5-2005

Cognitive Assessment of School Age Spanish Speaking English Language Learners

Casey Johnson
Utah State University

COGNITIVE ASSESSMENT OF SCHOOL AGE SPANISH SPEAKING ENGLISH LANGUAGE LEARNERS

by

Casey Johnson

A plan-B paper submitted in partial fulfillment
of the requirements for the degree

of

MASTER OF SCIENCE

in

School Psychology

Approved:

UTAH STATE UNIVERSITY
Logan, Utah

2005

CONTENTS

SECTION 1

INTRODUCTION

Background and Problem Statement

The number of students who speak a language other that English in schools across the U.S. is rapidly increasing. The Spanish speaking student population, in particular, has grown considerably in recent years. One way to examine the extent of this growth is to consider demographic data at the national and state levels.

According to data provided by the U.S. Bureau of the Census (2000), there are approximately 281 million Americans. Of this 281 million, approximately 75 percent are categorized as white. At first glance, it would appear that the U.S. population is predominantly white and English speaking. Further examination, however, of the actual numbers of ethnic and racial minorities shows a different picture of the U.S. population. It is important to note that in the U.S., the Hispanic population cuts across all racial groups, including white, African American, Asian or Pacific Islander, and Native American. The term "Hispanic" is used to denote people of various ethnic, racial, national, and cultural backgrounds whose ancestors lived in Spain or Latin America. There is great cultural, ethnic, and linguistic diversity among the Hispanic population in the U.S. When one takes this into consideration, the nation's diversity becomes more apparent.

The Hispanic or Latino population in the U.S. is a group that is growing at a much more rapid rate than other ethnic populations. It was shown to be the fastest growing population in the 1990s (U.S. Bureau of the Census, 2000). Data from the 2000 census indicated that while the white population grew by 6 percent during the

1990s, the Hispanic population grew by 58 percent. Hispanics grew in number from just over 22 million in 1990 to just over 35 million in 2000. More recent data showed that, as of July 1, 2004, the nation's Hispanic population reached 41.3 million (U.S. Bureau of the Census, 2004). Estimates are that by the middle of the twenty-first century 25 percent of the population in the U.S. will be Hispanic.

In addition to being ethnically diverse, the population of the U.S. is becoming more linguistically diverse as well. According to the U.S. Bureau of the Census (2000), 46.9 million, or 18 percent, of Americans speak a language other than English at home, an increase from 31.8 million, or 14 percent, a decade ago. Analysis of data from the past twenty years shows that the number of people in the U.S. who speak a language other than English in the home has doubled and continues to rise. Spanish is by far the most widely spoken non-English language. The number of those who listed Spanish as the primary language spoken in the home rose from 17.3 million in 1990 to 28.1 million in 2000. The U.S. Bureau of the Census reports that although many of these households also speak English, of those 28 million, 7.9 million reported speaking English "not well" or "not at all."

The dramatic growth of the Spanish-speaking population in the U.S. is largely due to a significant increase in the number of immigrants from Latin America in the past few decades. Data from the 2000 census showed the nation's immigrant population to grow by 11.3 million in the 1990s, faster than any other time in the history of the U.S. (Camarota & McArdle, 2003). In that time, immigrants from Spanish-speaking Latin America were shown to account for more than 60 percent of the growth in the foreign-born population nationally.

Several terms have been used to describe and categorize English learners in the schools. The term English language learner (ELL), as used in this paper, indicates a

person who is in the process of acquiring English and has a first language other than English (U.S. Department of Education, 1994). Another term often used is limited English proficient (LEP). LEP refers to individuals who were either not born in the United States and whose native language is other than English, or who come from environments in which a language other than English is dominant. Although the term LEP has frequently been used by educators and researchers in the past, there has been a gradual shift towards using the term ELL to aid in removing negative connotations regarding a student's abilities. The terms ELL and LEP are used synonymously in this paper, often depending on the term used by authors of a particular study or article.

The nation's ethnic and linguistic diversity is reflected in our school systems where educators work with growing numbers of children who come from monolingual or bilingual backgrounds and who are learning English as a second language. In the U.S. the LEP student population accounted for 9.3 percent of the school-age population (pre-kindergerten to 12th grade) in the 1999-2000 school year (Kindler, 2002). In states such as California and New Mexico, the LEP population accounts for as much as a quarter of the total enrollment. School district data regarding LEP populations mirrors census data in demonstrating dramatic growth over the past twenty years. Kindler reports that the LEP population more than doubled in 23 states during the 1990s.

According to data provided by school district LEP programs (Kindler, 2002), Spanish was found to be the native language of more than three quarters of LEP students (76.9 percent). No other language group exceeded three percent of the LEP population. In districts where Spanish was the most common language for LEP students the median percentage of students whose native language was Spanish was 90.9 percent.

Student diversity means not only that those working in education must accommodate those from different cultural backgrounds and nations of origin but also

that with increasing frequency they will find themselves working with students who either speak English as a second language or not at all. Working with English language learners (ELLs) can be a particular challenge for those who complete psychoeducational evaluations.

First of all, many psychologists enter the workplace having received insufficient training and experience in the area of bilingual assessment (Scribner, 2002). Results of a survey of school psychologists who conduct bilingual psychoeducational assessments indicated that the majority of respondents believed that they had received inadequate training (Ochoa, Rivera, & Ford, 1997). This included knowledge and training in the areas of second language acquisition factors, methods to conduct bilingual psychoeducational assessment, and the ability to interpret the results of bilingual psychoeducational assessments. In a separate study, a survey of directors of school psychology programs showed that 40 percent of the programs did not offer courses on minority issues (Rogers, Ponterotto, Conoley, & Weise, 1992). A more recent study presented somewhat more encouraging results. Loe (2001) examined school psychologists' professional training in the areas of family oriented services and cultural diversity. Ninety-four percent of school psychologists surveyed reported receiving some training related to cultural diversity. A sizeable portion of respondents, however, reported feeling dissatisfied with their competence (23 percent) and training (34 percent) in the provision of services to ethnically diverse students.

Other assessment challenges arise from the fact that many existing personnel in the field often lack dual language proficiency. Given the significant numbers of students who speak a language other than English, in addition to the variety of languages spoken, this is not surprising. Problems arise, however, when a psychologist's lack of proficiency in the student's primary language leads to the use of

assessment practices that do not coincide with legal and ethical guidelines. In addition, specific assessment practices may be of questionable validity. Historically, these practices have included testing in English only, using interpreters, using only nonverbal measures, and administering measures with unestablished validity and reliability with ELL populations (Lopez, 1995).

Several researchers have noted that inadequate or invalid psychoeducational assessment practices have contributed to inappropriate labeling and misplacement of many ethnic and language minority students in special education classes (Chinn & Hughes, 1987; Macias, 1998; Shinn, Collins, & Gallagher, 1998). The National Research Counsel reported that nationally, Hispanics had a 7 percent greater probability of being labeled learning disabled when compared to white students (2002). The *Executive Summary - Conference on Minority Issues in Special Education*, written by the Civil Rights Project (2000), states the following:

> Historically, special education has too often been a place - a place to segregate minorities and students with disabilities....To the extent that minority students are misclassified, segregated, or inadequately served, special education can contribute to a denial of equality of opportunity, with devastating results in communities throughout the nation. (p.1)

Clearly the stakes are high with regards to identifying and placing Hispanic or ELL students in special education programs. It is imperative that valid assessment techniques are developed and utilized so that educational decisions provide ELL students with equal access to appropriate educational opportunities.

Purpose of this Review

This paper will serve to address those challenges described above by providing

professionals with a resource for conducting cognitive assessments of Spanish speaking children in an empirically sound, nonbiased, defensible, and practical manner. The discussion will begin with a review of the ethical and legal guidelines relevant to conducting assessments of ELL students. Previous court cases that have relevance to current practice will be highlighted. Ethical guidelines from groups such as the American Psychological Association and National Association of School Psychologists will be presented. This will be followed by a review of various assessment methods and important considerations pertaining to the assessment of Spanish speaking ELL students. This section will include discussion on topics such as critical components of the assessment, language proficiency assessment, acculturation, competency of the examiner, and the use of interpreters. Next, a review of specific cognitive measures will be conducted. Comprehensive intelligence tests, nonverbal measures, and a measure of bilingual verbal ability will be examined. Measures were selected for review based on several criteria. One criterion was the widespread use of the measures by school psychologists and other professionals. Ochoa, Powell, and Robles-Pina (1996) offer data regarding several instruments often used by school psychologists to assess intellectual functioning with bilingual students. The measures most often used by school psychologists were considered for this paper. Other measures included were those found to be frequently and consistently mentioned and discussed by leading authors in the field of bilingual assessment (Athanasiou, 2000; Figueroa, 1990b; Gopaul-McNicol & Armour-Thomas, 2002; Lopez, 1997; Ortiz, 2002; Rogers, 1998; Willen & Sweeting, 1986). Cognitive measures that were normed within the last 15 years was another criterion for inclusion of tests. When examining assessment measures, studies that are empirical in nature and include research conducted on Spanish speaking or Hispanic youth will be included. Studies that examine outcomes,

test bias, reliability, and validity will be reviewed. Alternatives to traditional standardized measures will also be discussed. Finally, the paper will end with a conclusion that summarizes best practices in the area of cognitive assessment with Spanish speaking ELL children. Upon reviewing this information, it is hoped that professionals will be better prepared to meet the needs of an increasingly diverse student population by becoming better informed regarding specific assessment methods as well as the advantages and disadvantages of specific measures.

SECTION 2

ETHICAL AND LEGAL ISSUES

Various policies, laws, and judicial decisions have been designed to ensure that ELL students with and without disabilities receive an appropriate education. For the past several decades, psychoeducational assessment practices have been largely guided by federal, state, and local legislation, and by litigation outcomes. The courts and congress have become increasingly more involved in decisions that affect the direction of educational and psychological services in schools. The impact of these legal actions on children from varied cultural and linguistic backgrounds, in particular, has been significant. Because of the impact these cases and legislative acts have had on current assessment practices that pertain to ELL children, it is imperative that those who conduct assessments of ELL students understand their implications on current practice. What follows is a review of the pertinent court decisions and legislative acts that have had important consequences for the way ELL children are evaluated in U.S. schools today.

Legal Considerations

Brown v. Board of Education (1954)

In this landmark supreme court case, the court ruled that segregating students based on their ethnicity or race conflicted with the 14th Amendment of the U.S. Constitution. The 14th Amendment stipulates that no state shall "deny any person within its jurisdiction the equal protection of the laws." The court found that schools were arbitrarily discriminating against African American students by educating them separately from other students. This ruling set a precedent for future litigation and

legislation that limited discriminatory practices against students considered different due to race, culture, language, or disability.

Diana v. State Board of Education (1970)

The Diana decree may be the most influential court case decision concerning assessment practices and ELL children. The Diana case was named for one of nine plaintiffs in a class-action suit. The case addressed alleged disproportionate representation of bilingual, Mexican-American children who had been placed in programs for the mentally retarded in California. Diana, a Spanish speaking student, was assessed and placed in a program for mentally retarded students after test results showed an IQ score of 30. She was later reassessed using the same instrument by a bilingual school psychologist in both English and Spanish. The resulting IQ score was almost 50 points higher, indicating she was not disabled and no longer qualified for special education services. In this case, California was mandated by the court to correct bias in assessment procedures used with Mexican American students. This consent decree set broad guidelines for the assessment of linguistically different children. Namely, that these students be evaluated in their native language or with sections of tests that do not require knowledge of the English language.

Guadalupe Organization v. Tempe Elementary School District No. 3 (1972)

This case was heard by the Ninth Circuit Court of Appeals after an Arizona district court rejected the suit brought against the Tempe district. In this case, the plaintiff requested that the school district provide all non-English speaking Mexican American (Hispanic) and Yaqui Indian students with bilingual and bicultural education. Results were similar to those in the *Diana* case and indicated that students should be assessed in their primary language or through the use of nonverbal measures if the

student speaks a language other than English. The case further established that IQ tests could not be the sole criteria or primary basis for the diagnosis of mental retardation and that adaptive behavior must also be considered.

Larry P. v. Riles (1972)

This landmark case was a class action suit filed in California on behalf of African-American students who had been disproportionately placed in classes for students with mental retardation based on the results of standardized IQ tests. The judge ordered an injunction against the use of IQ tests that failed to take into consideration the cultural backgrounds and experiences of African American children. The state was ordered to reevaluate students in programs for the mentally retarded and to monitor racial and ethnic disparities in special education. Much like the *Diana* case, it provided a legal precedent against culturally biased assessment practices in the schools.

Lau v. Nichols (1974)

In this case the Supreme Court ruled that the San Francisco Unified School District violated Title VI of the Civil Rights Act by failing to provide services to help Chinese-speaking students learn English. Findings indicated that merely providing equal materials and resources did not represent equality of treatment if the students do not understand English. The court decision helped to foster programs which focus on the identification of linguistically diverse students, assessment of their language proficiency, and their placement in appropriate programs with bilingual instructional strategies (Gopaul-McNicol & Armour-Thomas, 2002).

Many of these landmark cases have generated court decisions that have translated into a series of federal dictates. This includes legislation such as the Civil Rights Act and the Individuals with Disabilities Education Act (IDEA).

Civil Rights Act of 1964

The Civil Rights Act (1964), along with the judicial interpretations that followed, prohibits programs that are federally funded from discriminating in their services on the basis of race, color, religion, or national origin. The act stipulates that programs cannot offer services that are different from, or less effective than those offered to other individuals unless it can be shown that to do so ensures that services are effective. In 1970, the US. Department of Health, Education, and Welfare issued a memorandum detailing that excluding children from participating in school because they cannot understand or speak English constitutes a violation of the Civil Rights Act (Artiles & Ortiz, 2002). School districts were instructed to take steps to rectify children's language deficiencies and avoid identifying students as mentally retarded based on criteria related to English proficiency.

Individuals With Disabilities Act of 1975

The Education for All Handicapped Children Act of 1975, or Public Law 94-142, was developed to ensure children with disabilities are provided access to a free appropriate public education and to improve educational results for children with disabilities. Various aspects of the law have implications for the assessment of linguistic minority children. First of all, the law mandated that nondiscriminatory assessment practices be employed when assessing students from culturally and/or linguistically diverse backgrounds. This included evaluating children in their native language or primary mode of communication unless it is clearly not possible to do so. Native language is defined as the language that the child understands best and is not necessarily the language spoken by the parents. The act also stipulated that assessment is to be done by a multidisciplinary team, using instruments that do not discriminate on

the basis of race or culture. Schools are further directed to provide information regarding the special education process to parents in their native language. This may include steps such as providing parents with an interpreter or translating IEP forms into parents' native language.

Various amendments have been made to the Education for all Handicapped Children Act. In 1986, Public Law 99-457 extended rights to all children with disabilities between the ages of 3 to 5 years. Congress again amended the act in 1990, when its name changed to the Individuals with Disabilities Education Act (IDEA). IDEA, which has since been revised in 1997 and 2004, further emphasizes the requirement that procedures used for evaluation and placement of children with disabilities not be discriminatory on a racial or cultural basis. The most recent revision of IDEA provides additional clarity by requiring that assessment materials are administered "in the form most likely to yield accurate information on what the child knows and can do academically, developmentally, and functionally" (IDEA, 2004). Schools must ensure that materials and procedures used to assess a child with limited English proficiency are selected and administered to ensure that they measure the extent to which the child has a disability rather than simply measuring the child's proficiency in the English language. Once an ELL student is identified as having a disability, the assessment team must consider the language needs of the child when developing and reviewing the individualized education program (IEP). IEPs should specify which instructional goals and objectives will be delivered in the native language of the student and which will be delivered in English, using strategies appropriate for ELL students (Artiles, & Ortiz, 2002).

Ethical Considerations

Although it is imperative that school-based practitioners have an extensive knowledge of federal law, federal regulations, and state regulations, legal requirements alone may not address the various complicated issues that tend to arise when working with students from diverse linguistic and cultural backgrounds. Thus, school psychologists should also be cognizant of the various ethical guidelines that relate to their practice. Various governing bodies and organizations have developed ethical codes and guidelines that relate to conducting assessments of ELL children. These guidelines represent ideal standards and principles that are generally intended to be aspirational in nature. Six ethical guidelines that are especially relevant to those working with ELL students are the *Ethical Principles of Psychologists and Code of Conduct* by the American Psychological Association (APA, 2002), the National Association of School Psychologists' (NASP, 2000) *Professional Conduct Manual*, the APA's (1993) *Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations*, the *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists (APA, 2003)*, the *Standards for Educational and Psychological Testing* (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999), and the *Code of Fair Testing Practices in Education* (Joint Committee on Testing Practices, 1995). Highlights from each of these standards will be presented.

Ethical Principles of Psychologists and Code of Conduct (APA, 2002)

The *Code of Conduct* of the American Psychological Association (APA) provides psychologists with general standards that help to define and regulate many

aspects of their professional practice. It states that psychologists have an ethical responsibility to consider the impact of age, race, gender, socioeconomic status (SES), language, disability, and national origin on individual functioning and psychological well-being. It also calls for professionals to strive to become culturally competent through training, supervision, or consultation with diverse groups. The *Code* clearly emphasizes psychologists' obligation to consider each individual's unique cultural and linguistic characteristics when providing psychological services.

Professional Conduct Manual (NASP, 2000)

Consistent with its mission to promote educationally and psychologically health environments for children, NASP has developed a set of ethical standards for school psychologists. In addition to standards on professional credentialing, training, and field placement, guidelines for the provision of school psychological services are also included in the *Manual.* Various sections address issues related to the provision of services to cultural and linguistic minorities. Practice Guideline 5 states that school psychologists "have the sensitivity, knowledge, and skills, to work with individuals and groups with a diverse range of strengths and needs from a variety of racial, cultural, ethnic, experiential, and linguistic backgrounds." School psychologists are encouraged to eliminate biases in themselves and in the tools they use and are instructed to enlist the assistance of other specialists when appropriate. Psychologists are also prompted to involve parents in aspects of assessment and intervention, taking into account language and cultural differences.

Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations (APA, 1993)

In addition to the general standards provided in the *Code of Conduct*, the APA

has developed more specific guidelines to assist psychologists in working with ethnic, linguistic, and culturally diverse populations. These are included in the *Guidelines for Providers of Psychological Services to Ethnic, Linguistic, and Culturally Diverse Populations.* The *Guidelines* encourage professionals to acknowledge the influence of ethnicity and culture on behavior and to take such factors into account when working with different ethnic groups. The authors also urge psychologists to consider the validity of assessment methods and measures when used with minority populations and to interpret assessment data within the context of the cultural and linguistic characteristics of the individual being assessed. Psychologists who do not possess knowledge and training about a specific minority groups are encouraged to seek consultation with knowledgeable professionals or to refer the individual to appropriate specialists.

Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists (APA, 2003)

The *Guidelines on Multicultural Education, Training, Research, Practice, and Organizational Change for Psychologists* were developed by the APA to provide psychologists with a framework for providing services to an increasingly diverse U.S. population. They provide professionals with several guidelines that address cultural awareness and knowledge of self and others. One guideline, for example, encourages psychologists to recognize that they are cultural beings and may hold attitudes and beliefs that can have an adverse affect on their perceptions and interactions with individuals who are ethnically and racially different from themselves. Rather than take a "color-blind" approach, psychologists are encouraged to use a multicultural approach that recognizes and appreciates group similarities and differences. Other guidelines emphasize the importance of diversity and multicultural instruction in psychology

training programs as well as the need for psychologists to use organizational change processes to support culturally informed policies and practices.

Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, & National Council on Measurement in Education, 1999)

The *Standards for Educational and Psychological Testing* provide perhaps the most authoritative set of ethical guidelines to be considered when conducting evaluations of ELL children. The 1999 edition of the *Standards* delegates an entire chapter to issues related to testing linguistically diverse children. It addresses issues related to the development, use, interpretation, and evaluation of tests. When it is feasible, test developers are encouraged to collect validity evidence for different linguistic subgroups as well as that of the population as a whole. Test developers are also instructed to provide the information necessary for appropriate test use and interpretation when a test is recommended for use with linguistically diverse individuals. Guidelines are provided for translating tests from one language to another, including reporting evidence of test comparability.

In addition to providing guidelines for test developers, the *Standards* also include specific recommendations for testing practices. Test users should seek to avoid bias in test selection, administration, and interpretation. Testing practices should be developed to reduce threats to reliability and validity that arise due to language differences. For example, a specially trained bilingual examiner may be able to use the test taker's primary language or bilingual speech to more effectively elicit test responses. The evaluator may also take into account language behavior that is considered socially acceptable and appropriate in the test taker's culture. Some children, for example, may demonstrate a tendency to be slow to respond that is typical

of their culture. Rather than interpret this tendency as a deficiency, these culturally learned speech patterns should be identified by the administrator and taken into consideration when interpreting test results. Generally, testing is to be done in the test taker's most proficient language, unless language proficiency in both languages is part of the assessment. The authors of the *Standards* noted that whenever students who are still in the process of learning English are tested in English, regardless of the content or intent of the test, their proficiency in English will also be tested. The *Standards* provide further instructions that when an interpreter is used in testing, he/she should have expertise in translating and should have a basic understanding of the assessment process. The *Standards* state that English language proficiency should not be determined solely with tests that require only a single linguistic skills and recommend that a wider range of skills be assessed. This last standard relates to cognitive assessment because the establishment of language proficiency is often the first step in determining the language to be used to administer cognitive measures.

Code of Fair Testing Practices in Education (Joint Committee on Testing Practices, 1995)

The *Code* was developed by the Joint Committee on Testing Practices as a supplement to the original *Standards for Educational and Psychological Testing.* It provides guidance separately for both test developers and test users. In general, test developers are instructed to provide information and supporting evidence that test users need to select appropriate tests. This includes providing evidence of what the test measures, the intended test takers, and evidence on the performance of diverse subgroups. Test developers are to provide guidelines for assessing individuals who need special accommodations or those with diverse linguistic backgrounds. The *Code* instructs test users to select tests that meet the intended purpose and are appropriate for

the test taker's individual characteristics. Accommodations or modifications that depart from standardized procedures are to be well documented. In addition, test results from modified test administrations are to be interpreted taking into consideration the impact modifications may have had on test results.

Summary

Today, practices in the area of assessment of ELL children are guided by a series of court decisions and legislation intended to safeguard the rights of all children and guarantee a free and appropriate education. In addition, several professional organizations have developed ethical standards for working with linguistically and culturally diverse children. Despite these guidelines, researchers and practitioners continue to struggle to address the various problems inherent in assessing ELL children. Professionals, burdened by practical limitations and often lacking sufficient knowledge and experience continue to have difficulty implementing the standards in their daily practice (Gopaul-McNicol & Armour-Thomas, 2002). Historically, there has been a significant shortage of instruments validated with a variety of language groups. In addition, ELL students continue to be disproportionately represented in special education programs (Macias, 1998). Thus, it is imperative that the assessment practices of school psychologists and other professionals continue to be evaluated and, when possible, improved in order to provide appropriate assessments of language minority students.

SECTION 3

ASSESSMENT METHODS AND CONSIDERATIONS

In order to conduct nondiscriminatory and nonbiased assessment of linguistically diverse individuals, practitioners must become well acquainted with the various methodological issues that affect assessment validity with this population. Because of the various complicating factors that are associated with language and culture, there are many ways that assessment of ELL students differs from assessment of children whose native language is English. Practitioners must consider internal factors such as the student's language, academic achievement, and cognitive ability as well as external factors such as the impact of culture, educational history, and family issues. This is often a complicated and difficult task. The following guidelines are provided to help practitioners avoid potential bias in the various stages of the assessment process.

Important Assessment Components

Cognitive assessment of children is most often completed as part of a more comprehensive psychological or psychoeducational evaluation. In order to accurately interpret cognitive assessment data, it is imperative that results be examined taking into account data and information from a variety of sources and assessment methods.

Review of Records

The first of these assessment components is often the process of reviewing existing data. In the case of an English language learner who is evaluated in the school setting, generally a large amount information is gathered over the course of the student's school career. School records may include information such as academic and language

proficiency test results, work samples, decisions made by bilingual education and ESL committees, language(s) of instruction at each grade level, grades and teacher reports, health history, and individualized education plans (Ortiz & Yates, 2002). Especially close attention should be paid to the student's school history. For example, interruptions in schooling, location and number of schools attended, grades enrolled in and completed, history of retention, and special services previously received should be noted. These can all have a significant impact on students' academic progress. For example, students may experience academic difficulties primarily due to frequent moves or disruptions in their academic program. Hispanic students' families may move back and forth between the United States and a Spanish speaking country. This makes it difficult to establish competency in either language. The situation is complicated when some students are totally out of school for an extended period of time while the family transitions from one place to another. Background information is crucial in order to distinguish between a student's lack of opportunity to learn and actual learning difficulties within the child.

Interviews

Further background information should be obtained through interviews conducted with parents, teachers, and the student (Rhodes, 2005b).
Information gained through a comprehensive interview can provide important data regarding the child's developmental, environmental, educational, linguistic background, and family history. Rhodes (2005b) recommends that practitioners establish a structured interview format to enable a translated version to be presented to parents in their native language and to ensure that important topics are not overlooked. Rogers (1998) emphasizes the importance of including the child in the interviews and recommends directly questioning the child about his/her academic skills, social

adjustment, motivation to learn, and instructional needs. Rogers also recommends conducting the parent interview with the child present so the evaluator is able to note parent-child interactions and conversations and identify differences between the child's use of language at home and in the school setting. It is important to ask questions regarding the parents' educational background and experiences, attitude toward education, and expectations for their child's education.

Observations

Observations are another essential component of assessments of ELL children. One of the functions of the observations is to allow the practitioner to evaluate the instructional environment (Lopez, 1995). Observations, along with teacher interviews and analysis of permanent products can be used to determine whether the instructional program and classroom setting is a good fit for the student given his/her cultural and linguistic background. Observations serve to answer questions such as whether the appropriate languages are being used for instruction, whether the language demands of the classroom are appropriate, and whether the teacher has realistic expectations for the student.

Another purpose of observations is to compare the student's behavior with that of other children in the same environment. Rogers (1998) advises that observations should include comparisons with same-age, linguistically similar and linguistically different peers. Through this procedure evaluators can obtain a good deal of information about the match between the student's behaviors, the task at hand, and the behaviors of others in the same environment.

Language Proficiency

As discussed earlier, the IDEA (2004) dictates that assessment of English

language learners be conducted in their native language or primary mode of communication. This establishes language proficiency assessment as another integral part of the assessment of an ELL student. Language proficiency information is crucial not only in order for the examiner to select the language(s) of cognitive assessment, but also to identify appropriate measures and accurately interpret test results. Practitioners often have access to formal language proficiency test results contained in the student's educational records. In addition, several formal Spanish and English language proficiency measures are available to practitioners and can be useful in establishing language dominance and proficiency. Two widely used measures are the Woodcock Language Proficiency Battery - Revised, Spanish and English Forms, and the Peabody Picture Vocabulary Test - Revised and its Spanish equivalent, the Test de Vocabulario en Imagenes Peabody. Formal measures have been criticized, however, for their overemphasis on discrete aspects of language as well as their questionable validity and reliability (Lopez, 1995; Lopez, 1997; Ortiz & Yates, 2002). Best practices, as well as legal mandates, call for assessment of language proficiency using tools that measure a wide range of language skills while using informal as well as formal assessment measures (American Educational Research Association et al., 1999; Gopaul-McNicol & Armour-Thomas, 2002; Holtzman & Damico, 1991; Lopez, 1995; Ochoa & Ortiz, 2005; Rogers, 1998).

One informal assessment method is to collect a series of oral language samples via interviews with the student (Rogers, 1998). The language samples could be recorded either through the use of a tape or voice recorder or by taking written notes on the student's responses during the interview. The child's teacher may be in the best position to obtain these data due to his/her rapport with the student. Observations across various settings and natural situations are another form of informal language

assessment. By observing the student's interactions across a variety of settings, such as the classroom, playground, and family interactions, the assessor will be more likely to obtain a complete picture of the student's language profile. A student learning English might be observed to be very quiet in an English speaking classroom, for example, while observations of familial interactions in Spanish show the same student to be talkative and proficient. Questionnaires have been developed to allow parents to provide information on language use in the home. Finally, as was mentioned previously, parent interviews are often crucial to gain an understanding of language dynamics and proficiency in the home.

Lopez (1995) relates that language proficiency data should be interpreted taking into consideration several key issues related to language acquisition. First of all, it is important to understand that language proficiency includes both Basic Interpersonal Communicative Skills (BICS) and Cognitive Academic Language Proficiency Skills (CALPS) (Cummins, 1984). BICS is the level of proficiency needed to engage in casual conversation. CALPS, on the other hand, is the language proficiency someone needs to comprehend more challenging, academically related tasks. According to Cummins, it takes approximately two years to develop BICS in the second language while it takes five to seven years to develop CALPS proficiency. Both BICS and CALPS should be evaluated as part of a comprehensive assessment. A second issue is that as children are exposed to a second language, it is not unusual for them to show a loss of receptive and expressive language skills in their primary language. This language loss should not be confused with a language disability. In addition, as bilingual children acquire fluency in their second language, due to the variability seen in children's language skills acquisition, frequent assessments of their language abilities are warranted. Researchers recommend against using language proficiency assessments

that are more than six months old (Holtzman & Wilkinson, 1991; Rogers, 1998). Another important concept to keep in mind is that being dominant in one language does not necessarily imply proficiency in that language, as is the case for many ELL students. An ELL student could be dominant in Spanish, for example, yet because of language loss or limited use of Spanish in school could still be somewhat limited in Spanish, especially CALPS. English language learners' proficiency in each language skills often vary depending on the context in which the language is being used. A student might demonstrate stronger conversational skills in Spanish, his/her primary language, while showing stronger CALPS skills in English due to having received academic instruction in English.

Acculturation

In addition to taking into account linguistic factors when assessing Spanish speaking children, it is important to consider cultural factors as well. Acculturation, the process of adopting the cultural traits or social patterns of another group, often has a significant effect on ELL students' academic progress and performance on assessment measures (Gopaul-McNicol & Armour-Thomas, 2002). In general, intelligence tests tend to sample behaviors that are typical or valued by the culture of the test developers. Examinees who do not come from the mainstream US culture are likely at a disadvantage when given these tests. Traditional cognitive assessment measures have been criticized based on test items that may tap information that culturally different children may not be familiar with due to their lack of exposure to certain concepts (Lopez, 1997). Therefore, it is imperative that examiners be aware of children's level of acculturation as well as aspects of their culture that may adversely affect their performance on traditional measures. Although it may be impossible to totally eliminate bias using traditional measures, professionals can reduce the chance that

children are misidentified by considering acculturation factors. Practitioners are encouraged to consult with other professionals and review multicultural literature to become familiar with different cultures as well as issues related to acculturation and assessment.

As is the case in the assessment of language proficiency, evaluation of acculturation may involve both formal and informal measures. Assessment methods typically include interviews with the child and his/her family, direct observations, and questionnaires (Ortiz, 2005). Parent interviews may revolve around questions regarding the family's identification, participation, comfort, familiarity, knowledge, or affiliation with the customs, values, and language of mainstream US culture. Those interviewing children may ask questions such as what language they prefer using, who their friends are, what music they listen to, what television shows they watch, and what difficulties they may be having adjusting to the new culture. Drawings and play activities can also be useful tools when interviewing young children who are less verbal (Esquivel, 1988). Interviews with children should be conducted keeping in consideration that their level of acculturation may be different than that of their parents as they spend more time in public schools. Measuring acculturation via observations can be difficult as cultural variables are often latent and not easily measured. Nevertheless, observations can provide the examiner with such information as manner and style of dress, language use, and interactions with peers. In addition to observations in natural settings, practitioners are encouraged to observe behaviors during individual testing sessions. These may include the child's familiarity with test materials and procedures, language use patterns, conversational skills, statements regarding hobbies and interests, eye contact, and motivation.

Regarding acculturation questionnaires, there is no shortage of measures

available to practitioners. Chun, Organista, and Marin (2003) provide information on a number of scales of acculturation. Practitioners should ensure that the culture of the scale used matches that of the child's family. Examples of acculturation scales designed for use with Hispanics include the Acculturation Rating Scale for Mexican Americans-II (Cuellar, Arnold, & Maldonado, 1995) and the Bidimensional Acculturation Scale for Hispanics (Marin & Gamba, 1995). Scales such as these provide valuable acculturation information via questions on topics such as language, geographic history, identity, attitudes, work, and personal associations. Unfortunately, many acculturation scales, including those listed above, lack sufficient data examining their validity. In addition, scales may have been normed on specific subgroups (i.e. Cuban Americans) or on people living in specific geographic locations, limiting their utility with broader groups of children.

Academic Achievement

Poor academic performance is the primary reason ELL students are referred for special education assessment (Ortiz & Yates, 2002). Effective measurement of the student's levels of academic achievement, therefore, becomes an integral component of the assessment. Practitioners have a range of options regarding assessment measures and methods. In general, these include both standardized or norm-referenced measures and informal or alternative measures. Both have their advantages and disadvantages when used with ELL populations.

Standardized academic measures possess the advantage of allowing the examiner to compare the student's achievement to a specific peer group (Rogers, 1998). Norms are typically provided for the student's age group and grade level. Standardized academic tests allow for a prescribed administration and scoring format. This improves the objectivity of the evaluation. Another advantage is that many standardized measures

are considered to have a high degree of reliability and validity. Various test developers have created parallel Spanish versions of English achievement tests. Parallel English and Spanish achievement testing allows for comparisons of skills across languages, in the case of students who have received instruction in both languages. A good example of widely used parallel English and Spanish standardized achievement measures are the Woodcock-Johnson III Tests of Achievement (Woodcock, McGrew, & Mather, 2001a) and the Bateria III Woodcock-Munoz: Pruebas de Aprovechamiento (Munoz-Sandoval, Woodcock, McGrew, & Mather, 2005a).

Despite their widespread use, there are several disadvantages of using norm-referenced measures with ELL students. One criticism is that academic measures in English tend to measure bilingual students' language proficiency in English rather than assessing actual achievement or knowledge of academic content (Figueroa, 1990a). In order to have confidence in the validity of the academic test results, careful examination of the student's English proficiency needs to be conducted beforehand. The student must have the ability to understand the instructions and perform the various academic tasks. Another criticism of standardized achievement measures is that although some measures are available in the native language of the ELL student, the validity of their results are typically limited as many bilingual children have never received instruction in their primary language (Lopez, 1995). Finally, norm-referenced measures are typically not aligned with the student's curriculum (Baker & Good, 1995). Therefore, they may be inadequate in measuring how well students are acquiring the particular skills being taught in their classrooms.

Because of the limitations of standardized measures of academic achievement with ELL students, several alternative methods have been developed (Baker & Good, 1995; Lopez, 1995; Ortiz & Yates, 2002). One of the most common is curriculum-

based assessment (CBA). CBA is described as the process of determining a student's instructional needs by directly assessing specific curriculum skills (Lopez, 1995). CBA activities include tasks such as informal reading inventories and use the students' curriculum materials as the foundation of the assessment. Curriculum-based measurement (CBM) is a widely used form of CBA that involves the administration of brief fluency probes of reading, spelling, written language, and mathematics computation (Shinn et al., 1998). Preliminary research has shown CBM to be a valid and nonbiased measure of reading skills in Hispanic and Spanish speaking populations (Baker et al., 1995; Knoff & Dean, 1994; Shinn et al., 1998). An advantage of CBM is its sensitivity to small changes in performance. In addition, CBM probes are brief and have many alternate forms. These characteristics allow the examiner to use CBM probes on a repeated basis to track students' acquisition of basic academic skills over time and closely monitor progress.

Criterion-referenced assessment is another alternative to standardized academic measures. The aim of criterion-referenced assessment is to compare the performance of a student to a specific criterion rather than to the performance of a norm group (Rhodes, 2005a). An advantage of criterion-referenced measures is that they can be created by the examiner and can be easily adapted depending on the individual student and criterion. An example of a commercially produced criterion-referenced measure in Spanish is the Brigance Diagnostic Assessment of Basic Skills, Spanish (Brigance & Messer, 1984).

Another alternative academic assessment method is portfolio assessment. Portfolio assessment involves collecting samples of students' work over a period of time and evaluating the samples against specific criteria. An advantage of portfolio assessment is that it provides an analysis of achievement over time and in different

areas, including language development and achievement in both the student's native language and English (Ortiz et al., 2002). Another advantage is that students are involved in their own assessment as they are typically largely responsible for creating the portfolio.

Rhodes (2005a) identifies several disadvantages of using informal measures of academic achievement to assess the academic achievement of ELL students. One of these is that the development and application of criterion-referenced and curriculum-based assessments can vary widely from teacher to teacher. A second limitation is the teachers must be careful about "teaching to the test" or scores may be an inaccurate representation of true achievement levels. Lastly, the use of informal measures by themselves may not provide sufficient academic information necessary to make eligibility and service provision decisions.

Other Issues Related to Assessment

Because of the complexities introduced by cultural and linguistic factors, the assessment of ELL students is often a daunting task. Literature and discussion has grown over the past few decades, however, providing professionals with a framework for current practice. In addition to the various components of assessment already discussed, there are several important issues to consider in the evaluation of ELL children.

Competencies of the Examiner

In order to conduct accurate and nonbiased assessments of ELL children it is imperative that efforts be made to ensure professionals are qualified, having received appropriate instruction and practice in the areas of cross-cultural psychology and psychological assessment. It has been argued that experts in the field have focused only

on the development of reliable and valid assessment instruments for use with minority groups and not on the competencies of the professionals who are administering the particular instrument (Rogers, 1998). Several characteristics or qualifications of professionals working with linguistic and cultural minority groups have been suggested. These are outlined below. Before conducting an assessment of an ELL child, professionals should assess their own qualifications and determine whether they have the background to work effectively with this population. If they find they lack the necessary experience and skills, steps should be taken to seek consultation with other professionals or to refer the child to another evaluator (APA, 2002).

First, evaluators should possess a knowledge base in cross-cultural psychology (Esquivel, 1988; Ortiz et al., 2002; Ortiz, 2002; Rogers, 1998). They should be sensitive to ways culture affects learning and impacts assessment. Chamberlain and Medinos-Landurand (1991) relate that several cultural traits of the child being evaluated should be considered by the examiner. These include child-rearing and schooling differences, sociocultural position and role of the culture within society as a whole, attitudes in test-taking, value of competition, and adjustment to the artificiality of the testing situation. Chamberlain and Medinos-Landurand also discuss several problems related to cultural insensitivity. One complication is there may be misperceptions between the culturally or linguistically diverse student and the evaluator. This leads, in turn, to the evaluator and student having different understandings or expectations in the evaluation process. Immigrant children, for example, may not be familiar with testing situations and unlike most children, may not understand that testing is often used for evaluation to demonstrate learning and may be used for placement decisions. They may be less motivated to perform in testing situations. The student's unfamiliarity with testing situations and low test motivation may be perceived by the examiner as

indications of deficiencies. Such misperceptions may lead to inappropriate referrals for assessment, faulty test interpretations, and unfounded placement decisions. Another problem with cultural insensitivity relates to the issue of cross-cultural stereotyping and bias. Stereotyping can occur when students are identified as possessing particular intrinsic traits when they merely demonstrate behavioral differences. Professionals are encouraged to become more sensitive to cultural issues by evaluating their own value system, cultural backgrounds, and beliefs. This will lead to the identification of the degree to which stereotyping and bias are present in themselves and others, as well as the manner in which they negatively impact the students' school environment.

A second important qualification of examiners of ELL students is that they have received extensive coursework and training in the construction, selection, use, and interpretation of tests (Rogers, 1998). If evaluators are well-trained in the appropriate use of tests, including issues related to standardization, validity, reliability, and limitations of norm-referenced tests, they will be more prepared to conduct non-biased assessment.

A third qualification is that evaluators have firsthand exposure to and supervised casework experience with racial, ethnic, and linguistic minority children (Rogers, 1998). Professionals who do not have this opportunity in their university training should enhance their skills through self study, professional development, in-service training or through a mentoring relationships in the field (Scribner, 2002). Only through practical experience will examiners be able to synthesize theoretical information gained from their coursework with hand-on experiences.

A final qualification of evaluators of ELL children is that they be competent in the language of the individual being assessed (Esquivel, 1988; Ortiz, 2002). Ortiz describes linguistic competence as the ability to communicate effectively in an

individual's native language and possession of a knowledge base related to first and second language development. As discussed earlier, ethical and legal guidelines dictate that children be evaluated in their primary or native language (Diana v California, 1970; IDEA, 2004; American Educational Research Association et al., 1999). The number one option in meeting this guideline is for the evaluator to be bilingual. Unfortunately, there are a limited number of bilingual psychologists and other evaluators (Ochoa et al., 1997). In addition, the vast numbers of languages spoken by ELL students in U.S. schools (Kindler, 2002) make it seemingly impossible for evaluators to be available in the language of the student in every case. A solution to this dilemma has been to rely on the services of interpreters to assist in the assessment process.

The Use of Interpreters

Unfortunately, there exists a lack of research on the effect interpreters have on the assessment process. There is, however, agreement among experts in the field on various potential problems of using interpreters. Many problems arise when the interpreter is not properly trained in test administration procedures (Figueroa, 1990; Holtzman et al., 1991; Ortiz et al., 2002; Rogers, 1998). Results of one study indicated that inexperienced and untrained interpreters tend to make mistakes in the process of translating IQ test questions from English to Spanish (Lopez, 1994). Results of a separate study on the use of trained interpreters during diagnostic testing (Sanchez-Boyce, 2000) indicated that this practice adversely affects validity and reliability in the assessment of bilingual children. Researchers in this study found that the test administration directions were often not followed accurately. In translating test items on the spot, interpreters may omit, add, or substitute terms that may significantly alter the content of the question. In addition, interpreters may engage in subtle prompting behaviors that inadvertently help the examinee. An option is to have the interpreter

translate test items prior to administration (Lopez, 1997). Unfortunately, this practice is not problem-free as the interpreter still may alter the content of the test, adversely affecting its reliability and validity.

In order to minimize errors in assessment, the interpreter should be as fluent in Spanish as possible, understanding the pragmatics and nuances of the language (Plata, 1993). Section 9.11 of the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999) emphasizes this by stating, "When an interpreter is used in testing, the interpreter should be fluent in both the language of the test and the examinee's native language, should have expertise in translating, and should have a basic understanding of the assessment process." It is important for the interpreter to understand the importance of following standardized testing procedures, including the importance of accurately conveying an examinee's actual responses. Interpreters should be familiar with the Hispanic culture in particular regions. Finally, interpreters should be trained regarding ethical issues such as maintaining confidentiality.

When using interpreters in assessment, the examiner is encouraged to provide interpreters with opportunities to ask questions during the testing session (Lopez, 1995; Lopez, 2002). The examiner and interpreter should take time following the session to discuss any difficulties encountered in translation as well as cultural factors that may have influenced the child's behaviors. In addition, the use of an interpreter should be documented in the evaluation report. Information on how the interpreter was used, as well as possible impacts on the validity of results should be noted.

Problems with using interpreters exist even if they are properly trained and instructed. Regarding best practices, Figueroa (1990b) calls into question the validity of evaluations conducted by interpreters because of the lack of empirical evidence

supporting the practice. Translating a test that was developed and normed on an English-speaking population may not yield a technically equivalent form of the test. Various words in English do not have an equivalent Spanish translation. According to the Standards for Educational and Psychological Testing (American Educational Research Association et al., 1999), evidence of test comparability when tests are translated into a different language must be provided. No such evidence is provided for tests administered by an interpreter. Practitioners are cautioned to use interpreters in assessment only as a last resort, when a bilingual examiner is not available.

Summary

The assessment of ELL Spanish speaking students is accompanied by a variety of methodological and procedural issues. Based on the literature in this area, several recommendations appear warranted. First, evaluators working with ELL students must utilize a variety of assessment methods and sources of information. These include a review of records; interviews with parents, teachers, and students; observations in multiple settings; and standardized as well as informal assessment measures. It is important that the child's language proficiency in both English and Spanish be accurately evaluated. In addition, cultural factors, including the child and family's levels of acculturation, need to be considered. A range of measures of academic achievement are available to practitioners. Those who conduct evaluations of ELL students should possess certain characteristics or qualifications. These include knowledge of the student's culture and cross-cultural psychology in general, first-hand experience and training working with cultural and linguistic minorities, and general training in psychoeducational assessment practices. Many monolingual examiners find that they require the assistance of an interpreter during testing. If an interpreter is used,

steps should be taken to ensure interpreters are properly trained. Testing through the use of an interpreter is only recommended, however, as a last resort as its validity has not been established. By following these guidelines, evaluators will be better prepared to conduct non-biased assessments of ELL students.

SECTION 4

REVIEW OF COGNITIVE ASSESSMENT MEASURES

In the assessment of ELL students, the information gained through observations, interviews, and language proficiency assessment can be used to guide the selection of appropriate cognitive assessment measures. This section will serve to highlight various cognitive assessment measures that have been evaluated in the literature and show promise in their use with Spanish speaking children. In actual practice, measures should be selected keeping in mind the unique characteristics of the child as well as the specific referral questions. The issue of possible test bias as a result of using inappropriate testing instruments with ELL students is particularly important. Reynolds, Lowe, and Saenz (1999) define test bias as "systematic error in the measurement of a psychological attribute as a function of membership in one or another cultural or racial subgroup." Systematic error, or bias, will be addressed in this paper by examining the external or predictive validity as well as the internal or construct validity of the various cognitive measures when used with Hispanic or Spanish speaking individuals. Tests may be considered biased if they are shown to measure a different construct or lack predictive ability when used with Hispanic or Spanish speaking individuals compared to the general population. Consideration will also be given to test reliability, interpretation, and limitations. Independent empirical studies examining the validity and reliability of each measure when used with Spanish speaking populations will be reviewed, as well as the technical dimensions of the instruments presented by the test authors. By examining the psychometric properties of these measures, practitioners will be better prepared to conduct cognitive assessments in a manner that is defensible and as non-discriminatory as possible.

Examiners have several options when deciding upon a cognitive measure. One

option is to use traditional cognitive measures in English such as the Wechsler Intelligence Scale for Children - Fourth Edition (Wechsler, 2003) or the Woodcock-Johnson III Tests of Cognitive Abilities (Woodcock, McGrew, & Mather, 2001). Using a traditional intelligence test in English with a bilingual Spanish/English speaking student has several limitations (Armour-Thomas, 1992; Figueroa, 1990b; Holtzman & Wilkinson, 1991; Lopez, 1997; Ortiz & Ochoa, 2005). One criticism of traditional English tests has been their lack of representation of bilingual or ELL students in the norming samples. Their norms have been based largely on mainstream students in the United States and may be inappropriate for use with culturally or linguistically different students. Another criticism relates to test item bias. Items may tap information that bilingual children are unfamiliar with due to their linguistically or culturally different backgrounds or lack of exposure to particular concepts. In addition, a student with limited English proficiency may have difficulty understanding the nature of the various assessment tasks when given complex verbal directions. A third criticism of traditional intelligence tests administered in English is that they do not measure the same constructs when given to an ELL student as they do with monolingual English speaking student. Instead of measuring verbal cognitive ability, for example, various measures may be more accurately described as measures of English proficiency.

The difficulties associated with using tests developed for use with English speaking children with bilingual English/Spanish speaking students has led to the use of translated tests. This allows the individual to be assessed in his/her primary or dominant language. Two current comprehensive intelligence tests that have been translated into Spanish are the Wechsler Intelligence Scale for Children - Fourth Edition, Spanish (WISC-IV Spanish; Wechsler, 2005) and the Bateria III Woodcock-Munoz: Pruebas de Habilidades Cognitivas (Bateria III COG; Munoz-Sandoval,

Woodcock, McGrew, & Mather, 2005b). Historically, however, there has been a dearth of appropriate Spanish measures of cognitive ability for children living in the U.S. Many translated measures have been criticized for their reliance on the original English norms (Lopez, 1997). Other translated intelligence tests such as the Escala de Inteligencia Wechsler - Revisada para el Nivel Escolar (Wechsler, 1984) have been normed outside the United States on Spanish speaking populations (Lopez, 1997; Figueroa, 1990b). These tests are considered to have questionable content validity as they were not normed on children living in the U.S.

When a formal translated test in Spanish is not available, school psychologists have often resorted to translating test items "in session" or by intermixing the child's first language and English during administration (Ochoa et al., 1996). These practices are not recommended as they represent a departure from standardized procedures and invalidate test scores. Buitrago (1999) compared the performance of monolingual Spanish-speaking students on an informal, simultaneously translated Spanish version of the WISC-III and the Universal Nonverbal Intelligence Test (UNIT). Scores on the UNIT were consistently higher than scores on the informally translated WISC-III. Results suggested that differential performance between the two instruments may be attributable to the language and cultural loadings of the WISC-III and highlighted the difficulties of translating tests in-session. Although using an informal translation of a test may provide the examiner with valuable qualitative information, test scores should only be interpreted with caution, if they are used at all.

A third option is for the examiner to test the student in both English and Spanish, assuming the examiner is bilingual. An example of a unique measure designed to measure bilingual cognitive ability is the Bilingual Verbal Abilities Tests (BVAT; Munoz-Sandoval, Cummins, Sandoval, 1998a). Ortiz and Ochoa (2005) define

bilingual assessment as the "evaluation of a bilingual individual, by a bilingual examiner, in a bilingual manner... with both the examiner and the examinee free to use both languages as may be necessary or desired throughout the testing process" (p. 161). Bilingual testing is not simply assessing knowledge in the first and then the second language. Rather, it involves accessing information shared by the two languages as well as allowing the individual to freely code switch (shift from one language to another) as the situation indicates. Bilingual assessment is generally recommended as it allows for a more complete assessment of the student's verbal skills (Holtzman et al., 1991; Lopez, 1997). Testing bilingually is considered to minimize the risk of underestimating intelligence by allowing children to use their full range of knowledge. Unfortunately, testing bilingually may be considered a departure from standardized assessment procedures and there is little research to guide the practice of bilingual assessment. In addition, there are a limited number of bilingual school psychologists.

A fourth testing option available to practitioners are nonverbal tests of intelligence. These include unidimensional measures such as the Test of Nonverbal Intelligence - Third Edition (TONI-III; Brown, Sherbenou, & Johnsen, 1997) and comprehensive measure such as the Leiter International Performance Scale - Revised (Leiter-R; Roid & Miller, 1997) and the Universal Nonverbal Intelligence Test (UNIT; Bracken & McCullum, 1998). Unidimensional nonverbal tests measure a narrow aspect of intelligence through the use of progressive matrices while comprehensive tests measure multiple facts of intelligence (Bracken et al., 2001). Several experts in the field have indicated that the use of nonverbal measures with ELL students is appropriate, valid, and promising (Figueroa, 1990b; Holtzman & Wilkinson, 1991; Ochoa et al., 1996). Proponents of nonverbal tests indicate that by reducing the oral or spoken language requirements, nonverbal measures reduce or eliminate potential linguistic bias.

Also, it seems logical to use nonverbal measures in cases where students with limited English skills must be tested by English-speaking examiners. There are several disadvantages, however, to using nonverbal cognitive measures with ELL children. Particularly, their sole use to assess intelligence is questionable given that they typically measure a narrow range of abilities (Holtzman et al., 1991; Ortiz et al., 2005). This makes it difficult to accurately determine a student's global IQ as only a partial measure of the student's overall cognitive ability is obtained. In addition, although verbal cognitive abilities have been found to predict school achievement, there is little evidence to suggest a strong relationship between performance on nonverbal tests and academic success (Athanasiou, 2000; Lopez 1997). Consequently, using nonverbal IQ scores to predict ELL students' academic achievement should be done with caution.

Measures

Wechsler Scales in English

Since the development of the initial Wechsler Intelligence Scale for Children (WISC; Wechsler, 1949) the Wechsler scales have enjoyed widespread use within the field of psychological assessment. The Wechsler scales' use and popularity is apparent not only in the assessment of English speaking students but also ELL students. Ochoa et al. (1996) indicated that over half of the school psychologists surveyed reported using the WISC-R or WISC-III in English in their assessments of bilingual and LEP students.

The Wechsler scales have undergone several updates, revisions, and translations over the years. Following the development of the original WISC in 1949, the WISC-R (Wechsler, 1974) was published. The WISC-R was again revised in 1991 when it became the WISC-III (Wechsler, 1991). Many of items, subtests, and scales were retained in each revision. Seventy two percent of the WISC items, for example, were

retained for the WISC-R. In addition, changes to the basic structure, item content, and organization from the WISC-R to the WISC-III were relatively minimal, with most changes being cosmetic (Flanagan & Kaufman, 2004). The most recent version of the Wechsler scale, the WISC-IV (Wechsler, 2003), represents the most substantial revision to date; however, there remains a good deal of similarities between the scales. To a certain extent, this allows researchers and practitioners to take research conducted with the previous versions of the WISC into account when evaluating the most recent version. In order to gain an understanding of the usefulness of the WISC-IV with ELL students, it is helpful to know the history of the Wechsler scales, including advantages and criticisms of the previous English versions when used with linguistic minorities. While the WISC-IV has yet to be examined extensively with linguistic and cultural minorities, there is a generous amount of research available on earlier Wechsler scales.

A shortcoming of many IQ tests is that their norms are based on mainstream students and therefore may be inappropriate for use with cultural or linguistic minority students. Indeed the original WISC as well as the WISC-R were criticized for not including enough Hispanics and for having a disproportionate number of Hispanics with elevated socioeconomic status in their standardization samples (Holtzman et al., 1991). The developers of the WISC-III and WISC-IV took steps to ensure that the standardization samples were representative of the U.S. population according to race and parent educational level, among other variables (Wechsler, 1991; Wechsler, 2003). Children who were not fluent in English were not included in the standardization sample, rendering the WISC-IV inappropriate for use with students with limited English proficiency.

Various studies have shown that Hispanic children consistently exhibit characteristic and unique performance on the Wechsler scales. McShane and Cook

(1985) closely examined the performance of Hispanic children on the Wechsler scales by conducting a review of literature on the WISC and WISC-R. One of their findings was that Hispanic children, some of whom were identified as speaking English as a second language, consistently scored lower on the Full Scale IQ than white children in the standardization sample. Hispanic children included in the WISC-III standardization sample were reported to earn a mean Full Scale IQ score of 94, nine standard score points lower than the mean for white children (Wechsler, 1991).

It is important to note that mean scores differences between groups do not necessarily indicate test bias (Gutkin & Reynolds, 1980; Holtzman et al., 1991; Kaufman, 1994; McShane & Cook, 1985; Reynolds & Kaiser, 1990; Palmer, Olivarez, & Willson, 1989; Reynolds, Lowe, & Saenz, 1999). Current literature regarding test bias suggests it is more important to evaluate the differential construct and predictive validity across groups than to assume a test is biased based on mean score differences alone. Numerous studies have addressed the differential validity of the Wechsler scales for whites versus minority groups, including Hispanics. The majority of studies have found that the scales are not systematically biased against English-speaking minority group members (Ochoa et al., 2005a; Reynolds & Kaiser, 1990; Kaufman, 1994; Ortiz, 2004). Studies of the Wechsler scales with Hispanic children that support their use include studies of reliability (Dean, 1977), external or predictive validity (Cathers Schiffman, 2000; Johnson & McGowan, 1984; Weiss & Prifitera, 1995), and internal or construct validity (Gutkin et al., 1980; Reschly, 1978).

Another consistent research finding is that Hispanic children have consistently demonstrated a 10-15 point difference between the Performance and Verbal IQ scores on the Wechsler Scales (Figueroa, 1990a; McShane et al., 1985; Wilen & Sweeting, 1986). Performance scores have typically been shown to be higher than Verbal IQ

scores. Language proficiency is considered to adversely affect performance on verbal scales while having limited effect on performance tasks. Results of a recent study indicated that English language proficiency predicted the Verbal/Performance IQ discrepancy and also explained a significant amount of variance on the Verbal Comprehension scale of the WISC-III (Baldizon-de-Naclerio, 1999). Kaufman advises practitioners to not interpret bilingual and bicultural students' Full Scale IQ as it likely does not reflect their true intellectual potential (Kaufman, 1994).

The verbal/performance split shown by Hispanic children on the Wechsler scales has led to the recommendation to use only the subtests that make up the Performance IQ when assessing students who speak English as a second language (Bracken & McCallum, 2001; Figueroa, 1990b; Flanagan & Kaufman, 2004). This practice is problematic, however, for several reasons. First of all, as noted by Kaufman (1994), there is little empirical research on the nature and meaning of the verbal/performance discrepancy for ELL children. Other difficulties occur when psychologists assume that ELL students understand the verbal directions spoken by the examiner and therefore understand the nature of the task. Flanagan and Ortiz (2002) suggest that the characterization of the Performance IQ as a nonverbal measure is misleading because although the subtests do not require a verbal response, they often demand a high level of receptive language abilities in order to understand the test's instructions, as well as the examiner's expectations. Another limitation of the Performance IQ as an estimate of the intelligence of ELL children is that it measures a narrow range of abilities (Ortiz, 2004). The practice fails to take other into account other abilities that make up intelligence, potentially leading to the underestimation or overestimation of overall intelligence. Lastly, is it well documented that Performance IQ is not as strong as Verbal IQ in predicting academic achievement (Holtzman et al., 1991).

Although cross-cultural research with the Wechsler scales has been conducted for many years with a variety of ethnic groups, including children of Hispanic background, relatively few studies have examined the scales' validity with Spanish speaking students. Separate researchers in the early 1990s (Figueroa, 1990a; Holtzman et al., 1991) reviewed the literature at the time and concluded that no research data had been collected addressing the influence of limited English proficiency on test reliability and validity. Figueroa states, "The literature on bilingualism, second-language acquisition, bilingual education, and the measurement of language proficiency are generally overlooked or omitted from considerations of bias in intelligence tests" (p. 685). Unfortunately, there continues to be a dearth of studies conducted in these areas. Information from the limited studies conducted with Spanish speaking students on the Wechsler scales fails to provide practitioners with a clear picture regarding its utility.

In an encouraging study of the validity of the WISC-R, Lawlis, Stedman, and Cortner (1980) examined the WISC-R factor structure for a group of bilingual Mexican-American children. Students' bilingual status was established by means of a personal interview with each child's teacher. Results showed the general pattern of subtest loadings on the Perceptual Organization and Freedom from Distractibility factors was relatively similar to that of the standardization sample. The factor structure of the Verbal Comprehension factor was shown to be very similar to that of the standardization sample.

Other studies have provided data that calls into question the validity of the Wechsler scales with ELL children. Palmer, Olivarez, Willson, and Fordyce (1989) examined the predictive validity of the WISC-R with a sample of Anglo, Black, and Hispanic students using a test of regression slopes and intercepts. Approximately 38% of the Hispanic students in the study were identified as LEP. Results showed the

WISC-R to be biased with Hispanics and Blacks compared to Anglos as WISC-R results tended to overpredict scores for minority students on a measure of academic achievement. In addition, predictive bias due to limited proficiency in English was found for both the Performance and Verbal Scales on the WISC-R. The tendency of WISC-R results to overpredict academic achievement is problematic as there is an increased likelihood that referred ELL and Hispanic students will evidence a severe discrepancy between ability and achievement and, as a consequence, will be misidentified as learning disabled students.

Olivarez, Palmer, and Guillemard (1992) replicated the Palmer et al. (1989) study by using the WISC-R to predict achievement test scores in reading, math, and writing on the Woodcock-Johnson Achievement Test. Again, results provided evidence of bias across ethnic groups with Hispanic students' language dominance influencing the predictive relationship between IQ and achievement.

In a similar study, Mishra (1983) examined the validity of the IQ and factor scores from the WISC-R in their power to predict academic achievement on the Wide Range Achievement Test (WRAT). The sample consisted of children who predominantly spoke Spanish at home as well as in their conversations with friends and peers. Results showed low correlation coefficients between the WISC-R factor scores and achievement scores on the WRAT calling into question the predictive validity of the WISC-R with ELL students.

In a more recent study, Dicerbo (2003) examined the relationship between English language proficiency and performance on the WISC-III using a sample of Hispanic children. Students included in the sample showed relatively high levels of English language proficiency, as measured by the Woodcock-Munoz Language Survey (WMLS). Results of the study indicated that WMLS scores were a significant predictor

of WISC-III Verbal IQ and to a lesser extent, the Performance IQ. Dicerbo suggests that the WISC-III verbal scale, when used with LEP children, becomes a measure of language proficiency. Results also cast doubt on the validity of using the Performance IQ with ELL students. The authors noted that a specific level of English proficiency at which the WISC-III becomes valid could not be established.

In summary, research that addresses the validity of the Wechsler scales with Spanish speaking Hispanic students has led to mixed results and few conclusions. In general, there is evidence to suggest the scales are not systematically biased against Hispanics who are fluent in English. The scales demonstrate questionable validity, however, when used with Spanish speaking students, even those who have achieved moderately high levels of English proficiency. Studies also suggest that the use of the discrepancy model with ELLs may not be a valid practice as the relationship between IQ and achievement is not the same as it is for English speaking students. Instead of using the Wechsler scales in English with ELL Spanish speaking students, a more promising alternative may be the use of Spanish measures or nonverbal ability tests.

Previous Wechsler Scales in Spanish

Several Spanish translations and adaptations of the Wechsler scales have been developed. The first of these was the Escala de Inteligencia Wechsler para Ninos (EIWN), a Puerto Rican translation and adaptation of the WISC. To develop the EIWN, the order and presentation of items on the WISC was altered based on studies of item difficulty for Puerto Rican children (Wilen & Sweeting, 1986). Authors of the EIWN did not develop separate norms from the WISC. Very little is known about the psychometric properties of the EIWN. Practitioners have been recommended to interpret the EIWN results with caution as the mean IQ of the Puerto Rican sample was approximately 12 IQ points lower than the mean score of 100 for American children in

the WISC standardization sample (Wilen et al., 1986).

With the development of the WISC-R came the Spanish translation named the Escala de Inteligencia Wechsler para Ninos – Revisada (EIWN-R) in 1982. The EIWN-R was developed as an experimental translation of the WISC-R and as such, was not standardized when it was initially developed. Since its inception, isolated local norms have been developed for various groups. The EIWN-R was later standardized, for example, on 532 Cuban Americans living in Miami, Florida (Gass, Demsky, Martin, 1998). Like its predecessor, the EIWN, little is know about the psychometric properties of the EIWN-R. Gass, Demsky, and Martin (1998) compared the factor structure of the EIWN-R to that of the WISC-R using the EIWN-R standardization sample from Miami. Results of the factor analysis provided evidence for Verbal Comprehension and Perceptual Organization factors. The presence of a third factor, Freedom of Distractability, however, was not shown.

Other versions of the EIWN-R were normed outside the United States (Figueroa, 1990a; Lopez, 1997; McShane et al., 1985). A version of the EIWN-R has been used in Mexico, for example, for decades. The Escala de Inteligencia Wechsler – Revisada Para el Nivel Escolar (WISC-RM), was developed and standardized in 1983. The WISC-RM was normed on 1,100 students from Mexico City. Items from the Information, Vocabulary, and Comprehension subtests were revised to more accurately reflect Mexican culture. Mexican children included in the norming sample obtained a mean Verbal IQ of 89, a mean Performance IQ of 88, and a mean Full Scale IQ of 87. This is somewhat lower than the performance of Hispanic children on the Wechsler scales in English (Wechsler, 1991).

Another version of the EIWN-R that was normed outside of the continental U.S. is the Escala de Inteligencia Wechsler para Ninos – Revisada de Puerto Rico (EIWN-R-

PR). It was normed in 1992 on a sample of 2,200 Spanish-speaking children in Puerto Rico (Lopez, 1997). The authors of the EIWN-R-PR conducted concurrent validity studies with other intelligence measures used in Puerto Rico and examined the predictive validity of the scale using students' grade point averages (Jimenez, 2002). Results provided evidence to support the concurrent and predictive validity of the EIWN-R-PR. Although the EIWN-R-PR seems to be an adequately developed measure, it is likely only appropriate for use with Puerto Rican children or Puerto Ricans who have recently immigrated to the U.S. as this is the group on which it was standardized.

Because of limitations of previous Spanish translations of the WISC, namely outdated norms, lack of U.S. children in the norming sample, and unestablished reliability and validity, they are not appropriate for use with U.S. children. With the development of the Wechsler Intelligence Scale for Children – Fourth Edition, Spanish (WISC-IV, Spanish; Wechsler, 2005) many of these limitations have been addressed.

Wechsler Intelligence Scale for Children - Fourth Edition, Spanish

The WISC-IV Spanish is a translation and adaptation of the WISC-IV. Like the WISC-IV, the WISC-IV Spanish provides an overall full scale IQ score as well as the Verbal Comprehension Index, the Perceptual Reasoning Index, the Working Memory Index, and the Processing Speed Index (Wechsler, 2005). It includes 15 subtests, 14 of which were adapted from the WISC-IV, and one subtest adapted from the WISC-IV Integrated (Coding Copy). Although some of the verbal items were translated directly from the WISC-IV, others were developed solely for the WISC-IV Spanish to maintain levels of difficulty and clarity of item content. Authors of the WISC-IV Spanish took steps to incorporate language that would be familiar to the diverse Spanish speaking population in the U.S.

One of the characteristics of the WISC-IV Spanish that sets it apart from other

tests that have been translated into Spanish is its standardization and normative development. The WISC-IV was standardized on 851 Spanish-dominant children living in the U.S (Wechsler, 2005). The standardization sample was stratified by age, sex, parent education level, and primary guardian country of origin. Children who had completed more than 5 consecutive years of education in the continental U.S., as well as those that reported speaking or understanding English better than Spanish, were excluded from the standardization sample. To ensure that performances on the WISC-IV Spanish subtests were scaled to the norms developed for the U.S. population in general, subtest raw scores were calibrated to the total raw sores of the corresponding WISC-IV subtests.

Two studies were performed to evaluate the comparability of the WISC-IV Spanish and WISC-IV scores for Hispanic children (Wechsler, 2005). The first study compared children from the WISC-IV Spanish standardization sample with a group of Hispanic children from the WISC-IV standardization sample, matched on age, parent education level, and sex. Most composite scores for the different groups did not differ significantly. The mean FSIQ for the WISC-IV Spanish group was 92.1, while the mean score for the WISC-IV group was 94.1. Scores from the WISC-IV Spanish group were 2.5 points lower on the PRI, less than 1 point lower on the VCI, 6.1 points lower on the PSI, and 1.9 points higher on the WMI compared with scores of the WISC-IV control group. In a similar study (Wechsler, 2005), effect sizes for the mean composite scores were compared between the WISC-IV Spanish group and a control group of white/non-Hispanic origin children from the WISC-IV standardization sample. Again, most scores did not differ substantially between groups. The mean FSIQ score for the WISC-IV Spanish group was reported as 94.3 while the mean FSIQ of the white control group on the WISC-IV was 98.6. Differences between mean composite scores fell

between 3.6 to 5.5 points with the WISC-IV Spanish group scoring slightly lower. The exception was on the WMI where mean scores in the two samples were approximately equal. It should be noted that these results are in contrast with those of studies completed on previous Wechsler scales (Figueroa, 1990a; McShane et al., 1985; Wechsler, 1991; Wilen & Sweeting, 1986). Previous Wechsler scales have shown Hispanic children to earn FSIQ scores approximately nine points lower than white children while also demonstrating discrepancies of 10-15 points between verbal and performance scales.

In the WISC-IV Spanish manual the authors present a good deal of data to support its reliability. Reliability coefficients of the various subtests were shown to be good, with most coefficients ranging from .81 to .88 (Wechsler, 2005). The two exceptions were the Coding (.75) and Symbol Search (.74) subtests. As expected, reliability coefficients of the composite scores were higher, ranging from .82 (Processing Speed) to .97 (Full Scale). Test-retest reliability was examined using a sample of 55 children who were given the WISC-IV Spanish twice, with test-retest intervals ranging from 13 to 46 days with a mean interval of 27 days. Test-retest reliability for the various subtests ranged from .72 (adequate) on the Symbol Search subtest to .92 (excellent) on the Information subtest. In addition, stability coefficients for the composite scores ranged from excellent to good (.80s and .90s).

The WISC-IV Spanish test authors also provide evidence supporting the validity of the measure. First of all, intercorrelation studies generally showed that subtests of similar functioning correlated more highly with each other than with subtests measuring different types of functioning, providing evidence of construct validity. Secondly, exploratory and confirmatory factor analysis supported the factor model of the WISC-IV Spanish. Lastly, criterion-related validity of the WISC-IV Spanish was supported by

studies that examined the relationship of WISC-IV Spanish test scores with scores from measures such as the Universal Nonverbal Intelligence Test and Clinical Evaluation of Language Fundamentals - Third Edition. Unfortunately, the WISC-IV Spanish was not examined in relationship to measures of academic achievement.

A strength of the WISC-IV Spanish is that it provides two types of age-based percentile rank equivalents for composite scores. In addition to comparing the child's score to the general population, practitioners are given the option of evaluating a child's performance relative to Spanish-speaking children in the U.S. who are similar in terms of parent education and years of U.S. Educational experience (Wechsler, 2005). The authors provide as an example a child with 100 percent of his/her educational experience in the U.S. and parent education of 16 or more years. When compared to the general population, this child's FSIQ of 100 produces a percentile rank of 37. When compared to children with under 20% of their educational experience in the U.S. and parent education of less than 8 years, an identical score falls at the 90th percentile.

Although much additional research needs to be conducted with the WISC-IV Spanish, preliminary data provided by the test developers is encouraging. The WISC-IV Spanish seems to have overcome many of the limitations that plagued previous translations, namely, limited normative samples and questionable psychometric properties. A laudable feature that improves the accuracy and diagnostic utility of the WISC-IV Spanish is the inclusion of demographic tables that allow additional interpretation compared to all Hispanic children and subgroups of the Hispanic population. The authors provide a good deal of data to support the reliability and validity of the WISC-IV. Nonetheless, the lack of predictive validity studies, especially those that examine the correlation between the WISC-IV Spanish and measures of achievement, is a concern. Because of its standardized sample, the WISC-IV Spanish

should only be used with Spanish speaking students who have spent 5 or fewer consecutive years in school in the U.S.

Bateria III Woodcock-Munoz: Pruebas de Habilidades Cognitivas

Another comprehensive intelligence test that has been developed for use with Spanish speaking individuals is the Bateria III Woodcock-Munoz: Pruebas de Habilidades Cognitivas (Munoz-Sandoval et al., 2005b). The Bateria III COG is the third revision of a Spanish test originally published as the Bateria Woodcock Psico-educativa en Espanol (Bateria; Woodcock, 1982) and subsequently revised as the Bateria Woodcock-Munoz: Pruebas de Habilidad Cognitiva - Revisada (Bateria-R COG; Woodcock & Munoz-Sandoval, 1996). The Bateria III COG is the parallel Spanish version of the WJ III Tests of Cognitive Abilities (Woodcock et al., 2001b). Like the English version, the Bateria III COG is based on the Cattell-Horn-Carroll theory of cognitive abilities (Schrank, McGrew, Ruef, Alvarado, Munoz-Sandoval, & Woodcock, 2005).

A panel of professionally certified Spanish translators and native Spanish speakers from various countries provided assistance on the suitability of Bateria III COG item content, test translation, and adaptation (Schrank et al., 2005). The authors paid particular attention to ensure that items and test instructions were appropriate for all Spanish speaking regions.

The tests included in the Bateria III COG are translated or adapted versions of the WJ III COG tests (Schrank et al., 2005). Through the use of a calibration sample, Bateria III COG data were equated to the WJ III COG norms. Items for each WJ COG Spanish test were rescaled, or equated, to the WJ III COG according to the empirical difficulty of counterpart tasks in English. The calibrating and equating method used to equate the Bateria III COG and the WJ III COG involves several steps and is described

in the Overview and Technical Supplement. The calibration sample consisted of 1,413 native Spanish-Speaking individuals from inside and outside the U.S. Included in the sample were individuals from Mexico, the U.S., Costa Rica, Panama, Argentina, Columbia, Puerto Rico, and Spain. Mexico was the country with the highest representation in the sample, with 417 participants. Of the 279 participants from the U.S., 135 were born outside of the U.S. Compared to standardization samples of other cognitive measures, limited demographic, socioeconomic, and technical information is provided for the calibration-standardization sample of the WJ III COG.

In addition to providing the user with the same test and cluster scores as the WJ III COG, supplementary interpretation features of the Bateria III COG, such as the Comparative Language Index (CLI), are provided (Schrank, 2005). The CLI can be used when specific tests from both the WJ III COG and the Bateria are administered. The CLI score provides comparative information that provides evidence of language proficiency and illustrates which of the two languages is dominant.

Limited reliability and validity data are presented in the 28 page Overview and Technical Supplement (Schrank et al., 2005) that is provided with the Bateria III COG. Reliability data is presented for only 11 of 31 individual tests and 4 of 26 clusters or composite scores. Internal consistency reliability coefficients of the various tests fall between .80 and .93 while coefficients for the cluster scores fall between .88 and .94. Confirmatory factor analyses were conducted to evaluate the internal structure of the Bateria III COG. Results supported the organizational structure of the measure based on CHC theory. The test authors refer users to the Manual Tecnico, which is translation of the WJ III Technical Manual, for basic reliability and validity information on the WJ III. The authors state, "Because the Bateria III calibration data is equated to the WJ III norms, the underlying psychometric characteristics of the WJ III apply to the Bateria

III." (p. 17). The technical properties, including reliability and validity, of the WJ III have been described as exceptional (Cizek, 2003; Sandoval, 2003).

Saffron (2000) examined the validity of the Bateria-R COG, the predecessor of the Bateria III COG published in 1996. Saffron's research examined the predictive validity of select subtests of the Bateria-R COG, specifically, those that measure auditory processing (Ga) and Crystalized Intelligence (Gc). Results were inconsistent on the ability of the measures to predict reading skills of Spanish speaking students. Ga was found to be a strong predictor of reading in Spanish but not in English. Gc was shown to be a strong predictor of both English and Spanish reading.

Another study examined the construct comparability of the WJ III COG and the Bateria III COG (McCreith, 2005). This was done by evaluating whether the dimensionality and structure of each of the selected tests were the same and by examining whether specific items functioned differentially for English and Spanish speaking examinees. First, multiple bilingual reviewers completed a judgmental review process in which they compared the instructions and items of the Bateria III COG and WJ III COG, identified differences between the two versions, and judged whether this difference would provide one group with an advantage or disadvantage. The judgmental review process did not reveal significant differences in the items of the English and Spanish versions. Reviewers related that all the tests were translated well. Next, test equivalence was evaluated using factor analytic methods as well as item response theory analyses, including differential item functioning (DIF). Results indicated a high degree of comparability for the different language versions on the Concept Formation and Analysis-Synthesis tests. Empirical examination of the Spatial Relations test, however, indicated the two versions were not comparable. Analysis of the item level data for this test showed a relatively high number of DIF items. Six out

of seven items examined on the Spatial Relations test were shown to function differently between the language versions, with three items that were easier for the English-speaking examinees and three items that were easier for Spanish-speaking examinees. In addition, Spatial Relations was the only test on which there was a large difference between the internal consistency of scores for the two language versions.

One of the strengths of the Bateria III COG is the measure's alignment with a well-defined and empirically validated theory of intelligence. Yet another strength is that it is equated to the WJ III tests, which have historically demonstrated excellent reliability and validity. It is apparent that care was taken in the translation and calibration of the Bateria III COG. Caution should be used, however, when the Bateria III COG is used to evaluate children in other Spanish speaking countries as well as those who have recently moved to the U.S. It may not be appropriate to derive norm-referenced scores for these individuals based on a U.S. standardization sample. Unfortunately, little reliability and validity data is presented for the Bateria III COG. It may be erroneous to assume that because the Bateria III calibration data is equated to the WJ III norms, the validity of the two measures is equivalent. Additional research should be conducted to examine the psychometric properties of the Bateria III COG.

Bilingual Verbal Ability Tests

Historically, bilingual students' verbal cognitive ability has been tested in three ways: exclusively in English, exclusively in Spanish, or with separate measures in English and Spanish. As was discussed earlier, these practices have been problematic. The Bilingual Verbal Ability Tests (Munoz-Sandoval et al., 1998a) is a unique measure designed to provide equitable assessment of bilingual individuals by evaluating skills in both English and the child's primary language. It represents the first attempt to create a standardized procedure for combining verbal cognitive abilities in the first and second

language within the same instrument.

The BVAT has been developed for use in English and 17 other languages, including Spanish (Munoz-Sandoval, Cummins, Alvarado, & Ruef, 1998b). It contains three tests originating from the Woodcock Language Proficiency Battery - Revised which, in turn, were taken from the Woodcock-Johnson Revised Tests of Cognitive Ability. They include the Picture Vocabulary Test, the Oral Vocabulary Test, and the Verbal Analogies Test. All three tests were translated directly into the second language. In each English subtest, the level of difficulty gradually increases. The BVAT's standardized assessment procedure requires that the three English tests be administered first. The examiner then re-administers all items missed on the English test in the student's primary language. Testing continues until a new ceiling is established in the student's first language. The computerized scoring program provides scores for English Language Proficiency (ELP) and for Bilingual Verbal Ability (BVA). In addition to providing an estimate of verbal cognitive ability, comparisons of the ELP and BVA scores yields valuable information such as where the student is in the second language acquisition process.

The Comprehensive Manual (Munoz-Sandoval et al., 1998b) provides evidence regarding the reliability and validity of the BVAT. Norms and reliability data for the BVAT were based on a subset of the data used to standardize the WJ-R COG. The school-age sample data were gathered from 1986 to 1988. Subtest reliabilities were reported based on split-half analyses of the norming sample and were corrected for length by the Spearman-Brown formula. Median subtest reliabilities were reported to be strong, ranging from .89 to .90. The ELP median reliability was shown to be .96. The authors also reported a reliability index of .84 based on parallel form reliability for a bilingual Spanish/English speaking sample.

The authors of the BVAT made efforts to insure content validity by undertaking an 8-step procedure designed to ensure the comparability of translation (Munoz-Sandoval, 1998). Items that could not be translated equitably into the various languages were excluded. This includes three items on the Spanish tests. Five concurrent validity studies are reported in the Comprehensive Manual, using as criteria eight well-known tests of verbal abilities and language proficiency. Correlation coefficients fall within the range of .7 to .9. The authors also provide evidence indicating high correlations between the BVAT and academic achievement. Results from three separate validity studies indicated correlations between the BVAT and broad measures of achievement that range from .57 to .87., with most correlations falling in the mid .80s.

Alvarado (2000) conducted an independent study to evaluate the validity of the BVAT. The study compared and predicted associations of the BVAT with external criteria. Ninety bilingual Spanish/English speaking students were grouped into three bilingual categories: bilingual English dominant, bilingual Spanish dominant, and balanced bilingual. Test results from the BVAT were compared to those from the WJ-R COG, the Bateria-R COG, the WISC-III, and the TONI-III. Moderate to high intercorrelations were found between the three BVAT subtests, lending credibility to the construct validity of the measure. Comparisons between the BVAT and the other monolingual verbal ability scales showed the Bilingual Verbal Ability (BVA) score to be significantly higher for the total sample. While the BVA standard score mean fell at 94, mean scores of the monolingual verbal ability tests tended to fall in the low to mid 80s. Mean BVA scores were not consistently higher, however, in all bilingual groups. The mean score on the BVAT Picture Vocabulary test for the bilingual Spanish dominant group was 78, at least 15 standard score points lower than scores on the other two tests. This depressed the overall BVA score for this group. Alvarado noted that the

cultural content of the Picture vocabulary items, as well as the translation of English test items into Spanish, may explain the depressed Picture Vocabulary scores for Spanish dominant individuals. Alvarado related that caution appears warranted when using the BVAT Picture Vocabulary test with bilingual Spanish dominant students.

Other possible sources of bias when using the Spanish translation of the BVAT relate to the Oral Vocabulary and Verbal Analogies subtests. Administration of these items includes the presentation of written prompts that parallel the verbal questions. Students who do not read in Spanish may be at a disadvantage. Also, the Picture Vocabulary items on the Spanish version are presented in black and white compared to color for the English version. Presentation of the pictures in black and white may be less engaging to the student and may leave out visual cues that assist in the identification of the picture.

In summary, the BVAT is considered to be an original, influential, and effective measure of bilingual verbal ability. Its strengths lie in its groundbreaking design, well-written and comprehensive manual, ease of administration, and availability in numerous languages. Validity and reliability of the measure appear to be adequate. Correlational studies presented by the authors indicate the BVAT correlates highly with other measures of verbal ability as well as measures of academic achievement. Independent research should be completed to replicate results and further establish reliability and validity for the bilingual population for which the BVAT was designed. Unfortunately, a separate standardization sample was not gathered for the BVAT, which instead relies on the WJ-R COG sample for norming purposes. The BVAT test norms are outdated, having been collected from 1986 to 1988. Results of the BVAT with Spanish dominant students should be interpreted with caution as the Picture Vocabulary test may underestimate their true abilities. Lastly, the BVAT should not be considered a

comprehensive measure of cognitive ability as it only measures verbal ability.

Test of Nonverbal Intelligence - Third Edition

The unidimensional nonverbal test most often used by school psychologists in their assessment of bilingual and ELL students is the Test of Nonverbal Intelligence (Ochoa et al., 1996). The most current revision of the Test of Nonverbal Intelligence is the TONI-3 (Brown et al., 1997). The TONI-3 was designed to be a "language-free measure of abstract/figural problem solving" (p. 28). The TONI-3 has two equivalent forms, Form A and Form B, each containing 45 items. The TONI-3 administration and response format eliminates all language usage and attempts to reduce motoric and cultural factors. In general, it adheres to the guidelines set by Jensen (1980) for language-free and culturally reduced nonverbal tests. Namely, it is not timed, it uses novel problems to decrease the impact of prior exposure, it uses performance measures instead of paper and pencil tasks, it includes practice items, and includes instructions that are pantomimed to the examinee. These qualities enable the TONI-3 to be used effectively with students who are often not amenable to traditional measures such as the WISC-IV. This includes linguistic and cultural minority students, deaf children or those with significant language impairments, and students with motor impairments. However, the TONI-3's nonverbal testing procedures may make administration to gifted or nonhandicapped students unnecessarily awkward (Atlas, 2001).

The TONI-3 authors note that abstract/figural problem solving was selected as the core of the TONI-3 as it appears to be a general and important component or construct of intelligence (Brown et al., 1997). In addition, it is thought to be a pervasive activity that estimates the individual's level of overall intellectual functioning. The narrow focus in terms of abilities measured, however, is one of the limitations or criticisms of the TONI-3 (Bracken & McCallum, 2001; Lopez, 1997). The TONI-3

does not sample important cognitive dimensions such as memory that are components of most major theories of intelligence.

The TONI-3 manual provides a variety of data addressing its technical properties. The standardization sample consisted of 3,451 individuals chosen to represent the U.S. population according to geography, gender, community type, ethnicity and race, disabling condition, and socioeconomic status (Brown et al., 1997). Ninety individuals who speak English as a second language, or 2 percent of the standardization sample, were included in the standardization sample. This is a rather small number given that the LEP student population in the U.S. is estimated at 9.3 percent (Kindler, 2002).

To evaluate reliability, coefficient alpha and standard errors of measurement were calculated for 20 age intervals (Brown et al., 1997). The average coefficient for both form A and form B was high, falling at .93. Coefficients for ages 6, 9, and 10 on Form A and age 10 on Form B were shown to be somewhat lower, falling at .89. Standard errors of measurement ranged from 3 to 5. The coefficient alpha for the English as a second language sample was shown to be .95. Test-retest stability of the TONI-3 at one week intervals was shown to be between .89 and .94 for groups ages 13, 15, and 19 to 40. Evidence was also provided supporting the TONI-3's interscorer reliability.

The TONI-3 manual also presents a good deal of data regarding validity (Brown et al., 1997). Several correlational studies were reported by the authors. Correlations with other measures of intelligence were stated in the manual as moderate to high. Correlations with full scale or overall IQ scores ranged from .63 with the WISC-III to .76 with the Comprehensive Test of Nonverbal Intelligence (CTONI). Interestingly, correlations between the TONI-3 and the WISC-III Verbal Scale were .59 and .53 for

Form A and Form B while correlations with the Performance Scale were not significantly higher, at .56 and .58. The correlations between the TONI-3 and the WISC-III reported in the manual may be most accurately described as moderate (Atlas, 2001). Correlations between the TONI-3 and broad measures of academic achievement ranged from .55 to .76, suggesting a moderate relationship. The authors also reported data from seven studies correlating the TONI and TONI-2, predecessors of the TONI-3, to 40 different measures of academic achievement. Results indicated average correlations ranging from .36 in the area of written language to .49 in the area of reading. Finally, adequate content validity was established through classical item analysis and differential item functioning analysis. These procedures were applied to the ESL subgroup with resulting coefficients of .98 on both forms. These coefficients are described as being very high and provide evidence that the TONI-3 contains little or no systematic bias towards ESL individuals. A recent study evaluated and compared the psychometric properties of several nonverbal intelligence tests and found the TONI-3 to be technically adequate and psychometrically sound (Athanasiou, 2000).

To date, no independent studies have been conducted examining the validity of the TONI-3 with Spanish speaking or ELL students. However, Coleman, Scribner, Johnsen, and Evans (1993) examined the performance of a sample of Mexican-American students with learning disabilities on the TONI-2, the predecessor to the TONI-3. Coleman et al. compared students' scores on the TONI-2 with scores on the Wechsler Adult Intelligence Scale - Revised (WAIS-R). Correlation coefficients for the TONI-2 were .41 with the Verbal IQ, .44 with the Performance IQ, and .50 with the Full Scale IQ. The Mexican-American sample earned a mean Full Scale IQ of 83.1 on the WAIS-R while earning a mean score of 86.8 on the TONI-2.

In general, the TONI-3 appears to be a technically sound measure of nonverbal

intelligence. It is brief and easy to administer. Its nonverbal design lends itself to use with deaf students, those from diverse linguistic and cultural backgrounds, and students with significant motor and language disabilities. Preliminary studies conducted by the TONI-3 authors suggest the TONI-3 is a reliable measure and demonstrates adequate content validity when used with children who speak English as a second language. The TONI-3 has several shortcomings, however. Because it correlates only moderately with the WISC-III, and due to its unidimensional nature, it is best used as a screening measure or as one component of a more comprehensive battery. The TONI-3 correlates only moderately with tests of academic achievement when used with the general population. Predictive validity studies have yet to be conducted with ELL populations. Further studies will need to be conducted before determining that the TONI-3 is an unbiased measure when used with Spanish speaking individuals.

Leiter International Performance Scale - Revised

The Leiter International Performance Scale is a comprehensive nonverbal measure that has been widely used with ELL students (Ochoa et al., 1996). The original Leiter International Performance Scale was developed for use with children in the U.S. in 1948 and was subsequently revised in 1997 as the Leiter International Performance Scale - Revised (Roid & Miller, 1997). The authors of the Leiter-R describe it as a measure of general intellectual ability, memory, and attention that can be effectively used with groups of children who cannot be accurately assessed with traditional intelligence tests. This includes students with communication disorders, cognitive delays, hearing problems, motor impairments, attention deficits, and English as a second language. The Leiter-R is considered a truly nonverbal measure in that instructions and responses do not require the use of language by the examinee or testee. One of the strengths of the Leiter-R is its wide age range, which spans ages 2.0 to 20.11 years.

The Leiter-R is based on a three-level hierarchical model of intelligence that recognizes a general intelligence or "g" factor, as well as fluid, crystalized, and visual factors (Roid et al., 1997). The authors note that the Leiter-R focuses on fluid, as opposed to crystallized abilities, as they are less dependent on academic background or cultural factors. The Leiter-R consists of 20 subtests that make up two separate batteries, Visualization and Reasoning (VR) and Attention and Memory (AM). In addition to a composite IQ, various VR and AM composite scores are provided.

The Leiter-R VR Battery was normed on 1,719 individuals, all ranging in age from 2.0 to 20.11. The AM Battery was standardized on a subset of 763 of the same children. Roid et al., (1997) explained that the AM Battery, with its smaller role as a diagnostic tool in the areas of inattentiveness and memory span, did not require as large a sample size as the VR Battery, from which the IQ scores are estimated. The standardization sample was stratified by parent occupation, geographic region, community size, age, gender, and ethnicity. Hispanics are slightly over-represented, with 12.8 percent and 12.6 percent included in the AM and VR samples, respectively, compared to 11.6 percent in the 1993 U.S. census.

Extensive studies of internal consistency and test-retest reliability are reported in the Leiter-R's test manual (Roid et al., 1997). Internal consistency reliabilities for the VR subtests range from .75 to .90 across the various age levels. Internal consistency reliability estimates for the AM subtests are generally lower, ranging from .67 to .87. The FSIQ score reliabilities are reported as .91 and .93 for age groups six through ten and eleven through twenty, respectively. Test-retest reliability coefficients are reported based on samples of 143 children on the VR Battery and 45 children on the AM Battery. In general, scores on the AM Battery subtests were less stable than those on the VR Battery. AM Battery subtest coefficients ranged from .55 to .85 while VR Battery

subtests coefficients ranged from .65 to .90. Likewise, composite score test-retest correlations were higher on the VR Battery (.86 to .96) than on the AM Battery (.61 to .85). Unfortunately, the interval between testings is not reported. Also, reliability estimates for subgroups, such as ELL children, are not provided.

The authors of the Leiter-R also provide a good deal of data to support its content and criterion-related validity (Roid et al., 1997). Test items were analyzed and examined by a panel of experts in the field. Those items with high indices of item bias or poor ratings by examiners and experts, were not included in the final version. Subtests were developed that reflected major nonverbal cognitive factors with high internal consistency. Rasch item analysis was utilized to examine item bias of both the VR and the AM Batteries. Results showed the various subtests to be generally free from differential item functioning between Caucasian and Hispanic samples. Comparisons between the normative group and various criterion groups, such as ESL-Hispanic children, were conducted. The median score for the ESL-Hispanic group on the Full IQ was reported to be 92.5, compared to 101 for the normative group. Various correlations between the Leiter-R scores and scores from other cognitive measures are reported. Correlations between the Leiter-R Full IQ and WISC-III scores were shown to be .83 (Processing Speed), .78 (Freedom from Distractability), .85 (Performance IQ), and .86 (Full Scale IQ). Correlations were also reported between the Leiter-R Full IQ and measures of academic achievement. Correlations ranged from .62 (WRAT-3 Arithmetic) to .82 (WJ-R Reading and Broad Mathematics). Correlations with other cognitive and academic achievement measures were not reported for special groups, such as ELL populations.

Two independent researchers have studied the validity of the Leiter-R with ELL students. Koehn (1999) examined the performance of a sample of 28 ESL Hispanic-

American children on the brief IQ of the Leiter-R and the WISC-III VIQ, PIQ, and FSIQ. The children's mean Leiter-R score was 93, while their VIQ, PIQ, and FSIQ were 80, 93, and 85, respectively. Surprisingly, the correlation between the Leiter-R and the VIQ was shown to be moderately high at .71. Koehn noted that this unexpected result may indicate that a global factor or "g" is common between the two test batteries. Another explanation may be that subvocal language or "self talk" was used to solve the items presented on the Leiter-R, and therefore, verbal ability influences the score. The correlation between the Leiter-R and the PIQ was .65, and the correlation with the FSIQ was .74. These correlations, although not as strong as those presented by the authors of the Leiter-R, provide additional support for the validity of the Leiter-R as a measure of cognitive ability. This study is somewhat limited by the small sample size of 28 children.

Cathers-Schiffman (2000) examined the concurrent and predictive validity of the Leiter-R, Cross Battery Assessment (CBA), and the WISC-III for Hispanic and Anglo students, matched by age, grade, and gender. The study controlled for English ability and socioeconomic status. Compared to the Verbal IQ, the Performance IQ of the WISC-III, the CBA Fluid Intelligence measures, and the Full Scale IQ of the Leiter-R were shown to measure cognitive ability equally well across cultural groups, unconfounded by language ability. As expected, Verbal IQ was found to be highly influenced by English language ability. English language ability and socioeconomic status, rather than ethnicity, explained much of the relationships between the measures of cognitive ability. Predictive validity of the measures was examined using the Metropolitan Achievement Tests - Seventh Edition (MAT-7). Of the three measures, the Leiter-R was shown to be the weakest predictor of achievement. The authors noted that this is likely due to the nonverbal nature of the Leiter-R, as opposed to the WISC-III

and CBA which contain verbal components. Ethnicity did not account for variance in academic achievement criterion measures, especially when English language ability and socioeconomic status were controlled. This study presents mixed results regarding the utility of the Leiter-R when used to with ELL students.

In summary, the Leiter-R should be considered a promising alternative to traditional measures of intelligence when assessing ELL children. Careful attention was paid to its development. Its strengths lie in its nonverbal and comprehensive nature, wide age range, and technical properties. Preliminary evidence suggests the Leiter-R is a non-biased measure of intelligence when used with Hispanic individuals. In general, research conducted by the test authors as well as independent researchers supports the content and predictive validity of the Leiter-R. As is the case with other nonverbal measures, a weakness of the Leiter-R is its somewhat weak correlation with academic achievement, compared to traditional measures of intelligence. However, the Leiter-R was shown to more accurately predict achievement than the TONI-3, a unidimensional nonverbal measure. Unfortunately, reliability data are not presented by the test authors for ELL individuals. Much more data needs to be obtained to further establish the validity of the Leiter with ELL students.

Universal Nonverbal Intelligence Test

The Universal Nonverbal Intelligence Test (Bracken et al., 1998) is a comprehensive, nonverbal measure of intelligence. It was designed to "measure fairly the general intelligence and cognitive abilities of children and adolescents...who may be disadvantaged by traditional verbal and language-loaded measures" (p.1). This includes those with language-related learning disabilities, psychiatric conditions, sensory limitations, and language impairments. Like the TONI-3 and Leiter-R, the UNIT is completely nonverbal and does not require the use of either receptive or expressive

language from the examiner or the examinee. This allows Spanish speaking ELL students to perform without the interference of language issues.

As a comprehensive intelligence test, the UNIT was designed to measure both general intelligence as well as the underlying factors of memory, reasoning, symbolic skills, and nonsymbolic abilities (Bracken et al., 1998). While memory and reasoning are considered "primary abilities" by the UNIT authors, the Symbolic and Nonsymbolic scales are considered as "secondary" measures as they represent the inferred processes that facilitate task solution. The three subtests that are considered to be amenable to verbal mediation make up the Symbolic Scale while the three subtests that are not as amenable to verbal labeling comprise the Nonsymbolic Scale. The UNIT authors noted that the symbolic mediation adds an important verbal component to the nonverbal tasks, thereby increasing the power of the nonverbal tests to predict academic achievement.

Normative data on the UNIT were collected in 1996 on a sample of 2,100 children and teens (Bracken et al., 1998). The standardization sample was constructed to closely match U.S. census data regarding gender, race/ethnicity, Hispanic origin, geographic region, urban/rural residence, and parents' education level. A commendable feature of the UNIT is its inclusion of special populations in the normative sample to ensure representation of individuals for whom the test was intended. This includes those with learning disabilities, speech and language delays, emotional disturbance, hearing impairments, giftedness, bilingual education, and English as a second language. Although 1.8% of the students in the sample were bilingual and 2.0% were designated as LEP, these percentages are somewhat lower than the 3.1% and 4.0%, respectively, reported in the U.S. census data.

The authors provide a good deal of evidence to support the reliability of the UNIT. Internal consistency reliability estimates for the full scale score range from .84

to .95 for the Abbreviated Battery, Standard Battery, and Extended Battery (Bracken et al., 1998). Compared to older children, reliability estimates for younger children tend to be somewhat lower. Reliability figures for the various scale scores also tended to fall in the high .80s to low .90s. Coefficient alphas for the six subtests are reported in the UNIT manual as follows: Symbolic Memory .85, Cube Design .91, Spatial Memory .81, Analogic Reasoning .79, Object Memory .76, and Mazes .64. It should be noted that the Mazes subtest is not included in the standard battery. Test-retest reliability was evaluated over an interval of between 3 and 42 days with 197 participants ages 5 through 17. Coefficients for the Standard and Extended Battery Full Scale Score were equal to or greater than .85. An exception was the group of children ages 5-7, who showed a coefficient of .78 for the Extended Battery. Reliability data are not provided for special groups such as Hispanics and ELL individuals.

A wide range of validity data is presented in the UNIT manual (Bracken et al., 1998). Several concurrent validity studies were completed with traditional comprehensive intelligence measures such as the WISC-III and WJ-R, as well unidimensional nonverbal intelligence tests such as the Matrix Analogies Test and TONI-2. Full scale correlations with the comprehensive measures fell within the .83 to .88 range, with nonsignificant mean score differences between the UNIT and the criterion tests. In contrast, correlations with the unidimensional nonverbal tests were between .56 and .83. The low correlations with the unidimensional nonverbal tests may be due to the limited scope of intelligence assessed by these measures (Bracken et al., 2001).

The UNIT was also correlated with the Bateria-R using two samples of native Spanish speaking students (Bracken et al., 1998). One sample included 27 students in bilingual education classes while the other consisted of 26 students receiving services

for English as a second language (ESL). The bilingual education students' English proficiency was limited while the ESL students' was high. The resulting coefficients indicated little overlap between the Bateria-R and the UNIT. Correlations between the full scale scores was .39 for the bilingual education group and .17 for the ESL group. The authors noted that the Bateria-R scores for these groups was systematically and considerably lower than the UNIT scores. Mean scores from the Bateria-R Broad Cognitive Ability Early Developmental scale were 77 for the bilingual sample and 69 for the ESL sample. In contrast, the mean UNIT full scale scores on the Extended Battery were 93.41 and 96.88 for the bilingual and ESL sample, respectively. The authors noted that the stronger English language skills of the ESL group may have interfered with their performance on the Bateria-R as it was developed with monolingual or nearly monolingual Spanish-speaking examinees. Results of this study do not provide conclusive data as to the validity of the UNIT with Spanish speaking students.

The authors of the UNIT also present evidence that the UNIT adequately predicts academic achievement. The UNIT FSIQ correlated .62 with the WIAT (The Psychological Corporation, 1992) Total Composite Score. Correlations of the UNIT FSIQ with Basic Reading, Mathematics Reasoning, Language, and Writing were .70, .71, .48, and .55, respectively. Another study was reported in which the UNIT was examined in relation to achievement in Spanish as measured by the Broad Reading, Basic Reading Skills, and Reading Comprehension scales of the WLPB-R (Woodcock, 1991). The resulting correlations tended to be low, ranging from -.03 to .07 with an ESL sample and .12 to .39 with a bilingual sample. In contrast, correlations with the WLPB-R and the Bateria-R ranged from .28 to .91 with the same samples.

Other validity evidence presented by the UNIT authors include factor analyses,

discriminant validity studies, and item bias analysis. In general, results supported the internal structure of the UNIT and indicated it is not biased against any specific population. Comparison of the performance of whites (non-Hispanics) versus Hispanics and whites versus bilingual and ESL children did not show significant differences. The mean Extended Battery FSIQ score for the Hispanic group was 99.41, compared to 100.85 for a demographically matched non-Hispanic group. The FSIQ score for the bilingual/ESL group was 93.30, compared to 97.03 for the English-speaking comparison sample.

The UNIT is a test that shows promise as a measure of nonverbal intelligence. Unlike other nonverbal measures that measure a narrow range of abilities, the UNIT is more comprehensive in nature. Standardization appears well done and the results of a number of reliability and validity studies are impressive. The fact that bilingual and English as a second language students were included in the standardization sample is commendable and is a practice that all test developers should consider. Although the UNIT is considered comprehensive in nature, practitioners are encouraged to use the UNIT in combination with verbal measures to obtain a more accurate picture of the student's overall cognitive functioning. Caution is encouraged when interpreting results for children ages 5 to 7 because of concerns about reliability at these ages. Although the UNIT appears to have strong internal or content validity, it may be biased when used to predict achievement among Spanish speaking children. Research has failed to demonstrate a strong relationship between ELL individuals' academic achievement and performance on the UNIT. Future studies to further establish the utility of the UNIT with ELL students should be conducted.

Alternative Cognitive Assessment Methods

Although a number of promising traditional measures of cognitive ability have been developed over the last decade, due to the traditional limitations of such measures several researchers have proposed alternative methods. These include the dynamic approaches such as the Learning Potential Assessment Devise (LPAD) and cross battery assessment.

Dynamic Assessment

Proponents of dynamic assessment argue that children from racially and ethnically diverse backgrounds have not had learning experiences comparable to their mainstream peers and consequently perform poorly on IQ measures that assume equivalent experiences (Samuda, 1998). The guiding principle of dynamic assessment is that in order to understand how a child learns, you need to engage the child in the learning process (Lidz, 1997). Dynamic assessment, therefore, sets up a situation in which the student engages in the learning process and the examiner actively attempts to facilitate the student's cognitive competence. It most often takes place in a test-intervene-retest format. The intent is to gain an understanding of how to facilitate the learning of the child, instead of focusing on the child's demonstration of ability.

One of the best known methods of dynamic assessment is Feuerstein's Learning Potential Assessment Device or LPAD (Samuda, 1998). The LPAD consists of fifteen tests for individual administration and nine for group administration. The testing instruments facilitate a series of testing-in-the-act-of-learning procedures (Gopaul-McNicol et al., 2002). The task of the examiner is to observe the examinee's response to tasks and use this information to elicit positive changes in the performance of the examinee. Unlike traditional measures, the LPAD does not include norms. A strength

of dynamic assessment is its ability to provide information about the student's learning needs that can be linked with instruction. The LPAD may be particularly useful with ELL students who come from different educational and cultural backgrounds than their mainstream counterparts. Haywood, Brown, and Wingenfeld (1990) state that the LPAD hold promise as nondiscriminatory assessment methods because it is capable of distinguishing between lack of knowledge and lack of ability to acquire knowledge. Unlike traditional norm-referenced measures, it does not assume ELL students have had similar opportunities to learn as mainstream students. Unfortunately, dynamic assessment procedures such as the LPAD have not been systematically employed or researched (Lopez, 1995; Rogers, 1998). Consequently, dynamic approaches lack empirical evidence supporting their validity.

Cross Battery Assessment

Another method that has been proposed for the assessment of ELL students is the cross-battery approach to psychoeducational assessment (Ortiz, 2004; Saffron, 2000). Cross battery assessment (CBA) involves a systematic approach to selecting and interpreting subtests from major cognitive batteries. Measures are selected depending on the characteristics of the examinee and the questions that the assessment attempts to answer. Ortiz (2004) presents a cross-battery approach that involves examining the relative influence of language and culture on test performance through the use of a matrix. Current tests of intelligence are classified according to the degree to which they require expressive or receptive language skills (linguistic demand), and the degree to which a particular test requires familiarity, specific knowledge, or understanding of U.S. mainstream culture (cultural loading). Next, tests that are considered less culturally and linguistically loaded can be selected and administered. Knowing the degree to which a particular measure is affected by cultural and linguistic factors guides interpretation of

the student's performance. On the WISC-IV, for example, the Matrix Reasoning, Cancellation, Block Design, Symbol Search, Digit Span, and Coding subtests can be selected based on low degrees of linguistic demand and cultural loading. In contrast, if the Information, Similarities, Vocabulary, Comprehension, and Word Reasoning subtests are given, the evaluator can interpret results of these subtests taking into consideration the high degree of cultural loading and linguistic demand inherent in these tests. Using tests that are less linguistically and culturally loaded with ELL students places practitioners in a position to better defend the validity of conclusions and inferences drawn from the obtained data.

Cathers-Schiffman (2000) conducted a study to examine the validity of CBA as a measure of intelligence for Anglo and Hispanic Spanish/English speaking children. Select subtests from the Leiter-R, WISC-III, and the Woodcock Johnson Tests of Cognitive Ability-Revised (WJTCA-R) were utilized. The CBA method was shown to account for more variance in the criterion variable, academic achievement test scores, than performance on the Leiter-R. That is, CBA was shown to be a more accurate predictor of academic achievement. CBA and the WISC-III were shown to be comparable predictors of academic achievement.

Although cross-battery assessment is a promising alternative to traditional approaches, this method is in its relative infancy. A good deal of research will need to be conducted to establish the method's utility with ELL students.

SECTION 5

SUMMARY AND CONCLUSIONS

The purpose of this paper is to provide school psychologists and other professionals with information necessary to conduct cognitive assessments of ELL Spanish speaking children in an empirically sound, nonbiased, defensible, and practical manner. Ethical standards, legal findings, various assessment practices, and specific assessment measures were examined. Based on a review of best practice literature pertaining to the psychoeducational assessment of ELL students, as well as ethical and legal guidelines, a number of recommendations regarding assessment practices with ELL students are warranted.

First of all, important intervention and placement decisions should not be based on results of a single test, including cognitive measures. Rather, abilities in multiple areas should be evaluated using multiple methods. Practitioners should utilize a wide variety of information sources to obtain a full history, description, and explanation of the child's current functioning across settings. Data gained through a review of records, observations, and interviews should be carefully collected and considered. In addition, assessment in the areas of language proficiency, academic assessment, and acculturation should be used to provide essential information. Results of cognitive measures need to be interpreted taking into account the various cultural, linguistic, and environmental factors that may have an effect on the student's learning.

Second, best practice calls for cognitive assessment in both the child's primary language and English. At the very least, language proficiency assessment should be conducted in both languages in order to ascertain in what language or languages cognitive testing should be given. At that point, assessment may continue in either the student's dominant language or in both languages in the case of bilingual students.

Nonverbal assessment can also be conducted. A variety of promising measures of cognitive ability are now available to practitioners for use with Spanish speaking children.

Various methods of assessing cognitive ability that have been widely used in the past appear to have limited validity with ELL students. Historically, school psychologists in the U.S. who work with Spanish speaking children have been limited in their choice of tools to assess cognitive ability. Many have simply administered the same measures they use with English speaking students. Others have relied solely on nonverbal measures, tests normed outside of the U.S., or informal, on-the-spot translations. These practices all have questionable validity with ELL students. In addition, the use of an interpreter during assessment should be considered only as a last resort in cases where a bilingual assessor is not available. This is due to the questionable validity of test scores obtained via measures translated by an interpreter.

Finally, when selecting norm-referenced measures, tests should be carefully examined to ensure that they are appropriate for the individual test taker. Tests should be normed on a sample that matches the characteristics of the child, and the reliability and validity of the measure should be well documented. Fortunately, several measures of cognitive ability have been developed for use with ELL and Spanish speaking children over the last decade. Recently developed measures address many of the shortcomings of previous assessment tools and show promise as valid and reliable measures.

As full-scale or broad measures of intelligence, the WISC-IV and WJ III Tests of Cognitive Abilities have been translated and adapted into Spanish as the WISC-IV Spanish and the Bateria III Woodcock-Munoz: Pruebas de Habilidades Cognitivas. Both appear to be well-developed measures and benefit from their association with the

English versions, which are widely esteemed in the field of intellectual assessment. Both include norms that have been calibrated or equated to the norms on the English versions. As both of these scales were recently released, outside studies have yet to be conducted examining their psychometric properties. Authors of the WISC-IV Spanish present much more data regarding reliability and validity than the authors of the Bateria III COG, suggesting the WISC-IV may be a more appropriate measure at this time. The inclusion of children from a variety of countries outside the U.S. in the Bateria III COG calibration sample may limit its use to children who have recently immigrated to the U.S. Similarly, because the WISC-IV Spanish was standardized using a sample of Spanish-dominant students who had five years or less of education in the U.S., it should only be used with students from similar backgrounds. A strength of the WISC-IV Spanish is the opportunity it provides examiners to not only compare the child's performance relative to English-speaking children in the U.S. population but also to Spanish-speaking children in the U.S. who are similar in terms of U.S. educational experience and parent educational level. A serious shortcoming of both the WISC-IV Spanish and the Bateria III COG is the lack of predictive validity studies examining their relationship with measures of achievement.

Another cognitive measure available in English and Spanish is the Bilingual Verbal Ability Tests. The BVAT is currently the only measure available designed to measure the combined bilingual verbal ability of children in a variety of languages. Reliability and validity of the measure appear to be adequate, though further studies should be completed to establish its validity for different languages and levels of language proficiency. The BVAT appears to be a valid measure of verbal ability and seems to correlate with academic achievement. Unfortunately, although the BVAT has been in print since 1998, few independent studies have examined its use with Spanish

speaking children. Other weaknesses of the BVAT include its outdated norms and possible bias of the Picture Vocabulary test with Spanish dominant student. Scores on the Picture Vocabulary test may underestimate Spanish dominant students' true abilities.

The use of nonverbal measures of intelligence in the assessment of ELL students continues to be recommended, especially with children with language impairments or severe motor deficits. Three nonverbal measures that demonstrate utility with ELL students are the TONI-3, the Leiter-R, and the UNIT. A commendable feature of all three measures is their inclusion of students who speak English as a second language in their standardization samples. All three measures demonstrate reliability and validity supporting their use with Hispanic students. Very little data are available, however, to support their use specifically with Spanish speaking individuals. In fact, there are data to suggest that nonverbal measures are even less accurate in predicting the achievement of Spanish speaking children than English speakers. Because of its unidimensional nature, as well as its moderate correlations with full-scale intelligence tests, the TONI-3 seems most appropriately utilized as a screening measure or as a supplementary scale used with a battery of cognitive measures. The UNIT and Leiter-R have the advantage of being more comprehensive in nature, measuring a broader range of cognitive abilities.

Although the assessment recommendations provided in this paper represent what may be considered best practice, practical experience suggests that they may not always be feasible. Often, members of assessment teams find that they are expected to complete a high number of psychoeducational assessments in a limited time frame. This problem is compounded when they work with populations with a high number of Spanish speaking students as bilingual assessments tend to take more time than a typical evaluation. It is often helpful to enlist the assistance of personnel that may not normally

be as involved in the assessment process. Para-professionals or teachers, for example, can be enlisted to conduct parent interviews or classroom observations and complete rating scales or record reviews. Some school districts find it helpful to train bilingual para-professionals specifically to perform language proficiency or academic testing. Assessment teams may rely heavily on data collected from parents and teachers at a student success team (SST) or at-risk meeting. This includes information regarding health and developmental history, language proficiency, academic history, acculturation, response to intervention, classroom functioning, and home environment. Regarding cognitive assessment measures, it may be beneficial to begin with the BVAT as it provides a measure of oral English proficiency as well as bilingual verbal ability. If the student has been in the U.S. for five or fewer years and appears to have adequately developed Spanish skills, a comprehensive measure such as the WISC-IV Spanish could then be administered. If the student has attended school in the U.S. for more than five years or does not appear to have adequately developed Spanish skills a nonverbal measure such as the UNIT could be given. To confirm the presence of a psychological processing disorder, other measures, such as select subtests of the Bateria III COG, may then be administered. Although the assessment of Spanish speaking individuals tends to take more time and effort than a typical evaluation, with the assistance of team members and a well-developed pre-referral system, a comprehensive evaluation can be done in a timely manner.

Care should be taken in evaluating cognitive assessment data in light of factors such as language proficiency and acculturation. Caution should be taken in using a discrepancy model to identify a specific learning disability with ELL students as linguistic factors are likely to result in a discrepancy between ability and achievement. This is especially true for ELL students in the early grades or for those who have

recently moved to the U.S. For example, it is not unusual for a Spanish dominant student who has limited English proficiency to demonstrate academic skills that are well below his/her measured cognitive ability, especially as measured by nonverbal measures. It would be erroneous to automatically assume the discrepancy is due to a learning disability instead of linguistic factors.

As many experts in the field have noted, much research needs to be done in order to establish best practices in the assessment of ELL students. Future research should continue to focus on establishing sound and non-biased assessment methods with those of diverse linguistic and cultural backgrounds. In addition, researchers need to more closely examine the psychometric properties and the differential item functioning of the most commonly used instruments. The effect of bilingualism and language proficiency on students' performance on traditional measures should be studied more closely. Finally, more research needs to be conducted in the area of alternative assessment.

Although many unanswered questions remain concerning the best assessment practices with ELL students, this paper has outlined several guidelines that may minimize bias during assessment activities. It is hoped that by becoming more sensitive to the special considerations that must be given to ELL Spanish speaking children in the evaluation process, professionals will conduct more accurate and meaningful assessments. This, in turn, will hopefully lead to better educational planning and outcomes for Spanish speaking students.

REFERENCES

Alvarado, C. G. (2000). A theoretical and empirical study of the English/Spanish Bilingual Verbal Ability Tests. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 60,* p. 2336.

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

American Psychological Association. (1993). Guidelines for providers of psychological services to ethnic, linguistic, and culturally diverse populations. *American Psychologist, 48,* 45-48.

American Psychological Association. (2002). *Ethical principles of psychologists and code of conduct.* Washington, DC: Author.

American Psychological Association. (2003). Guidelines on multicultural education, training, research, practice, and organizational change for psychologists. *American Psychologist, 58,* 377-402.

Armour-Thomas, E. (1992). Intellectual assessment of children from culturally diverse backgrounds. *School Psychology Review, 21,* 552-565.

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

Atlas, J. A. (2001). Review of the Test of Nonverbal Intelligence. In B. Plake & B. Impara (Eds.), *The fourteenth mental measurements yearbook.* Lincoln, NE: Buros Institute of Mental Measurements.

Baker, S. K. & Good, R. (1995). Curriculum-based measurement of English reading with bilingual Hispanic students: A validation study with second-grade students.

School Psychology Review, 24, 561-578.

Baldizon-de-Naclerio, M. (1999). Language proficiency and stress: Impact on measured intelligence and anxiety for Latino children. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 60,* 357-477.

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

Bracken, B. A. & McCallum, R. S. (2001). Assessing intelligence in a population that speaks more than two hundred languages: A nonverbal solution. In L. A. Suzuki, J. G. Ponterotto, & P. J. Miller (Eds.), *Handbook of Multicultural Assessment - 2nd Edition* (pp. 405-431). San Francisco: Jossey-Bass.

Brigance, A. H., & Messer, P. (1984). *Brigance Diagnostic Assessment of Basic skills: Spanish edition.* North Billerica, MA: Curriculum Associates.

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

Brown, L., Sherbenou, R. J., & Johnsen, S. K. (1997). *Test of Nonverbal Intelligence - Third Edition.* Austin, TX: Pro-Ed.

Buitrago, R. (1999). Concurrent validity between a simultaneously translated Spanish version of the Wechsler Intelligence Scale for Children - Third Edition (WISC-III) and the Universal Nonverbal Intelligence Test (UNIT) with a monolingual Spanish-Speaking sample. *Dissertation Abstracts International: Section A:Humanities and Social Sciences, 60,* p. 1427.

Camarota, S. A. & McArdle, N. (2003). Where immigrants live: An examination of state residency of the foreign born by country of origin in 1990 and 2000. Retrieved December 21, 2005, from *www.cis.org/articles/2003/back1203.html#table1*

Cathers-Schiffman, T. A. (2000). A validation of three measures of intelligence for Hispanic Spanish/English-speaking and Anglo English-only speaking children.

Dissertation Abstracts International: Section A: Humanities and Social Sciences, 61, p. 4744.

Chamberlain, P. & Medinos-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. 112-156). Austin, TX: Pro Ed.

Chinn, P.C., & Hughes, S. (1987). Representation of minority students in special education classes. *Remedial and Special Education, 8,* 41-46.

Chun, K. M., Organista, P. B., & Marin, G. (2003). *Acculturation: Advances in theory, measurement, and applied research.* Washington DC: American Psychological Association.

Civil Rights Act of 1964. Title VI, 42 U.S.C., 200d et seq.

Civil Rights Project. (2000). *Executive summary: Conference of minority issues in special education.* Retrieved June 25, 2005, from www.law.harvard.edu/civilrights/conferences/SpecEd/exsummary.html.

Cizek, G. J. (2003). Review of the Woodcock-Johnson III. In B. Plake, J. Impara, & R. Spies (Eds.), *The fifteenth mental measurements yearbook.* Lincoln, NE: The Buros Institute of Mental Health.

Coleman, M., Paredes, S., Sohnsen, S., Evans, M. K. (1993). A comparison between the Wechsler Adult Intelligence Scale-Revised and the Test of Nonverbal Intelligence-2 with Mexican-American secondary students. *Journal of Psychoeducational Assessment, 11,* 250-258.

Cuellar, I., Arnold, B., & Maldonado, R. (1995). Acculturation rating scale for Mexican Americans-II: A revision of the original ARSMA scale. *Hispanic Journal of Behavioral Sciences, 17,* 275-304.

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

Dean, R. S. (1977). Reliability of the WISC-R with Mexican-American children. *Journal of School Psychology, 15,* 267-268.

Diana v. California. No. C-70-37 (N.D. Cal. 1970)

Dicerbo, K. E. (2003). English language proficiency and tests of intelligence and academic achievement. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 64,* p. 793.

Education for All Handicapped Children Act of 1975 (P.L. No. 94-142. 20) U.S.C. Sec. 401.

Education of the Handicapped Act Amendments of 1986. (Pub. L. No. 99-457).

Elliot, C. D. (1990). *Differential Ability Scales.* San Antonio, TX: The Psychological Corporation.

Esquivel, G. B. (1988). 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). Washington, DC: National Association of School Psychologists.

Figueroa, R. A. (1990a). Assessment of linguistic minority group children. In C. R. Reynolds & R. W. Kamphaus (Eds), *Handbook of psychological and educational assessment of children: Intelligence and achievement* (pp. 671-696). New York: The Guilford Press.

Figueroa, R. A. (1990b). 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.

Flanagan, D. P. & Ortiz, S. O. (2002). Best practices in intellectual assessment: Future

directions. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1351-1372). Bethesda, MD: National Association of School Psychologists.

Flanagan, D. P. & Kaufman, A. S. (2004). *Essentials of WISC-IV Assessment.* New Jersey: John Wiley & Sons, Inc.

Gass, C. S., Demsky, Y. I., & Martin, P. C. (1998). Factor Analysis of the WISC-R (Spanish Version) at 11 age levels between 6 ½ and 16 ½ years. *Journal of Clinical Psychology, 54,* 109-113.

Gopaul-McNicol, S. & Armour-Thomas, E. (2002). *Assessment and culture.* San Diego: Academic Press.

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

Gutkin, T. B., & Reynolds, C. R. (1980). Factorial similarity of the WISC-R for Anglos and Chicanos referred for psychological services. *Journal of School Psychology, 18,* 34-39.

Haywood, H. C., Brown, A. L., & Wingenfeld, S. (1990). Dynamic approaches to psychoeducational assessment. *School Psychology Review, 19,* 411-422.

Holtzman, W. H. & 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. 248-280). Austin: Pro Ed.

Individuals with Disabilities Education Act of 1990 (Public Law No. 101-476), 20 U.S.C. Sec. 1400

Individuals with Disabilities Education Act Revision of 1997 (Public Law No. 105-17), 20 U.S.C. Sec. 1400

Individuals with Disabilities Education Improvement Act of 2004 (Public Law No. 108-446), 20 U.S.C. Sec. 1400.

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

Jimenez, S. (2002). An analysis of the reliability and validity of the Universal Nonverbal Intelligence Test (UNIT) with Puerto Rican children. *Dissertation Abstracts International: Section B: The Sciences & Engineering, 62,* pp. 524.

Johnson, D. L. & MCGowan, R. J. (1984). Comparison of three intelligence tests as predictors of academic achievement and classroom behaviors of Mexican-American children. *Journal of Psychoeducational Assessment,* 2, pp. 345-352.

Joint Committee on Testing Practices. (2004). *Code of Fair Testing Practices in Education.* Washington, DC.

Kaufman, A. S. (1994). *Intelligence testing with the WISC-III.* New York: John Wiley & Sons, Inc.

Kindler, A. L. (2002). *Survey of the states' limited English proficient students and available educational programs and services 1999 - 2000 summary report.* Washington, DC: National Clearinghouse for English Acquisition and Language Instruction Educational Programs.

Knoff, H. M. & Dean, K. R. (1994). Curriculum-based measurement of at-risk students' reading skills: A preliminary investigation of bias. *Psychological Reports,* 75, 1355-1360.

Koehn, R. D. (1999). WISC-III and Leiter-R assessments of intellectual abilities in Hispanic-American children with English as a second language. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 59,* p. 3350.

Larry P. v. Wilson Riles. No. C-71 2270 RFP, U.S. District Court, Northern District of California (1972).

Lau v. Nichols. 414 U.S. 563. (1974).

Lawless, G. F., Stedman, J. M., & Cortner, R. H. (1980). Factor analysis of the WISC-

R for a sample of bilingual Mexican-Americans. *Journal of Clinical Child Psychology,* 57-58.

Lidz, C. S. (1997). Dynamic assessment approaches. In D. Flanagan, J. Genshaft, & P. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues.* NY: The Guilford Press.

Loe, S. A. (2001). An examination of family oriented practice and cultural diversity in school psychology: A national survey of school psychology practitioners. *Dissertation Abstracts International Section A: Humanities & Social Sciences, 61,* p. 4298.

Lopez, E. (1994). *A preliminary investigation of errors made by interpreters during on the spot translations of WISC-R questions.* Paper presented at the National Association of School Psychologists Conference, Washington, DC.

Lopez, E. C. (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. C. (1997). The cognitive assessment of limited English proficient and bilingual children. In D. P. Flanagan, J. L. Genshaft, & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (pp. 503-516). New York: The Guilford Press.

Lopez, E. C. (2002). Best practices in working with school interpreters to deliver psychological services to children and families. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp1321-1335). Bethesda, MD. National Association of School Psychologists.

McCreith, T. M. (2005). A construct comparability analysis of cognitive ability tests in different languages. *Dissertation Abstracts International Section A: Humanities &*

Social Sciences, 65, p. 2962.

Macias, R. F. (1998). *Summary report of the survey of the states' limited English proficient students and available educational programs and services 1996-1997.* Washington, DC: National Clearinghouse for Bilingual Education.

Marin, G. & Gamba, R. J. (1995). A new measurement of acculturation for Hispanics: The Bidimensional Acculturation Scale for Hispanics (BAS). *Hispanic Journal of Behavioral Science, 18,* 297-316.

McShane, D. & Cook, V. J. (1985). Transcultural intellectual assessment. In B. B. Wolman (Ed.), *Handbook of intelligence: Theories, measurements and applications* (pp. 737-785)

Mishra, S. P. (1983). Validity of WISC-R IQs and factor scores in predicting achievement for Mexican-American Children. *Psychology in the Schools, 20,* 442-444.

Munoz-Sandoval, A. F., Cummins, J., Alvadado, C. G., & Ruef, M. L. (1998a). *Bilingual Verbal Ability Tests.* Itasca, IL: Riverside Publishing.

Munoz-Sandoval, A. F., Cummins, J., Alvarado, C. G., & Ruef, M. L. (1998b). *Bilingual Verbal Ability Tests, Comprehensive Manual.* Itasca, IL: Riverside Publishing.

Munoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005a). *Bateria III Woodcock-Munoz: Pruebas de Aprovechamiento.* Itasca, IL: Riverside Publishing.

Munoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2005b). *Bateria III Woodcock-Munoz: Pruebas de Habilidades Cognitivas.* Itasca, IL: Riverside Publishing.

National Association of School Psychologists. (2000). *Professional conduct manual.*

Bethesda, MD: Author.

National Research Council. (2002). *Minority students in special and gifted education.* Washington, DC: National Academy Press.

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.

Ochoa, S. H., Rivera, B., & Ford, L. (1997). An investigation of school psychology training pertaining to bilingual psycho-educational assessment of primarily Hispanic students: Twenty-five years after Diana v. California. *Journal of School Psychology*, 35, 329-349.

Ochoa, S. H. & Ortiz, S. O. (2005a). Conceptual measurement and methodological issues in cognitive assessment of culturally and linguistically diverse individuals. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students* (pp. 153-167). New York: The Guilford Press.

Ochoa, S. H. & Ortiz, S. O. (2005b). Language proficiency assessment: the foundation for psychoeducational assessment of second-language learners. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students* (pp. 137-152). New York: The Guilford Press.

Olivarez, A., Palmer, D. J., & Guillemard, L. (1992). Predictive bias with referred and nonreferred black, Hispanic, and white pupils. *Learning Disability Quarterly, 15,* 175-186.

Ortiz, S. O. (2002). Best Practices in nondiscriminatory assessment. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp1321-1335). Bethesda, MD. National Association of School Psychologists.

Ortiz, A. A. & Yates, J. R. (2002). Considerations in the assessment of English

language learners referred to special education. In A. J. Artiles & A. O. Ortiz (Eds.), *English language learners with special education needs* (pp. 73-93). Washington, DC: Center for Applied Linguistics.

Ortiz, S. O. (2004). Bilingual-multicultural assessment with the WISC-IV. In D. P. Flanagan & A. S. Kaufman, *Essentials of WISC-IV Assessment* (pp. 245-254).

Ortiz, S. O. (2005). Acculturational factors in psychoeducational assessment. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing culturally and linguistically diverse students* (pp. 124-136). New York: The Guilford Press.

Palmer, D. G., Olivarez, Jr., Willson, V. L., & Fordyce, T. (1989). Ethnicity and language dominance - influence on the prediction of achievement based on intelligence test scores in nonreferred and referred samples. *Learning Disability Quarterly, 12,* 261-274.

Plata, Maximino. (1993). Using Spanish-speaking interpreters in special education. *Remedial & Special Education,* 14, 19-27.

The Psychological Corporation. (1992). *Wechsler Individual Achievement Test: Manual.* San Antonio, TX: Author.

Reynolds, C. R., & Kaiser, S. M. (1990). Test bias in psychological assessment. In T. B. Gutkin & C. R. Reynolds (Eds.), *The handbook of school psychology - Second Edition* (pp. 487-525). New York: John Wiley & Sons.

Reynolds, C. R., Lowe, P. A., Saenz, A. L. (1999). The problem of bias in psychological assessment. In C.R. Reynolds & T. B Gutkin (Eds.), *The handbook of school psychology - Third edition* (pp. 549-594). New York: John Wiley & Sons.

Rhodes, R. L. (2005a). Assessment of academic achievement. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing Culturally and Linguistically Diverse*

Students (pp. 103-123). New York: The Guilford Press.

Rhodes, R. L. (2005b). The interview process. In R. L. Rhodes, S. H. Ochoa, & S. O. Ortiz (Eds.), *Assessing Culturally and Linguistically Diverse Students* (pp. 103-123). New York: The Guilford Press.

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

Rogers, M. R., Ponterotto, J. Gl, Conoley, J. C., & Weise, M. J. (1992). Multicultural training in school psychology: A national survey. *School Psychology Review*, 21, 603-616.

Rogers, M. R. (1998). Psychoeducational assessment of culturally and linguistically diverse children and youth. In H. B. Vance (Ed.), *Psychological assessment of children* (2nd Ed., pp. 355-384). New York: John Wiley and Sons , Inc.

Saffron, Y. M. C. (2000). The prediction of reading achievement by Ga and Gc cognitive abilities for Spanish-speaking students. *Dissertation Abstracts International: Section A:Humanities and Social Sciences, 61,* p. 498.

Sanchez-Boyce, M. (2000). *The use of interpreters in special education assessments.* Unpublished doctoral dissertation, University of California at Davis .

Sandoval, J. (2003). Review of the Woodcock-Johnson III. In B. Plake, J. Impara, & R. Spies (Eds.), *The fifteenth mental measurements yearbook.* Lincoln, NE: The Buros Institute of Mental Health.

Samuda, R. J. (1998). *Psychological testing of American minorities: Issues and consequences* (2nd ed.). Thousand Oaks: SAGE Publications Inc.

Scribner, A. P. (2002). Best assessment and intervention practices with second language learners. In A. Thomas & J. Grimes (Eds.), *Best practices in school psychology IV* (pp. 1485-1499). Bethesda, MD. National Association of School

Psychologists.

Shinn, M. R., Collins, V. L., Gallagher, S. (1998). Curriculum-based measurement and its use in a problem-solving model with students from minority backgrounds. In M. R. Shinn (Ed.), *Advanced applications of curriculum-based measurement* (pp. 143-173). New York: The Guilford Press.

U.S. Bureau of the Census. (2000). *Population by race and Hispanic origin for the United States: 2000.* Washington DC: Author.

U.S. Bureau of the Census. (2000). *Language spoken at home and ability to speak English for United States, regions and states: 2000.* Washington DC: Author.

U.S. Bureau of the Census. (2004). *U.S. interim projections by age, sex, race, and Hispanic origin.* Washington DC: Author.

U.S. Department of Education. (1994). *Summary of the bilingual education state educational agency program survey of states' limited English proficient persons and available educational services 1992-1993): Final report.* Arlington, VA: Development Associates.

Wechsler, D. (1949). *Wechsler intelligence scale for children.* New York: The Psychological Corporation.

Wechsler, D. (1974). *Wechsler intelligence scale for children-revised.* New York: The Psychological Corporation.

Wechsler, D. (1982). *Escala de inteligencia de Wechsler para ninos-revisada.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1984). WISC-RM *Escala de inteligencia Wechsler-revisada para el nivel escolar.* Mexico City: El Manual Moderno.

Wechsler, D. (1991). *Wechsler intelligence scale for children-third edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (1992). *Escala de inteligencia Wechsler para ninos-revisada de Puerto Rico.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2003). *Wechsler intelligence scale for children-fourth edition.* San Antonio, TX: The Psychological Corporation.

Wechsler, D. (2005). *Wechsler intelligence scale for children-fourth edition, Spanish.* San Antonio, TX: The Psychological Corporation.

Weiss, L. G. & Prifitera, A. (1995). An evaluation of differential prediction of WIAT achievement scores from WISC-III FSIQ across ethnic and gender groups. *Journal of School Psychology, 33*(4), 297-304.

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

Woodcock, R. W. (1982). *Bateria woodcock Psico-educativa en Espanol.* Itasca, IL: Riverside Publishing.

Woodcock, R. W. (1991). *Woodcock Language Proficiency Battery - Revised: English and Spanish Forms.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., & Munoz-Sandoval, A. F. (1996). *Bateria Woodcock-Munoz: Pruebas de Habilidad Cognitivas – Revisada.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001a). *Woodcock-Johnson III Tests of Achievement.* Itasca, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001b). *Woodcock-Johnson III Tests of Cognitive Abilities.* Itasca, IL: Riverside Publishing.