ ), cognitively disabled ( $r = .72$ ), and intellectually gifted ( $r = .54$ ) individuals. Construct validity was established by correlating UNIT scores with the Woodcock Johnson Test of Educational Achievement-Revised ( $r = .72$ ) and the Wechsler Individual Achievement Test ( $r = .62$ ) (McCallum & Bracken, 1998).

#### Summary of Nonverbal Scales of Intelligence.

The TONI-3, the Leiter-R, and the UNIT are all recently-developed nonverbal instruments of intelligence. The theoretical foundation of each instrument is relatively solid. In addition, factors of ability identified in the theoretical structure of the Leiter-R and the UNIT are supported through factor analysis. The developmental procedures used in the collection of normative data were outstanding for both the UNIT and the TONI-3 but somewhat weak for the Leiter-R. This weakness is primarily due to a small standardization sample for the Attention and Memory battery and the lack of accountability for disabilities and socio-economic status in the stratifying procedures. The psychometric properties of the TONI-3, the Leiter-R, and the UNIT are adequate to support their utility as measures of intellectual functioning. Since all three of these measures are relatively new, little research has been published on their psychometric

properties apart from that reported in the manuals. Therefore, future research will be required to assess the characteristics of each of these instruments.

Overall, all three of these instruments appear to be solid measures of intellectual functioning. However, the UNIT may have some distinct advantages over the other instruments. First, unlike the Leiter-R and the UNIT, the TONI-3 is primarily a screening device and cannot assess intellectual functioning in a comprehensive manner. In addition, the TONI-3 does not provide a profile of individual strengths and weaknesses in functioning. Therefore, the Leiter-R and the UNIT have a distinct advantage over the TONI-3. Second, the standardization sample of the Leiter-R is smaller and was not stratified according to as many variables as the standardization sample of the UNIT. Third, the UNIT is a shorter instrument, consisting of only 6 subtests versus the 10 to 17 used in the Leiter-R, and consequently would require far less administration time. Using tests that are less time consuming enables the time-crunched examiner to free up time for additional methods of assessment for the child. In addition, shorter tests prevent examinee fatigue and loss of motivation, resulting in more valid test scores. Because of these distinct advantages, the UNIT may be a favorable nonverbal instrument of intellectual functioning.

#### IQ Tests for the Hmong Population

To date, only a single published study evaluates the efficacy of a psychometric instrument for the Hmong population. As part of a comprehensive assessment plan for

Hmong children, Irwin and Madden (1986) developed normative data for 1) the Raven Coloured Progressive Matrices and 2) the Mazes, Block Design, and Coding subtests of the WISC-R. Irwin and Madden hypothesized that testing the Hmong population in their native language would enable them to score at equivalent levels to that of the American and British norms. Consequently, both of these instruments were translated into the Hmong language. The sample in the norm development procedure consisted of 110 nonreferred Hmong students from a small mid-western city enrolled in a summer school program ranging in age from 6 years 0 months to 16 years 11 months.

The Raven Coloured Progressive Matrices data obtained from the Hmong students is best interpreted when split into two separate age groups. Children from 6 to 11 years of age ( $n = 55$ ) achieved scores that did not differ significantly from that of the British normative data. However, children between 11 and 16 years of age ( $n = 48$ ) achieved scores that were significantly lower than the British normative data. In fact, the mean score obtained from the Hmong sample was equivalent to that of the 35 percentile of the British norms.

The Hmong students in the study performed remarkably better on the WISC-R subtests than on the Raven Coloured Progressive Matrices. Mean standard scores for Mazes, Block Design, and Coding were 11.54, 10.05, and 10.63 respectively. Interestingly, each of these scaled scores is equal to or greater than that of the overall population mean score. Overall, Irwin and Madden contend that the WISC-R's Mazes, Block Design, and Coding subtests provide a valid means of assessment based

on the minimal discrepancy in scores between the Hmong norms and that of the American norms.

Unfortunately, using the Mazes, Block Design, and Coding subtests of the Wechsler scales does not provide a comprehensive assessment of intellectual functioning. It is important to recognize that these three subtests only measure a portion of intellectual functioning. Consequently, such procedures have little applicability in the assessment process of Hmong children. No published research to date supports the efficacy of a test of comprehensive intellectual functioning for the Hmong population. Further research needs to be conducted to evaluate the efficacy of intelligence tests for the Hmong population.

As previously noted, nonverbal scales of intelligence appear to reduce language-based biases typically experienced by ESL and LEP populations similar to the Hmong. A likely next step in the research would be to evaluate the fairness of a nonverbal measure of intelligence, such as the UNIT, for the Hmong population. Such research would be invaluable to school psychologists who either occasionally or routinely are involved in assessment of Hmong children.

### Summary

Research on the validity and reliability of intelligence tests for ESL and LEP populations has confirmed that traditional, language-loaded tests may yield scores that do not accurately reflect intellectual abilities (Cummins, 1984). This can be attributed,

in part, to language-based biases. Fairness data for nonverbal scales of intelligence suggests that minimal differences exist between the scores of ESL populations and the overall population (Brown et al., 1997; McCallum & Bracken, 1998; Roid & Miller, 1997). Consequently, nonverbal scales of intelligence may be useful in eliminating language-based biases.

Inherent flaws in traditional, nonverbal intelligence tests, including outdated norms, weak theoretical foundations, and unidimensional focuses (McCallum & Bracken, 1997), have been recently addressed. Contemporary nonverbal scales of intelligence including the UNIT, the TONI-3, and the Leiter-R have excellent theoretical foundations, standardization procedures, and psychometric properties. Of these instruments, the UNIT may have distinct advantages in that it is comprehensive in nature, has an outstanding normative population, and requires minimal administration time.

Little empirical evidence has been established to support the utility of an intelligence test for the Hmong population. Currently, there are no published studies that evaluate the efficacy of an entire, comprehensive battery of intelligence for Hmong people. Future research needs to fill this "gap" in order to assist clinicians and school psychologists in meeting the needs of their Hmong clients.

## CHAPTER III

### Discussion

The following chapter will review and synthesize the information presented in the first two chapters of this paper. Such information is organized into three broad categories, including current practices in the intellectual assessment of the limited English Proficient (LEP) population, a comparison of the Universal Nonverbal Intelligence Test (UNIT) to other modern nonverbal scales of intelligence, and the utility of the UNIT for Hmong children. Next, the relevance of such information for school psychologists will be discussed. Finally, this chapter will detail weaknesses in the literature and provide suggestions for future research.

#### Intellectual Assessment for the LEP Population

Legislation, litigation, and professional standards have routinely identified the inappropriateness of conventional, verbally-loaded intelligence tests for English as a second language (ESL) and LEP populations and suggested that such measures must be administered in the child's native or primary language (Esquivel, 1985; Figueroa, 1990; Figueroa et al., 1984; Lopez, 1997). Consequently, school psychologists throughout the nation have complied with such "recommendations" by either using translators during the process of assessment or by using translated, re-normed versions

of popular tests. It is important to recognize that such procedures have not been supported by research as valid methods of assessment. In addition, hypothesized methodological flaws within the interpretation and translation process may have devastating impacts on assessment outcomes. Using interpreters during the assessment process creates a possibility for indirect translations and inappropriate reinforcement (Lopez, 1995). Examiners should also be cautious when using translated, re-normed versions of popular tests. Such measures are plagued with problems, including small and inappropriate standardization samples (Figueroa, 1990; Olmedo, 1981) and unknown predictive validities (Figueroa, 1990).

Attempting to assess children in their primary language may have an additional inherent flaw. Such practices assume that the child is proficient in his or her primary language. However, it is important to recognize that primary language proficiency may actually decline during the process of learning the English language, termed subtractive bilingualism (Cummins, 1984; Dao, 1991). Consequently, it is quite possible for the child to be lacking proficiency in both languages. In such cases, the child would not have the adequate language skills to be validly tested with a verbally-loaded intelligence test, regardless of whether it was administered in English or the child's primary language.

Hmong individuals are more likely to be LEP than many other linguistic groups, including Spanish-speaking populations. In addition, Hmong children are more likely than Spanish-speaking children to maintain LEP status through their junior

year of high school (Ima & Rumbaut, 1995). Low levels of parental education and literacy (Morrow, 1989; Westermeyer, Bouafuely-Kersey, & Her, 1997) as well as several incongruencies between the Hmong and English languages (Lewis, Vang, & Cheng, 1989) may have a tremendous impact on the acquisition of the English language. Consequently, concerns relating to identifying appropriate intellectual assessment instruments for LEP populations have particular relevance for the Hmong population.

#### The Universal Nonverbal Intelligence Test

A primary purpose of this paper was to determine how the Universal Nonverbal Intelligence Test (UNIT) compares to other modern nonverbal scales of intelligence. As previously noted, several of the concerns addressed in the research about historical nonverbal tests, including outdated norms, weak theoretical foundations, and unidimensional focuses (McCallum & Bracken, 1997) have been addressed in contemporary nonverbal tests including the UNIT, the Leiter International Performance Scale-Revised (Leiter-R), and the Test of Nonverbal Intelligence-3 (TONI-3). All three of these measures have adequate theoretical foundations, standardization procedures, and psychometric properties. The UNIT may have a distinct advantage over these two main competitors in the nonverbal testing market in that it has superior standardization procedures, is comprehensive in nature, and requires minimal time to administer.

### The Utility of the UNIT for the Hmong Population

Another primary purpose of this paper was to determine if the UNIT is a valid and fair measure of intellectual functioning for the Hmong population. Although not specifically tested with empirical research, several conclusions can be drawn by inferentially combining information about the characteristics of the Hmong population with the properties of the UNIT.

To date, there is no empirical support validating the use of a comprehensive intelligence test for the Hmong population. Evidently, research must be conducted to determine the efficacy of such measures. In the meantime, one must evaluate the characteristics of the Hmong population to determine which instruments are most appropriate. As detailed in chapter 1, a good measure of intellectual functioning for the Hmong population should 1) have limited verbal administration or content, 2) offer plenty of demonstration and training items, 3) have minimal emphasis on time and speeded responses, 4) be relatively culture-free, and 5) de-emphasize the use of paper and pencil-based tasks.

Recently developed nonverbal measures of intellectual functioning appear to address these issues in an adequate manner. The UNIT was designed to be entirely nonverbal, eliminating the use of language in both the content and the administration of the test. In addition, the standardization population of the UNIT included children who were classified as ESL (McCallum & Bracken, 1998). Consequently, the UNIT is an ideal instrument for populations with language deficiencies. In fact, research

supports the effectiveness of the UNIT for LEP populations (McCallum & Bracken, 1998). Therefore, it can be assumed that the UNIT will reduce or eliminate any language-based biases that the Hmong children may experience during testing. The UNIT also provides training and practice items for each type of task, ensuring that the child has a thorough understanding of the task before engaging in scored items. Consequently, Hmong children, who may be reluctant to ask for help during times of uncertainty (Williams, 1987), will have plenty of opportunity to develop an understanding of the required tasks. Similar to many current nonverbal scales of intelligence, the UNIT was designed to be relatively culture-free. The stimuli and materials used throughout the entire test do not appear to be culturally-specific. Consequently, it is assumed that Hmong children who are unfamiliar with American culture will not be at a disadvantage. Finally, the UNIT does not rely heavily on pencil and paper-based tasks.

In summary, it appears as if the UNIT can function as an appropriate and valid measure of intellectual functioning for the Hmong population. Hmong children have several unique needs in intellectual assessment including limited verbal administration or content, an abundance of training items, and culture-free content. The UNIT was designed in a manner that accounts for these considerations.

### Implications for School Psychologists

The information presented in this paper has particular relevance to school psychologists who serve Hmong children. It is important to recognize the limitations of traditional, verbally-loaded scales of intelligence when working with LEP populations. Such measures may not accurately reflect the actual abilities of the child (Cummins, 1984). It is important to assess the language proficiency of Hmong children (both in English and in Hmong) to assist in the selection of an appropriate intelligence test. If the child does not have proficiency in the English language, nonverbal IQ tests such as the UNIT should be used in the assessment of intellectual functioning. Incorporating the UNIT in the assessment process of LEP Hmong children should provide valid assessment results, leading to appropriate diagnostic and placement decisions. The ultimate goal is to identify the unique needs of the child. The UNIT is an instrument that will assist school psychologists in achieving this goal.

It appears, based on preliminary evidence, that the UNIT can function as an appropriate instrument in the cognitive assessment of Hmong children. However, future research needs to evaluate the validity and reliability of the UNIT for the Hmong population. Such empirical support will provide practitioners with greater insight into the effectiveness of nonverbal measures of intelligence in the assessment process of Hmong children.

### Gaps in the Literature and Recommendations for Future Research

Very little research has evaluated the efficacy and validity of intelligence tests for LEP populations. The effectiveness of using interpreters in the process of assessment is yet to be extensively evaluated. In addition, the psychometric properties of translated and renormed intelligence tests, including predictive validities, remain unknown. A lack of such research is detrimental if school psychologists continue to use these methods in the assessment of ESL and LEP children.

Very little is known about the performance of Hmong children on intelligence tests. In fact, no study to date has evaluated the effectiveness of a comprehensive instrument of intellectual functioning for the Hmong population. It is important to determine how Hmong children perform on these tests in comparison to the total population. Such information will provide insight into potential biases that may exist within the content, structure, or administration of the test.

The following recommendations are intended to guide researchers who are interested in the intellectual assessment of the Hmong and other populations characterized by high levels of LEP:

- 1) Evaluate the validity and reliability of nonverbal scales of intelligence, such as the Universal Nonverbal Intelligence Test, for Hmong and LEP populations.
- 2) Compare the efficacy and validity of nonverbal scales of intelligence, such as the UNIT, for Hmong children receiving ESL services and Hmong children who are not receiving ESL services. Assuming that these two groups have similar abilities, both

groups should perform at approximately equivalent levels. This would support the notion of the absence of language biases in such measures.

3) Develop normative data for the Hmong population with nonverbal measures of intelligence, such as the UNIT, to assist in valid assessment and placement decisions.

This may be particularly important for geographic regions in California, New York, Minnesota, Illinois, and Wisconsin where there are larger Hmong populations. Such data can be used to compare the performance of a particular Hmong child to that of his/her peers within the city or local region.

4) Further evaluate the psychometric properties of the UNIT to provide support for the validity and reliability of the measure for various populations.

5) The authors of the UNIT contend that the Symbolic component of their measure provides insight into an individual's verbal abilities. Empirical support is required to justify this claim.

6) Evaluate the effectiveness of translators and translated tests in the process of intellectual assessment. Empirical data is necessary to justify the continued use of such procedures in the assessment of ESL children.

## REFERENCES

- Brown, L. Sherbenou, R., & Dollar, S. (1982). Test of Nonverbal Intelligence: Examiner's manual. Austin, TX: Pro-Ed.
- Brown, L., Sherbenou, R., & Johnsen, S. (1997). Test of Nonverbal Intelligence (3rd ed.):Examiner's manual. Austin, TX: Pro-Ed.
- Carroll, J. (1993). Human cognitive abilities: A survey of factor-analytic studies. New York: Cambridge University Press.
- Clark, P. (1985). [Review of the Test of Nonverbal Intelligence]. In J. Mitchell (Ed.), The ninth mental measurements yearbook (pp. 1580-1581). Lincoln, NB: Buros Institute of Mental Measurements.
- Cummins, J. (1984). Bilingual special education: Issues in assessment and pedagogy. San Diego, CA: College Hill.
- Dao, M. (1991). Designing assessment procedures for educationally at-risk southeast Asian-American students. Journal of Learning Disabilities, 24, 594-601.
- Esquivel, G. (1985). Best practices in the assessment of limited English proficient and bilingual children. In A. Thomas, & J. Grimes (Eds.), Best practices in school psychology (pp. 113-123). Kent, OH: National Association of School Psychologists.
- Figueroa, R. (1990). Assessment of linguist minority group children. In C. Reynolds, & R. Kamphaus (Eds.). Handbook of psychological and educational assessment of children: Intelligence and achievement (pp.671-696). New York: Guilford Press.
- Figueroa, R., Sandoval, J., & Merino, B. (1984). School psychology and limited-English-proficiency children: New competencies. Journal of School Psychology, 22, 131-143.
- Ima, K., & Rumbaut, R. (1995). Southeast Asian refugees in American schools: A comparison of fluent English-proficient and limited-English proficient students. In D. Nakanishi, & T. Yamano (Eds.), The Asian American educational

- experience: A source book for teachers and students (pp. 180-197). New York: Routledge.
- Irwin, D., & Madden, C. (1986). A psychoeducational assessment procedure for Southeast Asian refugee students. In C. Williams, J. Westermeyer et. al. (Eds.), Refugee mental health in resettlement countries: The series in clinical and community psychology (pp. 189-204). Washington, DC: Hemisphere Publishing.
- Jensen, A. (1980). Bias in mental testing. New York: Free Press.
- Kowall, M., Watson, G., & Madak, K. (1990). Concurrent validity of the Test of Nonverbal Intelligence with referred suburban and Canadian native children. Journal of Clinical Psychology, 46, 632-636.
- Leiter, R. (1979). Instruction manual for the Leiter International Performance Scale. Wood Dale, IL: Stoelting.
- Lewis, C., & Lorentz, S. (1994). Comparison of the Leiter International Performance Scale and the Wechsler Intelligence Scales. Psychological Reports, 74, 521-522.
- Lewis, J., Vang, L., & Cheng, L. (1989). Identifying the language-learning difficulties of Hmong students: Implications of culture and context. Topics in Language Disorders, 9, 21-37.
- Liu, T., & Li, C. (1994). How can school psychologists best meet the needs of Hmong-Americans? Article submitted for publication.
- Lopez, E. (1995). Best practices in working with bilingual children. In A. Thomas, & J. Grimes (Eds.), Best practices in school psychology III (pp. 1111-1121). Washington, DC: National Association of School Psychologists.
- Lopez, E. (1997). The cognitive assessment of limited English proficient and bilingual children. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), Contemporary intellectual assessment (pp. 503-516). New York: Guilford Press.
- Martin, J., Blair, G., & Bledsoe, J. (1990). Measures of concurrent validity and alternate-form reliability of the Test of Nonverbal Intelligence. Psychological Reports, 66, 503-508.

- McCallum, S., & Bracken, B. (1997). The Universal Nonverbal Intelligence Test. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), Contemporary intellectual assessment (pp. 268-280). New York: Guilford Press.
- McCallum, S., & Bracken, B. (1998). The Universal Nonverbal Intelligence Test: Examiner's manual. Itasca, IL: Riverside Publishing.
- McGhee, R., & Lieberman, L. (1990). Test-retest reliability of the Test of Nonverbal Intelligence (TONI). Journal of School Psychology, 28, 351-353.
- Morrow, R. (1989). Southeast-Asian parental involvement: Can it be a reality? Elementary School Guidance & Counseling, 23, 289-297.
- National Association of School Psychologists. (1995). Standards for the provision of school psychological services. In A. Thomas, & J. Grimes (Eds.), Best practices in school psychology III (pp. 1161-1172). Washington, DC: Author.
- Nuttall, E. (1987). Survey of current practices in the psychological assessment of limited-English proficiency handicapped children. Journal of School Psychology, 25, 53-61.
- Olmedo, E. (1981). Testing linguistic minorities. American Psychologist, 36, 1078-1085.
- Reynolds, C., & Kaiser, S. (1990). Bias in assessment and aptitude. In C. Reynolds, & R. Kamphaus (Eds.), Handbook of psychological and educational assessment of children: Intelligence and achievement (pp. 611-649). New York: Guilford Press.
- Roid, G., & Miller, L. (1997). Leiter International Performance Scale-Revised: Examiner's manual. Wood Dale, IL: Stoelting.
- Sattler, J. (1988). Assessment of children: Revised and updated (3rd ed.). San Diego, CA: Author.
- Schiele, J. (1991). An epistemological perspective on intelligence assessment among African American children. The Journal of Black Psychology, 17, 23-36.
- Testerman-Reed, M., & McCallum, R. (1995). Construct validity of the Universal Nonverbal Intelligence Test. Psychology in the Schools, 32, 277-290.

- Thorndike, R. (1997). The early history of intelligence testing. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), Contemporary intellectual assessment (pp. 3-16). New York: Guilford Press.
- Wang, L. (1995). Lau v. Nichols: History of a struggle for equal and quality education. In D. Nakanishi, & T. Yamano (Eds.), The Asian educational experience: A source book for teachers and students (pp. 58-85). New York: Routledge.
- Westermeyer, J., Bouafuely-Kersey, M., & Her, C. (1997). Hmong children. In G. Johnson-Powell, J. Yamamoto, et. al. (Eds.), Transcultural child development: Psychological assessment and treatment (pp. 162-181). New York: Wiley.
- Williams, C. (1988). Issues surrounding psychological testing of minority patients. Hospital & Community Psychiatry, 38, 184-189.