EXAMINING THE CULTURAL LOADING AND LINGUISTIC DEMAND OF THREE NEUROPSYCHOLOGICAL ASSESSMENT BATTERIES FOR CHILDREN IN A MIXED CLINICAL POPULATION

A DISSERTATION

SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS

FOR THE DEGREE OF DOCTOR OF PHILOSOPHY

IN THE GRADUATE SCHOOL OF THE

TEXAS WOMAN'S UNIVERSITY

DEPARTMENT OF PSYCHOLOGY AND PHILOSOPHY

COLLEGE OF ARTS AND SCIENCES

BY

PAMELA E. CIOFFI, B.A.

DENTON, TEXAS

DECEMBER 2015

DEDICATION

For my loving family, friends, and for my boyfriend, Chad Hensley, all of whom have provided unconditional support and inspiration throughout this arduous journey.

ACKNOWLEDGMENTS

I would like to take this opportunity to acknowledge and extend a thank you to all of the individuals who contributed to the creation of this dissertation. The support and guidance of my academic advisor and committee chair, Dr. Maricle, has been instrumental in my completion of this dissertation, as well as in my overall growth as a professional. I would like to thank Dr. Miller for allowing me the opportunity to use data collected from the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program for this dissertation, as well as his assistance through constructive feedback and thoughtful advice throughout the writing process. A special thanks also goes out to the other members of my dissertation committee, Dr. DeOrnellas and Dr. Scott, both of whom provided invaluable support throughout this process. The time and endless encouragement imparted by the members of my dissertation committee is most appreciated, as it was with their support that the completion of this dissertation was made possible. I would also like to express my gratitude to the dear friends I made during my graduate school journey. For it was the special bond we shared that gave me the strength and confidence to persevere even in times of self doubt. Finally, I am thankful for the ongoing support that my family has provided during the production of this dissertation.

ABSTRACT

PAMELA E. CIOFFI

EXAMINING THE CULTURAL LOADING AND LINGUISTIC DEMAND OF THREE NEUROPSYCHOLOGICAL ASSESSMENT BATTERIES FOR CHILDREN IN A MIXED CLINICAL POPULATION

DECEMBER 2015

There are a disproportionate number of culturally and linguistically diverse (CLD) children and adolescents referred for and placed in special education programs (Guthrie, 2004). These CLD individuals are tested using neurocognitive measures and methods plagued with issues ranging from culturally loaded test content and linguistically loaded test instructions to psychometric issues caused by poor normative reference samples and/or failure to remove error variance attributed to differences in performance for reasons other than cognitive ability level (i.e., level of acculturation or level of language proficiency; Ortiz, Ochoa, & Dynda, 2012). There is a need to examine the influence of cultural loading and linguistic demand on the test performance of CLD individuals. The Culture-Language Test Classifications (C-LTC) and the Culture-Language Interpretive Matrix (C-LIM) are an integrated system that uses classifications of cognitive test batteries on the two dimensions of cultural loading and linguistic demand to determine whether a given test performance reflects differences on these dimensions or a true measurement of ability (Flanagan, Ortiz, & Alfonso, 2007, 2013). However, research on the validity of the C-LTC and C-LIM is limited. A study was conducted using archival data from a mixed clinical sample (n = 520) of children and adolescents from school neuropsychology case studies. This sample included test scores obtained by the participants for selected tests from three standardized neuropsychological test batteries: the *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update* (WJ III COG NU; McGrew & Woodcock, 2001); the *NEPSY-II: A Developmental Neuropsychological Assessment* (NEPSY-II; Korkman, Kirk, & Kemp, 2007); and the *Delis-Kaplan Executive Function System* (D-KEFS; Delis, Kaplan, & Kramer, 2001a). A multivariate analysis of variance (MANOVA) revealed differences in performance existed between ethnicity groups on selected subtests from the WJ III COG NU, NEPSY-II, and D-KEFS. A discriminant function analysis (DFA) was used as a post-hoc measure. Findings appear to suggest that the C-LTC ratings have some validity for use in CLD individuals. However, there were instances in which subtests rated highly for both cultural loading and linguistic demand were not found to be statistically different across ethnicity in this study. Limitations and implications of this study are presented.

TABLE OF CONTENTS

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

CHAPTER I

INTRODUCTION

While there is the appearance that psychologists today have easy access to a robust arsenal of neurocognitive assessment tools with which to evaluate children, a more critical look at these tools reveals an abundance of psychometric flaws and limitations for their use. A review of the history of intellectual assessment reveals longstanding cultural biases, racial biases, and linguistic biases that influenced the development of early intelligence theories (Skiba et al., 2008; Wasserman, 2012). These psychometric issues have been acknowledged by researchers in the field of psychology since the inception of intelligence testing and intelligence test battery development in the early twentieth century (Gottfredson & Saklofske, 2009).

Conceptualizations of intelligence have historically been construed by groups of people who were motivated to differentiate among individuals perceived to be superior based on some set of beliefs (Guthrie, 2004; Wasserman, 2012). This desire to discriminate "us" versus "them" has been a mentality that has inspired investigation among scientific communities for potential differences among individuals from different cultural backgrounds. Definitions of race and ethnicity have been derived from the interpretations from such research, as well as from social constructions within a given society (Guthrie, 2004; Schaefer, 2006). Thus, theories of intelligence have been tainted with bias and subjectivity associated with these streams of thinking, and the bias woven into these theories appears to persist, despite a century of research and modifications. Furthermore, these theories of intelligence have influenced the methods and tools used to measure cognitive ability (Wasserman, 2012). Consequently, there are flaws inherent to the use of these methods and tools in the cognitive assessment of culturally and linguistically diverse (CLD) individuals.

The first tests used to measure intelligence, such as the Binet-Simon Intelligence Scale (Binet & Simon, 1905), Army Alpha, and Army Beta (Yerkes, 1921), were developed for use in determining placement and eligibility for program-level decisions (Guthrie, 2004; Wasserman, 2012). For instance, the Binet-Simon Intelligence Scale was used to make eligibility decisions for entrance to public education in France, and the Army Alpha and Army Beta were used to make placement decisions for the United States (U.S.) military. Problems associated with these early tests were detected by interpreters tasked at translating the Binet-Simon Intelligence Scale to immigrants who were just arriving in the U.S. at Ellis Island (Ortiz, Ochoa, & Dynda, 2012). In spite of one interpreter's contention that he could not have performed successfully on the test upon his own first arrival to the U.S., Henry Herbert Goddard, the American psychologist leading this investigation at Ellis Island, rebuffed the notion and "convinced him that the boy was defective" (Goddard, 1913, p. 105).

The strong influence that preconceived notions of intelligence has had on the development of testing methods and tools used to measure it has also had consequences for those affected by subsequent interpretations. In particular, CLD individuals have consistently received performance scores an average of one standard deviation below the mean on tests of intelligence, been disproportionately diagnosed with intellectual and psychological disorders, and been disproportionately placed in special education programs (Guthrie, 2004). Furthermore, CLD individuals have historically had difficulty gaining entrance to graduate level psychology programs, finding employment opportunities once a doctoral degree was earned, and gaining recognition for the contributions that they have made within the field of psychology. The disproportionately low number of CLD psychologists contributes to the difficulty in developing and implementing valid assessment practices for CLD individuals.

Issues pertaining to neurocognitive assessment with CLD individuals stem from several main factors, such as the use of test batteries based on intelligence theories that are tainted with cultural and linguistic bias (Guthrie, 2004; Wasserman, 2012). Other main factors are the lack of appropriate tests and procedures for assessment with CLD individuals and the inappropriate decisions (i.e., referrals to special education) that are made, which are guided by the interpretations of the tools and procedures that are currently in use by practitioners (Skiba et al., 2008). Specific problems with neurocognitive assessment batteries range from culturally biased test content and linguistically biased test instructions to psychometric issues caused by poor normative reference samples and/or failure to remove error variance attributed to differences in performance for reasons other than cognitive ability level (i.e., level of acculturation or level of language proficiency; Ortiz et al., 2012).

There is a collection of literature pertaining to the influence of culture on one's cognitive development. For example, Wang (2008) postulates that autobiographical memory is constructed and organized differently, depending on the type of culture one adopts (i.e., individualistic or collectivistic). The potential implications of a finding such as this could pose a serious threat to the validity of any tests based on theories that claim to be all-inclusive. If, for example, children from opposing cultures were administered a measure that tapped into a memory store that was influenced by this differentiation, there could be a significant difference observed between their performances. As with any other neurocognitive construct, differences in performance should be interpreted within a biological-psychological-sociological (bio-psycho-social) framework (Meyer & Melchert, 2011). If any variation exists in the organization or functioning of any one neurocognitive process across cultures, differences in performance should be at least partially attributed to this aspect.

In addition to the effect of one's level of acculturation, or familiarity with mainstream culture, on cognitive test performance, the linguistic demands of a neurocognitive test can also influence the performance of a CLD individual (Ortiz et al., 2012). Linguistic demands of a test include the language in which the test is given, the language in which responses must be provided during the test, the complexity of language used, any nonverbal language commands that must be known or used, and the extent to which language is needed in order for the test to be administered by the examiner and completed by the examinee (Flanagan, Ortiz, & Alfonso, 2013; Ortiz et al., 2012). The influence of language on test performance was recognized as early as 1921 by Robert Yerkes, who used this realization to create the Army Beta as an alternative version to his Army Alpha test for use with individuals who did not speak English or who could not read or write.

In response the to issues identified with neurocognitive assessment of CLD individuals, several methods and tools have been developed. Methods employed for assessment with individuals who are linguistically diverse (i.e., limited English proficiency, bilingual, nonverbal) include native language testing using a translator or interpreter, the use of "nonverbal" or language-reduced measures, and the use of alternate forms (e.g., *Wechsler Intelligence Scale for Children – Fourth Edition, Spanish*, WISC-IV Spanish; Wechsler, 2005). Another approach test developers have used in attempt to address the psychometric validity of neurocognitive test batteries has been through modifications to the procedures used to create representative normative samples (Ortiz et al., 2012). These test developers have sought to bridge the performance gap created by cultural and linguistic differences with the inclusion of more individuals with diverse racial and ethnic backgrounds. However, the cultural categories of race and ethnicity are typically the only variables addressed in the updated standardization samples, which leaves many other critical cultural and linguistic variables unaddressed.

The Cattell-Horn-Carroll (CHC) cross-battery assessment approach is another method that is being used by practitioners in the assessment of CLD individuals (Flanagan et al., 2013; Ortiz et al., 2012). Cross-battery assessment provides practitioners with a research-based, reliable method for the creation of individualized assessment protocols to meet the needs for cognitive evaluations. The examiner is able to do this through systematic selection of cognitive subtests to tap into broad and narrow CHC abilities, as guided by the cross-battery assessment procedures. This selective method for putting together an individualized assessment protocol allows examiners the opportunity to choose subtests that would best measure cognitive abilities of CLD examinees. Samuel Ortiz and Dawn Flanagan (1998) took this idea and built on it with the development of the Culture-Language Test Classifications (C-LTC). The C-LTC is a classification system based on expert consensus, empirical studies, and available data from published tests that organizes the influence of two dimensions, cultural loading and linguistic demand, for a collection of commonly used cognitive test batteries.

To further investigate the degree to which a test performance is affected by the cultural loading or linguistic demand of a test battery, Flanagan, Ortiz, and Alfonso (2007) developed the Culture-Language Interpretive Matrix (C-LIM). Creation of the C-LIM allows examiners to record subtest scores from individual cases into the C-LIM Microsoft Excel program in accordance to the specified classifications assigned in the respective C-LTC (Flanagan et al., 2007, 2013). The program then calculates mean scores and assists the examiner in determining if a pattern of scores is reflective of cultural and linguistic bias or a true measurement of the examinee's abilities. With the use of this tool, practitioners may be able to make more defensible decisions as to whether the observed low performance of a CLD individual is due to cultural and linguistic differences or the presence of an underlying disorder. While the introduction of the C-LTC and the C-LIM offers a promising approach to solving the issues of neurocognitive assessment with CLD individuals, more research is needed to substantiate the reliability and validity of their use.

Rationale and Purpose for this Study

The current state of cognitive and neuropsychological testing with CLD individuals is plagued with issues that stem from test selection, test administration, and test interpretation, and/or through the interventions that are carried out as a result of testing. Differences in neurocognitive test performance that result from factors relating to one's level of acculturation or language proficiency should be interpreted with recognition that this critical background information may be representative of cultural or linguistic differences, rather than the presence of a disability. The literature review presented in this dissertation summarizes the diagnostic, interpretive, and psychometric issues that current measures of neuropsychological functioning possess with regard to the assessment of CLD individuals.

There is a need to investigate the impact of cultural loading and linguistic demand on measures of neurocognitive ability. To address this need, a research study was conducted using archival data from a mixed clinical sample of children and adolescents from case studies submitted by students of the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program. This sample included test scores obtained by the participants for selected tests from three standardized neuropsychological test batteries: the *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update* (WJ III COG NU; McGrew & Woodcock, 2001); the *NEPSY-II: A Developmental Neuropsychological Assessment* (NEPSY-II; Korkman, Kirk, & Kemp, 2007); and the *Delis-Kaplan Executive Function System* (D-KEFS; Delis, Kaplan, & Kramer, 2001a).

Research Questions and Hypotheses

The following research questions were posed:

1. Do differences in performance exist based upon ethnicity for the three standardized (WJ III COG NU, NEPSY-II, D-KEFS) measures of neurocognitive functioning?
2. How much of the difference in performance is attributed to ethnicity?

Based upon previous literature and research, it was expected that the measures investigated in this study would reveal statistically significant differences in performance across ethnicity groups. Further, it was expected that those who reported being of minority group status (African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic) would demonstrate a significantly lower performance than those who reported being of majority group status (Caucasian/White). A literature search and relevant research led to the hypothesis that the differences in performance based on ethnicity would be due to the cultural loading and linguistic demand of the neurocognitive measures used (Cormier, McGrew, & Evans, 2011). Specifically, the observed differences across ethnicity groups were expected to be found for those subtests that are rated high for cultural loading and linguistic demand, according to Flanagan and colleagues' (2007) C-LTC classifications for the WJ III COG NU, NESPY-II, and D-KEFS.

CHAPTER II

REVIEW OF THE LITERATURE

Currently, there is an overrepresentation of CLD children and adolescents who are identified and placed in special education in the United States (Reschly, 2009; Shifrer, Muller, & Callahan, 2011). This disproportionality is discussed by Reschly (2009), who argues that, “inappropriate special education participation harms students and diminishes educational attainment and career opportunities” (p. 57). Many factors influence this disparity, including cultural and contextual factors that influence cognitive development and educational opportunities, as well as the use of cognitive assessment batteries that are biased towards a Westernized, Caucasian, socioeconomically advantaged population of individuals (Nampija et al., 2010).

As such, it is imperative to explore these issues as they relate to the assessment of neurocognitive functioning in children with CLD backgrounds. Performance on such measures is typically interpreted along with other information gathered for neuropsychological evaluations, and any impact a child’s background might have on their observed performance should be taken into consideration when conclusions and recommendations are developed. A thorough investigation of the differences in performance observed due to linguistic demand or cultural loading may reveal associated implications that could threaten the validity of neuropsychological test batteries currently in use. This chapter will summarize the relevant research pertaining to these issues and introduce the purpose and rationale for the current study.

Cultural and Linguistically Diverse Individuals in the Field of Psychology

Definitions

Culturally and linguistically diverse (CLD) refers to a broad group of individuals that vary with regard to cultural and linguistic backgrounds, such as level of proficiency in the dominant language, level of acculturation, socioeconomic status, race, ethnicity, or education level (Cormier et al., 2011). Defining the cultural categories of "race" and "ethnicity" has long been a subject of disagreement and controversy among professionals in scientific fields, including psychology, anthropology, and physiology (Guthrie, 2004). Lack of continuity for definitions of these terms is one reason that there is also inconsistency in the application of categories for research purposes, such as the U.S. Census. For the purposes of this study, definitions of culture, majority, minority, race, ethnicity, acculturation, language, and language proficiency will be articulated in the following subsections.

Culture. Culture is described as, "the totality of learned, socially transmitted customs, knowledge, material objects…behavior….ideas, values, customs, and artifacts…of groups of people" (Schaefer, 2006, p. 34). It includes the mode of communication, family structure, and ways of promoting standards of right and wrong within a group of people. Culture becomes embedded and learned by an individual through interactions with other members of the group from the time of birth. Aspects of culture vary across groups of people, and one's level of identification within a culture may also vary.

Acculturation. Gasquoine (2009) defines acculturation as, "the socialization process whereby members of minority groupings gradually learn and adopt certain elements of mainstream culture from continuous first-hand contact" (p. 255). Acculturation can be viewed as a process that does not have a set or universal path, or end result across individuals (Van de Vijver & Phalet, 2004). Psychologists, anthropologists, and sociologists initially viewed acculturation as a unidimensional model in which immigrants gradually adapted and assimilated from their original culture to that of the mainstream culture. However, upon recognition that an increasing proportion of migrants were not following this model towards complete adjustment or adoption of mainstream culture and instead opting to maintain their original culture or developing a bicultural identity, bidimensional models of acculturation were introduced. For example, Berry's (1997) model of acculturation looks at two dimensions: adaptation and cultural maintenance. The process of acculturation may look different across individuals, and may also be influenced by length of residence or generational status (i.e., first generation immigrant versus second or third generation immigrant).

Majority and minority. Majority status refers to a group of people within a population whose classification in some cultural or demographic category (i.e., race, ethnicity, primary language) falls within the largest proportion in comparison to other classifications in said category (Schaefer, 2006). Minority status can be assigned to any group of people whose classification in some cultural or demographic category is not that of the majority (Schölmerich, Leyendecker, Citlak, Caspar, & Jäkel, 2008). The definition of minority status described here also applies to the other demographic and cultural characteristics referred to throughout this research paper, including primary language spoken and level of acculturation. What is critical to point out here is that the distribution of people within different geographic locations or cultures varies, and those who might be in the majority group in one environment might fall into the minority group in another. For example, a United States (U.S.) citizen who speaks English as a first language falls in the majority group for this category whilst in the U.S., but may fall in the minority group for this category if he or she were to relocate to another country with a different dominant language (i.e., Germany). Schaefer (2006) also points out that a numerical minority, which is a group making up less than half of a larger population, does not necessarily equate to the sociological conceptualization of a minority group, which also considers the subordinate position of its members in relation to the dominant or majority group.

Race and ethnicity. It is a common opinion among scientific literature of the 21st century that the concept of race is a socially constructed classification system based on physical traits/characteristics and subject to influence from historical, cultural, and economic factors within a given society (Guthrie, 2004; Schaefer, 2006). For example, traditional ideas about race led to the de facto assignment of individuals with dark skin tones, dark eyes, and dark hair to be labeled *Black*, and individuals with pale skin tones, light eyes, and light hair to be labeled *White*. However, this method of classification does not serve to differentiate among any characteristics individuals possess that are not physical, such as nation of origin.

One's ethnicity "can be viewed as the heritage, nationality group, lineage, or country of birth of the person or person's parents or ancestors" (Humes, Jones, & Ramirez, 2011, p. 2). Within race categories, individuals may self-identify with differing ethnicities, or origins. Ethnicity groups commonly used in the U.S. include Caucasian, Hispanic/Latino, Asian American, Native American, and African-American. However, these classifications have also been described as too broad, because there is generally a wide range of more specific groups comprising each of them, all with their own unique cultural characteristics. For instance, the ethnic designation of Hispanic/Latino may be endorsed by groups of people from nations or regions of origin such as Spain, Mexico, Guatemala, and Chile. Nonetheless, these broad ethnicity group classifications are frequently used in multicultural research with the assumption that they are sufficient in being representative of the individual cultural and linguistic characteristics each group possesses. Furthermore, it is also common to observe the use of both race and ethnicity classifications alongside its socially agreed upon counterpart (i.e., White/Caucasian, Black/African-American). Despite the unreliable nature of these racial and ethnic group classifications, they continue to be used for research and other census related purposes. For the purposes of this study, race refers to the socially constructed classifications based on physical traits, and ethnicity refers to the commonly used classifications that are based on nation/region of origin. When describing the participant sample used in this study, ethnicity was the term applied; however, the classifications used in the archival sample of data used were labeled with a combination of race and ethnicity designations. The implications of this system of classification is addressed in the Methodological Issues and Limitations section of chapter five.

Language. A definition of language provided by Garrett (2009) posits that: "[Language] includes the generation and understanding of written, spoken, and gestural communication" (p. 262). Schaefer (2006) describes language as, "an abstract system of word meanings and symbols for all aspects of culture. It includes speech, written characters, numerals, symbols, and gestures and expressions of nonverbal communication" (p. 38). Research into the neurobiology of language has implicated multiple regions and neural circuitry systems that are involved in its acquisition, comprehension, and production. The production of language requires both cognitive and motoric functions by an individual (Carter, Aldridge, Page, & Parker, 2009). Furthermore, it has been found that some language functions, such as word recognition, are specialized to just one hemisphere of the brain. Because language has many components, differentiation of processing in the brain is the anticipated scenario. Thus, depending on the type of language system used, there may be implications for neurocognitive assessment. For example, assessment of individuals who use spoken language (i.e., English) will differ from individuals who use hand signals to communicate (i.e., sign language). Spoken language relies primarily on left hemispheric activation, while sign language uses more right hemispheric activation.

Another example of differential brain activation is in the case that an individual has experience and/or fluency in more than one language (Carter et al., 2009; Garrett, 2009; Kempert, Saalbach, & Hardy, 2011). Language proficiency refers to the level of language development achieved with regard to understanding, speaking, reading, and writing (Olvera & Gomez-Cerrillo, 2011). There are two dimensions in which language proficiency can be measured: cognitive academic language proficiency (CALP) and basic interpersonal communication skills (BICS). CALP refers to the level of complexity of language development needed for academic learning, and BICS is comprised of the language skills needed for informal types of communication for more social types of settings (Olvera & Gomez-Cerrillo, 2011). Proficiency may be obtained in more than one language for an individual, in which case he or she would be considered bilingual. Depending on the age at which the second language is introduced, an individual may utilize different streams of neural processing. Kempert and colleagues (2011) conducted a study in which monolinguals were compared to bilinguals in their ability to learn mathematical word problems. Findings of the study suggested that the bilingual students have a cognitive advantage over their monolingual peers with regard to attentional control, due to their relatively higher need for it to switch between their two known languages. In light of the complexity of language processing in the brain, this is a critical factor to take into consideration in the development of an assessment plan for a client who is linguistically diverse.

The need to clarify definitions related to cultural and linguistic diversity for the purpose of this research study exemplifies just how much variation exists in the interpretation of such terms. Furthermore, the lack of consistency has implications for situations in which individuals are required to identify or ascribe to an attribute by selecting among some set of choices. For example, the U.S. Census requires individuals to select their race from a specific list of choices (Guthrie, 2004). This may lead to issues for individuals who identify with more than one of the choices provided or if none of the choices provided reflect their own resolute distinctiveness. What this discussion also elucidates is how the misunderstanding and inappropriate assignment of cultural and linguistic attributes permeates into larger systemic problems. A summary of related issues will be articulated in the subsequent sections.

Disproportionality in Special Education

Between fall of 2000 and fall of 2010, the percentage distribution for the race/ethnicity of U.S. public school students enrolled in prekindergarten through 12th grade underwent some notable changes (Aud et al., 2013). As reported in the 2013 publication prepared for by the National Center for Education Statistics (NCES), the percentage of White/Caucasian students enrolled decreased from 61 to 52 percent. Meanwhile, the percentage of Hispanic/Latino students enrolled increased from 16 to 23 percent in the same time period. Increases in percentage of students enrolled were also observed for the Asian/Pacific Islander group, as well as the group comprised of those with two or more races. The percentages for African-American/Black and American Indian/Alaska Native students enrolled showed only negligible fluctuations across this same time span. These data are reflective of the changes that have been observed in the overall U.S. population characteristics due to immigration throughout its history (Jones, Sander, & Booker, 2013).

In conjunction with the increasingly diverse racial/ethnic makeup of the U.S. population over the last century, so too has the diversity expanded for languages spoken, religions/belief systems upheld, and many other cultural customs (Schölmerich et al., 2008). However, multicultural competency among psychologists and other professionals working with CLD individuals remains considerably limited. Issues stemming from this area of weakness have been overlooked and/or trivialized by researchers and practitioners for decades (Skiba et al., 2008). It is speculated that reasons for this include a lack of specific knowledge for methods to ameliorate the issues, belief that current alternative procedures for cross-cultural assessment yield valid and reliable scores, inadequate methodologies applied in multicultural research studies, or the failure to recognize the significance of certain cultural and linguistic differences (Byrne et al., 2009; Ortiz, 2006; Vazquez-Nuttall et al., 2007). The consensus among recent researchers is that the underrepresentation of CLD students in gifted education and overrepresentation in special education may be largely attributed to these multicultural competency issues (Ford, Moore, Whiting, & Grantham, 2008; Jones et al., 2013; Skiba et al., 2008).

Disproportionality of CLD students in special education has been found to manifest from a variety of factors, which include psychometric test bias, economic disadvantage, decisions at the time of initial referral, differential teacher ratings, and the process by which eligibility decisions are made (Ford et al., 2008; Jones et al., 2013; Laing & Kamhi, 2003; Skiba et al., 2008). The Individuals with Disabilities Education Improvement Act (IDEA, 2004, Public Law No. 108-446) has provisions, which were initially put in place during the 1997 IDEA reauthorization (IDEA, 1997, Public Law No.

105-117), that specifically address the need to remediate disproportionality in special education. These provisions require States receiving funding under IDEA to monitor State and local education agencies (LEA) for disproportionate representation of racial and ethnic groups in special education and related services. Should data reveal disproportionality, the LEA is required to allocate the maximum amount of federal funding received under IDEA to early intervening services directed towards those groups found to be overidentified. This legislation came about following decades of research and numerous court cases (*Hobson v. Hansen,* 1967; *Diana v. California State Board of Education*, 1970; *Guadalupe Organization, Inc. v. Tempe Elementary School District No. 3*, 1972; *Larry P. v. Riles*, 1984; and *PASE v. Hannon*, 1980), all of which sought to reveal and eradicate ethnic and language-based disparities observed in special education. Research into these issues has consistently indicated that the U.S. history of racial and ethnic oppression and discrimination has greatly influenced the intelligence measures being used and subsequent eligibility decisions.

Disproportionality Among Professionals

The field of psychology has historically been dominated by professionals with a particular set of demographic characteristics: Caucasian, male, and upper-middle class (Benjamin, 2009). Since psychology became established in the United States (U.S.) at the beginning of the 20th century, U.S. psychologists have also been a predominantly English-speaking group. This biased distribution of psychologists has had extensive implications for research and practice in the field, including the lack of proliferation of multicultural training and competency, as well as a reputation for cultural bias (Byrne et al., 2009; Ford et al., 2008; Jones et al., 2013).

Historically, the contributions of minority and women psychologists have been largely ignored (Guthrie, 2004). The racial and ethnic discrimination within the U.S. in the early 20th century made it difficult for individuals with minority status (e.g., Black or African-American; Guthrie, 2004) to gain entrance into graduate level psychology programs, find employment opportunities once a doctoral degree was earned, and to be taken seriously among the field of psychology. Presently, although the distribution of psychologists in the U.S. has shifted due to an increasing volume of racial, ethnic, linguistic, and other cultural minorities, the field remains overwhelmingly populated with White/Caucasian, English-dominant, majority status individuals (Ford et al., 2008; Guthrie, 2004). Issues related to this disproportionality among psychologists contribute to the difficulty in valid assessment practices with CLD individuals. Further explanation of these issues will be delineated in a subsequent section.

History of Intellectual Testing with CLD Individuals

Theories of Intelligence

Conceptualizations of intelligence can be traced back well before psychology was established as a legitimate field of science in the mid-19th century (Benjamin, 2009; Gottfredson & Saklofske, 2009; Wasserman, 2012). In fact, mental faculties were a topic of contemplation over two millennia ago by ancient Greek philosophers, such as Plato and Aristotle (Burnham, 1888). At the time, the concept of dualism was essential to the conceptualization of human thought (Burnham, 1888; Packard & Chen, 2005). Religion and spirituality were dominant in the infrastructure of European society; therefore, philosophers incorporated the influence of God in their theories. For example, Pythagoras discussed memory as a function of the soul, originating from a former state of existence. Thus, a distinction is made between the physical body and the intangible soul. This dualistic view of the mind-body persisted through the centuries, influencing intellectual thought, as well as the political regimes of Europe during the Middle Ages (Packard & Chen, 2005). During this time, Protestantism was the framework for the organization of political power, and citizens strived to achieve the qualities of an exemplary Protestant. People believed that true perfection only existed in God, and that they must actively strive to reach a state of goodness. Thus, a culturally derived theory of intelligence and moral character was contrived.

It was not until the late 19th century and early 20th century that theorists began to place sole emphasis upon the body for all processes involved in cognition (Jääskeläinen, 1998; Packard & Chen, 2005). The pseudo-scientific theory of phrenology, introduced by Franz Joseph Gall in the 19th century, was closely aligned with the theoretical frameworks driving the racial classification methodology by anthropologists during that time (Guthrie, 2004). Phrenology was based on the accurate belief that the brain was the source of mental faculties, or cognitive processes, and that mental faculties were controlled in different locations within the brain (Guthrie, 2004; Wasserman, 2012). However, Gall mistakenly believed that the size and shape of the brain was reflected in the surface of the skull. Thus, phrenology was based on the idea that one could use the measurements of an individual's skull to determine the development of independent mental faculties. This method was used to support an already widely held assumption of racial differences based on physical characteristics; however, the added component of mental ability to a scientifically flawed system of measurement served to inaccurately adhere to the belief that individuals of the White race possessed superior intellectual abilities compared to all other races (i.e., Black individuals).

The introduction of Charles Darwin's concepts of evolution and natural selection and his application of these to human mental abilities served to influence conceptualizations of intelligence within the field of psychology (Wasserman, 2012). For example, Lewis Terman firmly believed that intelligence was predominantly determined by heredity and far less influenced by one's environment (Guthrie, 2004; Wasserman, 2012). Terman was outspoken with his presumption that low intellectual capacities, such as those possessed by individuals with intellectual disabilities, were more commonly present in racial minorities. He went on to insist that this level of functioning was innately characteristic of these populations. Edward Thorndike followed Terman's line of thought, proclaiming that intelligence was 80% genetic, 17% attributed to educational opportunity, and 3% was deemed accidental. This further marginalized individuals in minority groups, as this statement implied that they had inherently inferior intellectual capacities. These biased perspectives not only influenced conceptualizations of cognitive functioning, but also the development and standardization of intelligence measures.

Test Development

Early assessment batteries were standardized with normative samples that excluded CLD individuals, thereby accumulating no validity for use in assessing these populations (Gottfredson & Saklofske, 2009; Laing & Kamhi, 2003). Two forms of cultural bias are discussed here, that which is due to acculturative differences and that which is due to linguistic diversity (Fraine & McDade, 2009). While it became evident that factors, such as language or knowledge of mainstream culture, affected the performance of CLD individuals on intelligence measures from the time they were first introduced at the beginning of the 20th century (e.g., Binet-Simon Intelligence Scale; Binet & Simon, 1905), the finding was met with mixed responses (Gasquoine, 2009; Gottfredson & Saklofske, 2009; Rembis, 2004). In some cases, there was a sense of denial and attribution of low performances to the prevailing Zeitgeist of the early 1900s, which included rampant discrimination of CLD individuals and increasing favoritism and laws requiring sterilization of those deemed intellectually inferior, or “feeble-minded” (Flanagan et al., 2013; Gasquoine, 2009; Skiba et al., 2008; Wasserman, 2012). However, there was some recognition that alternative tests designed to be administered nonverbally (i.e., using pictures, puzzles, or blocks) would be useful in assessing immigrants who did not speak or understand English, as well as individuals who could not read or write.

In a chapter of Robert Yerkes’ (1921) book, entitled, “Memoirs of the National Academy of Sciences, Vol. XV: Psychological Examining in the United States,” he described the development, examination data, and revisions of the Army Beta. This test, Yerkes explained, was adapted from the Army Alpha for the purpose of designing a measure of intelligence that could be used with non-English speaking immigrants and individuals who were illiterate. While the creation of the Army Beta was a step in the right direction in the valid assessment of intelligence in CLD individuals, the test had its limitations (Wasserman, 2012). The instructions for the Army Beta, while with the addition of physical gestures (e.g., pointing to items), continued to be verbally presented in the English language. Furthermore, the Picture Completion Test included drawings that referenced objects or situations that were culturally biased towards U.S. citizens. Those taking the test would require some knowledge of U.S. culture in order to recognize certain missing elements from an illustration, such as the bowling ball missing from the hands of two people pictured in front of two rows of bowling lanes.

Tests of intelligence are influenced by the cultural and linguistic background of those responsible for their development, but they are also a reflection of the sample from which they were standardized (Ortiz et al., 2012). Development of the first test of intelligence, the Binet-Simon Intelligence Scale (Binet & Simon, 1905) is credited to Alfred Binet, with the contributions of Theophile Simon and Victor Henri (Wasserman, 2012). This test was created for use in public schools of France to determine eligibility for public education, and while it was touted as a scientific tool for assessing individual differences in intelligence, Binet warned that valid measurement could only be obtained if the individual being tested had the same or similar environmental and educational experiences (Guthrie, 2004).

The problem of norm group comparison on tests of intelligence continued to be an issue in the assessment of CLD individuals upon the introduction of intelligence test batteries following the Binet-Simon Intelligence Scale (Binet & Simon, 1905). For instance, the standardization sample used in Terman's translated and revised version of the Stanford-Binet in 1916 was comprised of 905 participants drawn from California, Nevada, and Oregon, none of which were of immigrant or minority status (Wasserman, 2012). While the Army Alpha and Army Beta tests were used to guide decisions regarding military selection and placement, tests such as the Stanford-Binet drove placement decisions regarding education. However, the advent of intelligence testing in the U.S. also contributed to legal, political, and public policy decisions regarding the practice of eugenics, or mandatory sterilization of those deemed mentally or intellectually "defective."

The Society for the Psychological Study of Social Issues (SPSSI) is an organization that was founded amidst the Great Depression, a time when the social problems of unemployment, hunger, racism, labor-management disputes, poverty, and impending war were brought into the awareness of the general U.S. population (Benjamin, 2009). This organization, which focused on research of social problems, protested on the issues of racial psychology, issuing a statement in 1938 (Guthrie, 2004). The primary message contained within this statement was that the experiments and theories pertaining to the assignment of intellectual ability or superiority of a race or group of people based on physical characteristics held no scientific basis or support (SPSSI, 1938).

Following World War II and the mass genocide by Nazi Germany on the premise of racial superiority, the eugenics movement ended and psychologists began to leave behind their beliefs that inherent racial differences were the reason for lower scores on measures of intelligence for CLD individuals (Benjamin, 2009; Wasserman, 2012). By the 1960s and 1970s, the U.S. civil rights movement and increasing concerns regarding fairness in testing and educational placement led to a resurgence of studies evaluating biased intelligence measures (Gottfredson & Saklofske, 2009; Skiba et al., 2008). Since that time, test developers have increasingly made attempts to address the issue of test bias by including more individuals from racial and ethnic minority groups into their norm samples, attempting to create “nonverbal” measures, and various other strategies, which will be described in a subsequent section of this chapter. However, there remain significant weaknesses with regard to intelligence test batteries, as well as assessment procedures and test interpretation.

Current Issues in Test Batteries and Assessment

To this date, there are flaws in the standardization procedures that are utilized to validate cognitive assessment batteries. While quota sampling may be used as a means for approaching demographically proportionate samples to match the distribution of age, gender, race/ethnicity, and education level according to the U.S. Census, there are still limitations (Williams & Cottle, 2011). It has been argued that this endeavor does not guarantee an adequate sample of each of the minority populations, as the normative sample will still be overwhelmingly comprised of those with majority group status. Thus, the weighted distribution of the sample precludes the individuals of the minority group from having much emphasis on the weighted norm for a test battery. Furthermore, Laing and Kamhi (2003) point out that even with adjustments to normative samples, CLD children will continue to appear as though they perform below the mean of their age-matched peers due to factors such as unfamiliarity with test content or limited English proficiency (LEP).

Issues pertaining to the cognitive assessment of CLD individuals move beyond just flaws with standardization procedures. As the previous discussion regarding the history of intellectual assessment illustrates, even the structure and content of the test batteries in use are plagued with cultural and linguistic bias. Furthermore, the actions taken during (i.e., modification of instructions) and following (i.e., referral to special education) a cognitive ability evaluation of a CLD individual have led to compromised validity for test score performance, as well as inappropriate interpretations and intervention strategies (Skiba et al., 2008).

Biased Theories of Intelligence

Increased knowledge of the underlying structures and pathways for neurocognitive processing has led to theoretical models that are increasingly convoluted. Nevertheless, there are some common themes in the evolution of these models. One movement in the literature has been towards the separation of processing streams based on sensory modality (Floyd & Kranzler, 2012). Additionally, most comprehensive conceptual models of cognitive functioning consist of a hierarchical structure, with higher-level cognitive processes being dependent upon lower-level, basic processes. For example, the CHC model of cognitive functioning is structured in this way, with narrow abilities (i.e., short-term memory span) clustering to form broad abilities (i.e., short-term memory; Gsm). All broad abilities are thought to contribute to an overall intelligence quotient (*g*) (McGrew, 2005).

In another example, the school neuropsychology conceptual model (SNP model; Miller, 2007, 2010) was recently updated based on current psychometric and theoretical research to become the Integrated SNP/CHC model (Miller, 2013). This model organizes cognitive functioning into four major classifications. These include: (1) basic sensorimotor functions, (2) facilitators and inhibitors for cognitive processes and acquired knowledge skills, (3) basic cognitive processes, and (4) acquired knowledge. However, less emphasis is placed on overall intelligence, as the distinctive strengths and weaknesses among the neurocognitive constructs are deemed more informative of an individual's cognitive profile.

While there continues to be disagreement in the literature regarding the most accurate model of neurocognitive processing, the models described here are alike in their emphasis on the division of pathways for information processing and sensory input modality. However, empirical support for them has been primarily established using behavioral experimental studies, which necessitate only theoretical inferences (Repovš & Baddeley, 2006). Additionally, the validation procedures of these models have relied primarily upon assessment with neuropsychological test batteries, whose standardization groups were prone to recruitment bias (Williams & Cottle, 2011). A representative group is necessary to establish a reference of comparison for identifying relative impairment levels on a measure. However, the standardization groups used in the validation of neuropsychological test batteries have historically been biased towards the culture in which the test was developed (Van de Vijver & Tanzer, 2004).

Neuropsychological test batteries are generally developed to assess cognitive functioning based on the foundation of some theoretical conceptualization, such as those previously discussed. By having a strong theoretical framework, these tests are believed to provide valid and reliable estimates of neurocognitive functioning (Kamphaus, Winsor, Rowe, & Kim, 2012). However, neuropsychological test batteries presently in use, despite decades of revisions, maintain theoretical foundations that arose at a time of largely biased views towards CLD individuals (Guthrie, 2004) or lack a theoretical foundation altogether (e.g., Delis-Kaplan Executive Function System, D-KEFS; Homack, Lee, & Riccio, 2005). Research is needed to determine construct comparability for neuropsychological test batteries currently in use with CLD individuals.

In order to ensure valid comparisons in cross-cultural assessment, the examiner must utilize testing instruments that possess measurement invariance. Measurement invariance is described as the instance in which a construct is consistently measured across different groups, such as those of different cultures (Chen, 2008). This is a necessity for researchers intending to make cross-cultural comparisons on any particular construct. A study by Chen (2008) was conducted to evaluate the consequences of measurement invariance in the assessment and interpretations derived from cross-cultural research. To evaluate the impact of assessment in the absence of measurement invariance, a series of regression analyses were conducted to examine the predictive relationship of self-esteem on life satisfaction across sample groups of Chinese and Caucasian-American students. Results indicated that a lack of measurement invariance was found when a scale developed for one culture or reference group (e.g., Caucasian-American students) was used with a sample from a different culture or focal group (e.g., Chinese students). Furthermore, this lack of measurement invariance led to bias in means of both the reference group and the focal group, suggesting the potential for erroneous conclusions to be drawn. Should this type of situation occur in the neurocognitive assessment of CLD students, it is possible that inappropriate conclusions might also follow that might have implications for their educational programming.

To compound the limitations associated with the assessment of neurocognitive abilities, Robert Sternberg (2007, 2012) points out that various demands of different cultures, with regard to the activities and skills that are important for survival, influence the composition and development of one's cognitive profile. He refers to the concept of "successful intelligence," which is defined as "what is needed for success in life, according to one's own definition of success, within one's sociocultural context" (Sternberg, 2007, p. 148). Hence, it becomes clear that the Western conceptualization of intelligence may not generalize to individuals from cultural backgrounds that contrast with the Western culture (Nampija et al., 2010; Van de Vijver & Tanzer, 2004; Walker, Batchelor, & Shores, 2009).

Sternberg's theory of successful intelligence was the theoretical foundation for the development of the Aurora Battery, which is an assessment battery developed in 2004 to serve as an alternative or supplementary tool to the existing methods and instruments for measuring cognitive abilities in the U.S. (Sternberg, 2007, 2012; Tan et al., 2009). Aurora was designed to measure intelligence, as defined by the theory of successful intelligence, in a more broad scope with the goal being that it might be a more valuable tool in the assessment of CLD individuals (Tan et al., 2009). Sternberg's theory of successful intelligence and related research on measurement bias of neuropsychological test batteries supports the notion that not all theories of intelligence are universal.

Furthermore, test batteries developed from theoretical frameworks that are culturally loaded cannot produce a valid measurement of cognitive abilities in CLD individuals who come from a culture that have a different conceptualization of intelligence (Sternberg, 2007, 2012; Tan et al., 2009; Van de Vijver & Tanzer, 2004).

Lack of Appropriate Tests/Procedures for CLD Individuals

Many factors influence the availability and use of intelligence tests within a country or cultural population (Oakland, 2009). Test development is costly, and can cost upwards of $500,000. It is for this reason that test development and use is more common in countries with sufficient financial resources to support production. Furthermore, there must be a demand for such instruments and the presence of professionals with the training and expertise to utilize them. Countries or cultures that place emphasis on individualism tend to have a more competitive focus in which people are judged based on personal traits and accomplishments, whereas those with a collectivistic structure place value on the cohesiveness of the group as a whole. As such, individualistic countries and cultures are more likely than collectivistic countries and cultures to have and use intelligence tests because they serve the purpose of identifying individual differences that may be used for educational or occupational decisions. The result of these kinds of factors is a significant limitation in the availability of intelligence test batteries for CLD populations.

The cultural and linguistic influences upon cognitive functioning and performance discussed here are critical to take into consideration when evaluating the validity of neuropsychological assessment batteries for CLD children, especially when deciding the populations for which they can generalize. Further investigation of the possible differences in performance across CLD individuals is warranted. In particular, the psychometric properties of test batteries with regard to bias in these populations would be an important step towards addressing associated issues.

Van de Vijver and Phalet (2004) discuss three types of bias that may result from assessment of CLD individuals: (1) construct bias, (2) method bias, and (3) item bias or differential item functioning (DIF). Construct bias is said to exist in the event that a construct has an incomplete identity across groups or there is an incomplete overlap of behaviors associated with it. Method bias also has more than one possible cause, ranging from sample incomparability and instrument characteristics to tester/interviewer effects and mode of administration. Item bias or DIF occurs if there is a discrepancy in performance on an item that measures a construct which individuals from diverse backgrounds have demonstrated equal competence. The researchers address these issues in multicultural assessment with emphasis on the factor of acculturation. Van de Vijver and Phalet describe seven approaches for revealing and attempting to make adjustments for any cross-cultural test bias identified across varying levels of acculturation. The purpose of these efforts is to take this information into consideration and also to potentially develop a method that precludes examiners from obtaining invalid scores that "cannot be interpreted in the standard way" (p. 228).

Another phenomena that has been investigated for its influence on valid test performance is Steele and Aronson's (1995) theory of stereotype threat, defined as:

> The existence of a negative stereotype about a group to which one belongs…in situations where the stereotype is applicable, one is at risk of confirming it as a self-characterization, both to one's self and to others who know the stereotype….and…this threat can be disruptive enough…to impair intellectual performance. (Steele & Aronson, 1995, p. 808)

This research by Steel and Aronson (1995) demonstrated that the presence of stereotype threat in a testing situation with a sample of Black and White undergraduate students from Stanford University was salient enough to depress the performance of Black participants on measures of intellectual ability. The first three studies showed that Black participants' performance was impaired on intelligence test measures when the test was presented as being a diagnostic tool, as compared to White participants with the same emphasis on the diagnostic nature of the test or the Black and White participants who were presented the same test as a non-diagnostic problem-solving task. Furthermore, the third study showed that Black participants were less willing to indicate their race on the pre-test questionnaire if they were in the condition that emphasized the test as diagnostic of intellectual ability. To isolate the effect of stereotype threat, the fourth study revealed that Black participants had impaired intellectual test performance when asked to self-identify their race on the pre-test questionnaire even when the test was presented as a non-diagnostic measure.

The research of Steele and Aronson (1995) on stereotype threat and its effect on assessment in CLD populations has contributed to the growing body of literature regarding the discrimination of skill deficit from performance deficit on intelligence measures. Wicherts, Dolan, and Hessen (2005) conducted a series of studies demonstrating that stereotype threat effects were found among a sample of ethnic minorities from a Dutch high school in the Netherlands on a test of intelligence, as well as in a sample of male and female psychology undergraduate students from the Netherlands on a collection of arithmetic/mathematic tests. Suppression of performance scores attributed to stereotype threat was found to be the most pronounced on the more difficult subtests and subtests with strong correlations to the underlying construct believed to be measured by the test. Implications proposed by the authors are that stereotype threat is a source of measurement bias and may also result in a lack of measurement invariance across cultural groups due to differential item functioning (DIF).

Inappropriate Decision-Making

Interpretation of testing with CLD individuals is tainted by the difficulty in determining the validity of test scores obtained (Flanagan et al., 2013). This is commonly referred to as the “difference versus disorder” dilemma. The question is whether the scores obtained are due to differences in cultural or linguistic background or if a true disorder is present. In an article by Helms (2006), results were presented from a series of studies examining the effect of construct-irrelevant variance on test scores in a sample of first-year Black college students. The outcome was that racial identity variance translated into construct-irrelevant variance that led to underestimated performance scores on a test of mathematic knowledge, ability, and skills. These findings demonstrate the need to recognize the presence of construct-irrelevant variance in the assessment of CLD individuals, as failure to do so can result in inaccurate interpretations (i.e., diagnosis of disability instead of culture-related difference) and inappropriate education placement decisions (i.e., placement in special education when not actually needed).

In addition to the problems associated with the tests and test procedures for intellectual assessment of CLD individuals, there are issues with regard to the decision-making process by the professionals involved in the process at a systemic level (Jones et al., 2013). For example, there are a disproportionate number of minority students who are referred for special education evaluations. Skiba and colleagues (2008) argue that this referral bias starts at the classroom level with the students' teacher. Students of racial, ethnic, and linguistic minority status are consistently referred for special education evaluations at higher rates than their non-minority status peers. Furthermore, the need to follow state and federal guidelines, such as the No Child Left Behind (NCLB) Act of 2001 (NCLB, 2001), has put pressure on public schools to use standardized testing to measure student performance (Duckworth, Quinn, & Tsukayama, 2012).

This motivation to demonstrate satisfactory, academic performance among students is because the NCLB (2001) legislation mandated that federal funding for public schools in the U.S. be contingent upon the performance scores obtained by such standardized assessments (Duckworth et al., 2012). As such, the lower performing students, such as those with intellectual disabilities, have an impact on federal funding for public schools. Decisions regarding placement of students in special education have thusly been affected, as only a small percentage of these students are able to be excluded from these standardized tests. The increased demand for special education evaluations has meant increased use of intelligence test batteries that have been shown to possess measurement bias with CLD individuals. As a result, the disproportionate number of CLD individuals inappropriately placed in special education has continued to be the pattern.

Methods to Assess CLD Individuals

Training programs in psychology have also sought to address issues relating to assessment, counseling, interventions, and consultation with CLD individuals (Jones et al., 2013). Ethical guidelines and competency standards for multicultural training have been appended to national psychology accrediting bodies, including the American Psychological Association (APA), American Counseling Association (ACA), and the National Association for School Psychologists (NASP). The aim of these efforts by accrediting bodies and university training programs is to bring awareness of the issues surrounding the practice of psychology cross-culturally, foster self-awareness of trainees' personal beliefs, values, attitudes regarding culture, and educate these future psychologists with ways they can be an advocate for CLD clients who might otherwise be impacted by multicultural issues at a systemic level (e.g., community or school). The incorporation of multicultural training in university programs provides a foundation for psychologists to be sensitive to issues, such as biased intelligence measures (Jones et al.). What remains to be established, however, is how to best carry out an assessment that minimizes the degree of cultural loading and linguistic demand.

A variety of alternative methods for assessment of CLD individuals have been employed by mental health practitioners across many settings throughout the world. These have included the use of interpreters, testing with translated versions of test batteries, testing in both English and their native language, using "nonverbal" measures, and various other nonstandardized approaches (Cormier et al., 2011; Lakin & Lai, 2012). While these efforts have received much attention and may reduce some of the bias impacting the performance of CLD individuals, language and dialect are not the only factors that necessitate special consideration.

In addition to linguistic influences, one's level of acculturation may potentially be the source of discrepant performances across individuals (Gasquoine, 2009; Van de Vijver & Phalet, 2004). Test bias due to acculturation would favor the cultural group from which the norm sample was derived, as opposed to individuals with little or no experience in the culture (Van de Vijver & Phalet). A test is culturally biased towards the attitudes, values, and beliefs of those responsible for its development (Mpofu & Ortiz, 2009). The vocabulary used, objects used, content chosen for illustrations, structure of item presentation, and tasks chosen to assess specific cognitive abilities are all examples of ways that a test developer's cultural background might influence a test of intelligence.

"Nonverbal" Testing

In response to the recognition of the need for assessment tools and procedures for those who do not speak the language of the examiner or the test, the concept of nonverbal testing was born (Wasserman, 2012). The first attempt at creating a measure of intelligence that resembled the notion of nonverbal testing was by Yerkes (1921) in his development of the Army Beta. However, just as it was apparent in the use of the Army Beta, the goal of having an intelligence test that is void of language or communication requirements is not obtainable (Ortiz et al., 2012). The reason for this is that any test administered directly from examiner to examinee requires some form of communication. In the absence of verbal communication, what is left is nonverbal communication (e.g., gestures, eye contact, facial expressions). As such, the term "nonverbal" is a misnomer and a more appropriate description might be "language-reduced." In addition to the failure to acknowledge the influence of nonverbal communication, use of these language-reduced measures (e.g., Universal Nonverbal Intelligence Test, UNIT; Bracken & McCallum, 1998) neglects the affect of the cultural content that is embedded within the test (Ortiz et al., 2012). Consequently, even if the use of a language-reduced measure decreases the effect of linguistic demand on performance of a CLD individual, issues pertaining to cultural loading of a test on performance remain unaddressed (Byrne et al., 2009; Ortiz et al., 2012).

Native Language

Practitioners are required to assess for the client's cognitive academic language proficiency (CALP) or basic interpersonal communication skills (BICS) for determining the dominant language in which to administer the test (Olvera & Gomez-Cerrillo, 2011). Various measures and screening procedures have been used by school psychologists or other evaluation specialists across the U.S. to complete this assessment. Some school districts may utilize a language survey or screener created by a professional within the district and others may choose to use a standardized measure such as the *Woodcock-*

Muñoz Language Survey (Woodcock & Muñoz-Sandoval, 2001). Regardless of the method used, cases in which the child's dominant language does not match that of the examiner and the tests available necessitate an alternative approach to conducting the evaluation.

A recent approach to the assessment of linguistically diverse individuals has been the use of native language testing (Ortiz et al., 2012). The theory supporting this tactic is that testing a child in his or her own language will minimize or eliminate the negative impact of language demand and language bias on performance. However, several issues arise as a result of the strategies employed by practitioners attempting to use native language testing. One approach to the native language assessment of children who are linguistically diverse is through the use of an interpreter or translator. A problem with this approach is that there is a false sense of security in the validity of such practices. Any time the administration of a test is modified or adapted from its standardized version the reliability and validity are compromised. The performance captured from a modified or adapted administration is no longer comparable to the test's standardization sample and the construct validity, or confidence that the test is measuring the construct intended to be measured, can no longer be assumed (Vazquez-Nuttall et al., 2007).

Another approach is the use of a translated version of a test. Of the issues associated with testing in a child's native language using a translated version of a test or an interpreter, the most important is with regard to the actual translation of the words and ideas from the test's standardized form to the alternate language (Vazquez-Nuttall et al., 2007). Dalen, Jellestad, and Kamaloodien (2007) assert that there is a lack of semantic equivalence when attempting to translate from one language to another; even using the same words from the same language across different cultures (i.e., the United States and South Africa) can result in this problem (p. 616). More so, there are some concepts and words that do not have comparable concepts and words across languages. For example, the English expression, “raining cats and dogs,” “can neither be translated literally nor adapted culturally in any meaningful way in Arabic because it has no real equivalent in the Arabic language or the Arab culture” (Tan et al., 2009, p. 444). The implication for these issues on neurocognitive test performance with a CLD individual is that the scores may actually be a reflection of something other than neurocognitive abilities and there is a risk of the examiner making inappropriate interpretations.

Alternate Forms

One approach by test developers to address the need for cognitive ability tests appropriate for use with CLD individuals has been to create measures normed specifically for use in diverse populations (Oakland, 2009). For example, some tests have included supplemental norms for bilingual individuals, such as the *Wechsler Intelligence Scale for Children – Fourth Edition, Spanish* (WISC-IV Spanish; Wechsler, 2005). However, factors such as language proficiency and level of acculturation, which can further influence validity, are typically not addressed (Flanagan et al., 2013). To elaborate, even in the case that a child demonstrates Spanish language proficiency and is able to take one of these alternate test batteries, the extent to which he or she is familiar with the cultural content of the test may also have an effect on his or her test performance. Attendance only to the influence of linguistic factors may not be sufficient enough to ensure a valid measurement of cognitive abilities. While the prospect of an intelligence battery that minimizes or eliminates the bias associated with cultural or linguistic demand is appealing, the availability of such measures is scarce (Gasquoine, 2009).

Norm Sampling

There is growing recognition within the scientific literature that one of the primary issues associated with cognitive assessment of CLD individuals is the non-representativeness of the norm samples used in the standardization of test batteries (Ortiz et al., 2012). In an attempt to remedy this issue, test developers have increasingly included participants from diverse racial and ethnic backgrounds in their standardization samples. The flaws of this approach are two-fold: the variables chosen to add are not sufficient enough to represent the diverse range of children evaluated with the tests and proportionate sample size still means that the actual number of participants in the minority groups is significantly lower than that of participants in the majority groups (Gasquoine, 2009; Ortiz et al., 2012). Flanagan and colleagues (2013) insist that the inclusion of participants from diverse racial and ethnic backgrounds does not account for all cultural differences that might affect test performance in intelligence measures. Furthermore:

> Representation within a test's norm sample on the basis of racial and ethnic categories is simply not a sufficient proxy for experimental differences that mediate the degree to which an individual is or is not familiar with the culture of the test. (Flanagan et al., 2013, p. 298)

Cross-Battery Approach

To address the issue with content and linguistic biases associated with cognitive tests, researchers urge the use of processing-dependent measures, which aim to decrease the contribution of prior knowledge on performance (Laing & Kamhi, 2003). Utilization of assessment tasks that do not rely upon prior knowledge of a culture or proficiency in a particular language is hypothesized to reveal a more accurate depiction of one's abilities. Indeed, research by Malda, Van de Vijver, and Temane (2010) demonstrated that familiarity with content of a test (i.e., culturally relevant material) predisposes an individual with an advantage during a task of short-term or working memory, regardless of cognitive complexity. Thus, access to a rich network of semantic information about the cultural content included on a test allows for a more efficient encoding and retrieval process for the individual.

Increasing recognition that some subtests from intelligence test batteries are more culturally and linguistically biased than others led to the application of the cross-battery approach for assessment with CLD individuals (Flanagan et al., 2013). Cross-battery assessment involves the amalgamation of cognitive subtests selected specifically to tap into broad and narrow CHC abilities by an examiner to individualize an evaluation. It offers practitioners an alternative method for determining the cognitive strengths and weaknesses of CLD children and adolescents that is grounded in a research-based theoretical framework.

C-LIM/C-LTC

Development

Motivated with the awareness of the issues plaguing valid intellectual assessment of CLD individuals, Samuel Ortiz and Dawn Flanagan (1998) developed the Culture-Language Test Classifications (C-LTC). The C-LTC was developed as a classification system for the influence of two dimensions based on expert consensus, empirical studies, and from data available from published tests (Flanagan et al., 2013). With the C-LTC, subtests were classified (low, moderate, or high) along the dimensions of cultural loading and linguistic demand for a collection of commonly used test batteries, including those applied in this study. Practitioners may use these as a guide to formulating an assessment protocol when concerned about the potential implications a child's cultural or linguistic differences may have on their test performance.

In addition to the C-LTC, Flanagan and colleagues developed the Culture-Language Interpretive Matrix (C-LIM) as a tool to evaluate the degree to which test performance is affected by the cultural loading or linguistic demand of a test battery (Flanagan et al., 2007, 2013). The C-LIM provides a means to evaluate whether performance is a reflection of cultural loading and linguistic demand or if it is a true measurement of ability. On an individual basis, scores are recorded into the C-LIM Microsoft Excel program according to the specified classification assigned in the respective C-LTC. If the pattern of cell averages in the C-LIM derived from these scores shows a decline in value as the degree of cultural loading and linguistic demand increases, the examiner may conclude that the performance is likely influenced by these factors (see Figure 1). Thus, the results do not reliably measure the examinee's true ability, and should be interpreted with caution.

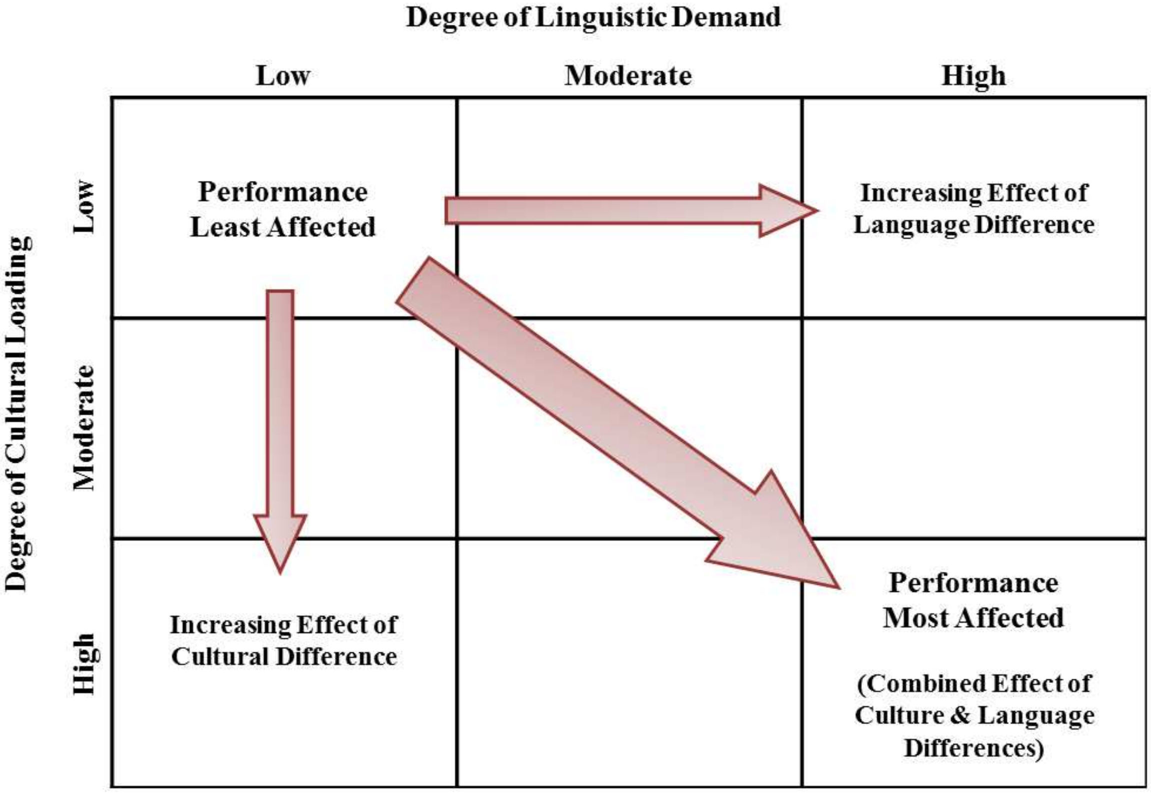

Degree of Linguistic Demand
Low
Moderate
High
Degree of Cultural Loading
Low
Moderate
High
Performance Least Affected
Increasing Effect of Language Difference
Increasing Effect of Cultural Difference
Performance Most Affected
(Combined Effect of Culture & Language Differences)

Figure 1. Pattern of predicted performance for CLD students based on C-LIM classifications (Flanagan et al., 2007)

Associated Research

Since the introduction of the C-LTC and C-LIM, there has been little research to validate their use. One research article described a proposed method for determining the degree of linguistic demand for the *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update* (WJ III COG NU; McGrew & Woodcock, 2001; Cormier et al., 2011). This study was instigated by researchers who recognized the need for a more psychometrically defensible method for ascribing cognitive subtests into the categories of the C-LTC. Few other studies have actually focused on applying the C-LTC and C-LIM with one of the neurocognitive test batteries that have C-LTC classifications in Flanagan and colleagues' (2013) third edition of *Essentials of Cross-Battery Assessment*. Thus, this gap in the literature was intended to be addressed through this study.

Rationale and Purpose of this Study

It is evident that the cognitive and neuropsychological test batteries available and test procedures utilized in current practice have important limitations with regard to their use with CLD individuals. There are also issues pertaining to research of CLD individuals and neuropsychological test development and use. For example, Ford and colleagues (2008) point out that even researchers with the best intentions towards conducting cross-cultural research are influenced by their own beliefs, experiences, attitudes, and values: "From the research focus to the research questions, to the research design and methodology, to the theoretical orientation, and to data interpretation, research is riddled with subjectivity" (p. 83). This viewpoint is shared by Byrne and colleagues (2009), who argue that failure to address issues relating to cross-cultural research and testing practices will lead to distortions, degradations, and limitations in the generalizations of research and interpretations of test performance with CLD individuals. The literature review illustrates the need to further evaluate the use of the C-LTC and C-LIM in the neurocognitive assessment of CLD children and adolescents.

CHAPTER III

METHODS

This chapter will outline the manner with which this study was investigated. Specific information will be provided with regard to the overarching research questions, hypotheses, method of data collection, and statistical analyses used to answer the research questions. This section will conclude with a summary of the design of the study.

Research Design

This study was conducted using an ex post facto criterion group research design for several reasons. First and foremost, there was no random assignment of the participants to any group, precluding the possibility for any type of true experiment (Jackson, Mitchell, & Jolley, 2006). Additionally, the design was not preexperimental, because there was no intervention or treatment implemented by the researcher, nor was it a quasi-experimental design, because there was no comparison of groups of individuals who had experienced some kind of natural disaster or trauma with another that has not had said experience. There was no baseline measure; the study simply consisted of a single posttest procedure. Finally, it was a criterion group design, because the intention was to investigate neurocognitive functioning in a preexisting, mixed, clinical group of children and adolescents. Essentially, each group within the sample must have met specific criteria. This design was the best method for investigating the research questions, because it allowed the researcher to utilize the performance data of children across a variety of ages and clinical diagnoses in a single analysis.

Participants

To investigate the influence of cultural loading and linguistic demand on performance on neuropsychological tests, a sample of children with an array of neurological and developmental disabilities was required. It was important to utilize a mixed clinical sample of individuals with neurological and developmental disabilities, as this is the population for which cognitive and neuropsychological assessment batteries are designed. As such, data were drawn from an archival set of case studies submitted by students of the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program. The archival data set consists of data culled from approximately 1,000 case studies. The data for this study were extracted from this larger data set. Only cases that met the criteria for the study (e.g., utilize the chosen tests, and provide the identified demographic information) were utilized. It was anticipated that the number of cases selected for inclusion in the study would result in about 375 participants, but ultimately the data yielded 520 participants. The participants ranged from eight to 16 years of age, due to the age constraints of the subtests used in the analysis.

Attempts were made to achieve a sample that was comprised of equivalent group sizes for all clinical populations in the data set that was included in this study. Clinical groups available in the archival data included children with attention-deficit/hyperactivity disorder (ADHD), learning disabilities (LD), autism spectrum disorder (ASD), and various others. Inclusion criteria for the present study was dependent upon the number of participants in each sample group; those with the largest numbers were used due to the expectation that they provided better psychometric validity than smaller groups (Cohen & Swerdlik, 2005). Additionally, the aim was to acquire a proportionate ratio of children based on gender and ethnicity. It was important to obtain a sample that was representative of the population for which the study intends to generalize (Gravetter & Forzano, 2009). Ethnicity served as a categorical, independent variable (IV) for this study, with four subgroups making up the multiple levels: Caucasian/White, African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic.

There was a significant portion of the case studies with ethnicity designations from various other backgrounds; however, there was not a sufficient volume of any one designation to comprise another subgroup to be used in this study. The inclusion of an "other" subgroup comprising these various ethnicity designations would only contradict the purpose of this study, which is to demonstrate the influence of varying cultural and linguistic differences on neurocognitive performance. Furthermore, while it would have provided useful discriminative information to also include language of the participant (i.e., English as a first language or English as a second language) as an independent variable, the archival data used in this study did not include this information for the participants.

Measures/Materials

The cognitive measures utilized in this study included the *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update* (WJ III COG NU; McGrew & Woodcock, 2001), the *NEPSY-II: A Developmental Neuropsychological Assessment* (NEPSY-II;

Korkman et al., 2007) and the *Delis-Kaplan Executive Function System* (D-KEFS; Delis et al., 2001a). Selected tests from the *Woodcock-Johnson III Tests of Cognitive Abilities, Normative Update* (WJ III COG NU; McGrew & Woodcock, 2001) were included in this analysis, due to the widespread use of the WJ III COG NU in the assessment of cognitive functioning in the pediatric population. Additionally, selected subtests from the *NEPSY-II: A Developmental Neuropsychological Assessment* (Korkman et al., 2007) and the *Delis-Kaplan Executive Function System* (D-KEFS; Delis et al., 2001a) were included due to their widespread application in the neuropsychological assessment of pediatric clinical populations. The scores derived from each of the tests and subtests for these three test batteries collectively served as the dependent variables (DVs) for this study.

NEPSY-II

The *NEPSY-II* is an integrated test battery that is designed to assess an array of neuropsychological cognitive domains (Korkman et al., 2007). It was revised from the original *NEPSY* (Korkman, Kirk, & Kemp, 1998) with the intent of creating a more clinically useful array of subtests to examine specific neurocognitive processes in children (Matthews, Riccio, & Davis, 2012). Both versions of this measure, as well as the original *NEPS*, are based on Luria's (1980) model of functional neurocognitive processes.

Psychometrics and standardization. The NEPSY-II was reportedly standardized using a random, normative sample (*n* = 1,200) of children and adolescents, between the ages of 3 and 16, who were stratified by age, race/ethnicity, geographic location, and parental education based on the October 2003 United States Census survey (Brooks, Sherman, & Strauss, 2010; Korkman et al., 2007). Additionally, clinical studies used to investigate the reliability and validity of the NEPSY-II included 10 clinical subgroups: ADHD, reading disorder, mathematics disorder, language disorder, intellectual disability, autism, Asperger's disorder, deaf and hard of hearing, emotionally disturbed, and traumatic brain injury. However, there is no evidence to support that each clinical subgroup was comprised of proportionate or representative distributions across all other demographic variables, including race/ethnicity. According to the *Clinical and Interpretive Manual* for the NEPSY-II, participants were excluded from both the normative and the clinical subgroup samples if English was not their primary language or if the child was not fluent in English. However, the subject as to whether participants were bilingual, multilingual, or came from a cultural background that was different from that of the United States was not addressed. The absence of studies to examine potential differences in performance based on additional cultural variables, such as primary language spoken or level of acculturation to the United States, is problematic. Without attending to these additional variables, the validity of scores derived may be diminished and interpretations thus affected.

The NEPSY-II test authors acknowledge that modifications of standardized test procedures prevents the use of normative data and urge examiners to use their "best clinical judgment when evaluating the effect of the modified procedures on the use of the normative data" (Korkman et al., 2007, p. 19). Furthermore, they state that an examiner qualified to use the NEPSY-II should have experience testing children whose cultural and linguistic backgrounds match that of the children they intend to test. An explanation of what this experience should entail is not provided by the authors. In summary, the authors acknowledge that some of the tests included in the NEPSY-II are “highly influenced by language skills” (p. 5), that modification of standardized test procedures may preclude examiners from using normative data, and that those testing children from CLD backgrounds should have the relevant experience with those populations in order to proceed with testing. However, what is missing is any kind of explanation or guidelines for which examiners should reference when testing children who are CLD, aside from suggesting the use of clinical judgment.

The current edition of the test battery was developed for use with children ages three to 16; however, the majority of the subtests included may only be administered to those at least five years of age or older (Matthews et al., 2012). In terms of reliability, the NEPSY-II demonstrates adequate to very high internal consistency among most of the subtests for both the normative and mixed clinical sample used to validate it (Brooks et al., 2010; Korkman et al., 2007). Marginal or low internal consistency reliability coefficients were observed for each age group with the standardization sample for Word Generation Total Score, Memory for Faces Total Score, Design Fluency, and Narrative Memory Free Recall. Within the mixed clinical sample, low internal reliability was only found for Word List Interference Recall. Brooks and colleagues (2010) also report the overall test-retest reliability of the NEPSY-II to be adequate to high.

A review of the validity of the NEPSY-II requires comparisons of the constructs and content presumed to be assessed in its subtests with other test batteries to evaluate the underlying theoretical framework (Brooks et al., 2010). Medium to large correlations were reported between the NEPSY-II subtests and theoretically related subtests from test batteries, such as the *Wechsler Intelligence Scale for Children – Fourth Edition* (WISC-IV; Wechsler, 2003), *Wechsler Nonverbal Scale of Ability* (WNV; Wechsler & Naglieri, 2006), *Differential Ability Scales – Second Edition* (DAS-II; Elliot, 2007), *Wechsler Individual Achievement Test – Second Edition* (WIAT-II; Wechsler, 2001), *Children's Memory Scale* (CMS; Cohen, 1997), and *Delis-Kaplan Executive Functions System* (D-KEFS; Delis et al., 2001a). Strong correlations between measures thought to assess a similar construct provide support that the subtests that make up the NEPSY-II have validity in measuring the neurocognitive functions for which they were intended.

The NEPSY-II *Clinical and Interpretive Manual* includes correlations between selected NEPSY-II scaled scores and selected D-KEFS subtest scores (Korkman et al., 2007). Because the D-KEFS is a collective battery of nine stand-alone subtests of executive functioning, convergent validity for the NEPSY-II would be supported by strong, positive relationships between subtests of the D-KEFS and subtests of the NEPSY-II that are purported to measure attention and executive functioning. Divergent validity can be established by strong, negative relationships between subtests of the D-KEFS and subtests of the NEPSY-II that are reported to measure other, unrelated cognitive processes.

According to the data reported in the NEPSY-II *Clinical and Interpretive Manual*, the NEPSY-II subtests in the Attention and Executive Functioning Domain (Animal Sorting, Auditory Attention and Response Set, Clocks, and Inhibition) have low to moderate correlations with the D-KEFS scores selected for this analysis (Korkman et al., 2007). Within this domain, the NEPSY-II Inhibition subtest demonstrates consistently moderate relationships with the D-KEFS Color-Word Interference subtest and Inhibition/Switching subtest. Word List Interference Repetition and Recall of the NEPSY-II also show a relationship with the Completion Time scores of those two D-KEFS subtests. The NEPSY-II subtest of Block Construction, on the Visuospatial Processing Domain, shows moderate correlations with Trail Making and Design Fluency from the D-KEFS, both of which require motor processing abilities. There are also moderate correlations between the NEPSY-II Visuomotor Precision Total Completion Time score of the Sensorimotor Domain and the D-KEFS subscores from Trail Making (Combined Number Sequencing and Letter Sequencing Completion Time), and Verbal Fluency (Letter Fluency and Category Fluency Total Correct Response). The NEPSY-II Visuomotor Precision Combined Scaled score also shows a moderate correlation with the D-KEFS Design Fluency Total Correct score and the Color-Word Interference Inhibition/Switching Total Errors score. Finally, within the Language Domain of the NEPSY-II, moderate to high correlations are shown between Word Generation and the D-KEFS subtest of Verbal Fluency. A moderate relationship is also observed between Comprehension of Instructions from the NEPSY-II and the D-KEFS subtests of Color-Word Interference and Verbal Fluency.

Further support for the validity of the NEPSY-II is demonstrated by the ability for relevant neurocognitive subtests and domains to effectively discriminate among clinical groups, as well as from a sample of matched controls (Brooks et al., 2010). For example, clinical groups that should characteristically demonstrate impaired performance on a particular cognitive domain (i.e., Autistic Disorder and Social Perception) have shown just this pattern. The subtests administered to the clinical subgroups included: Animal Sorting, Auditory Attention and Response Set, Clocks, Inhibition, Statue, Comprehension of Instructions, Phonological Processing, Speeded Naming, Memory for Designs, Memory for Faces, Narrative Memory, Sentence Repetition, Word List Interference, Fingertip Tapping, Visuomotor Precision, Affect Recognition, Theory of Mind, Arrows, Block Construction, Design Copying, Geometric Puzzles, and Picture Puzzles. Some of these subtests, however, were not administered to all clinical subgroups. Therefore, the clinical sensitivity of some subtests for certain clinical subgroups remains unknown as a result of these validity studies.

Scores derived. There are several types of scores that can be derived from the NEPSY-II (Korkman et al., 2007). Because this measure is comprised of a collection of subtests chosen to fit within a theoretically based model of neurocognitive functioning, it does not offer any kind of comprehensive score for the combination of all subtests administered. However, scores that can be obtained on the NEPSY-II include primary scores for individual subtests, process scores, contrast scores, as well as domain scores. Many of the subtests from the NEPSY-II provide more than one primary score, allowing for the examiner to isolate performance for specific parts of a subtest. For the purposes of this study, the subtest scaled scores were utilized as quantitative dependent variables, while those providing only percentile ranks or cumulative percentile ranks were excluded. Subtests included in this study were Finger Tapping – Dominant Hand, Imitating Hand Positions, Visuomotor Precision, Response Set, Block Construction, Arrows, Geometric Puzzles, Picture Puzzles, Speeded Naming, Comprehension of Instructions, Phonological Processing, Word List Interference, Narrative Memory – Free Recall, Memory for Faces – Immediate and Delayed, Memory for Designs – Immediate and Delayed, Memory for Names – Immediate and Delayed, Animal Sorting – Combined, Design Fluency, and Inhibition – Switching/Combined. Statue was not be included in this study because it has only been validated for use with children age five to six years old, and there were not a large number of participants who took this subtest. Table 1 provides a short description for each of the subtests from the NEPSY-II that were included in this study.

Table 1

NEPSY-II Subtests Included in this Study and Their Descriptions

Subtest	Description
Finger Tapping – Dominant Hand	Assesses fine motor coordination and motor programming of the fingers.
Imitating Hand Positions	Assesses the ability to imitate the examiner's static hand position, using visuospatial analysis, motor programming, and kinesthetic feedback.
Visuomotor Precision	Assesses graphomotor speed and accuracy.
Auditory Attention and Response Set	This subtest has two parts. The first part, Auditory Attention, assesses selective and sustained auditory attention. The second part, Response Set, maintains the selective and sustained attention requirements of Part 1 and adds a shifting attention component.
Block Construction	Assesses visuospatial constructional ability for 3-dimensional representation.
Arrows	Assesses the ability to judge line orientation on a visual stimulus.

(Continued)

Geometric Puzzles	Assesses mental rotation, visuospatial analysis, and attention to detail.
Picture Puzzles	Assesses nonmotor aspects of visual perception from photos of everyday scenes and objects.
Speeded Naming	Assesses rapid access to and production of names of colors, shapes, letters, numbers, or sizes.
Comprehension of Instructions	Assesses the ability to perceive, process, and execute oral instructions of increasing syntactic complexity.
Phonological Processing	Assesses awareness and analysis of auditory phonological elements of words.
Word List Interference	Assesses verbal working memory, repetition, and word recall following interference.
Narrative Memory – Free Recall	Assesses narrative memory under free recall, cued recall, and recognition conditions.
Memory for Faces	Assesses immediate and delayed visual memory of facial features, as well as face discrimination and recognition.
Memory for Designs	Assesses spatial and visual, nonfigurative content memory for novel visual material. The delayed task assesses long-term visual-spatial memory.
Memory for Names	Assesses name learning and delayed memory for names.
Animal Sorting - Combined	Assesses a child's ability to formulate basic concepts, to transfer those concepts into action, and to shift from one concept to another.
Design Fluency	Assesses ability to initiate and produce unique designs.
Inhibition Switching – Combined	A timed test that assesses the ability to inhibit automatic responses in favor of novel responses.

D-KEFS

The *Delis-Kaplan Executive Function System* (D-KEFS) is a neuropsychological assessment battery that was developed to measure executive functioning for individuals ages eight to 89 (Delis, Kaplan, & Kramer, 2001c). It is typically utilized for assessment and diagnostic purposes in children and adults with neurodevelopmental and neurodegenerative disorders affecting the frontal cortex (Dugbartey, 2011). The domains of executive functioning that the D-KEFS was designed to assess include, “flexibility of thinking, inhibition, problem solving, planning, impulse control, concept formation, abstract thinking, and creativity in both verbal and spatial modalities” (Homack et al., 2005, p. 599). A brief description for each of the subtests that were included in this study can be found in Table 2. The Proverbs subtest was not used in this study because there is no normative data available for children younger than 16 years old.

Psychometrics and standardization. The D-KEFS was reportedly standardized using a stratified normative sample (*n* = 1,750) of individuals from age 8 to 89 years, based on the United Stated Census of 2000 (Baron, 2004; Delis et al., 2001c; Homack et al., 2005). This normative sample was stratified according to age, sex, race/ethnicity, years of education (parental education for those ages 8-19 years), and geographic region. A separate study to standardize the two alternate forms available with the D-KEFS included 286 individuals, ages 16-89 years.

Validity and reliability of the D-KEFS is discussed in the *Technical Manual* (Delis et al., 2001c). The authors report that many of the tests included, although modified from their original format, were drawn from measures that had already established their validity. However, very few studies have been conducted since the introduction of the D-KEFS to confirm the validity of the tests in their current state (Shunk, Davis, & Dean, 2006). Of those that have been published, the D-KEFS demonstrated convergent validity with moderate correlations with the *Wisconsin Card Sorting Test* (WCST; Berg, 1948; Heaton, Chelune, Talley, Kay, & Curtiss, 1993) and demonstrated discriminant validity with a lack of correlation with the *California Verbal Learning Test – Second Edition* (CVLT-II; Delis, Kaplan, Kramer, & Ober, 2000).

Table 2

D-KEFS Subtests Included in this Study and Their Descriptions

Subtest	Description
Trail-Making Test – Condition 4	This is a visual-motor task designed to measure flexibility in thinking. It is comprised of five conditions that are intended to help the administrator discriminate and interpret levels of functioning across the cognitive processes of visual scanning, visual attention, basic numeric sequential processing, letter sequential processing, shifting attention/cognitive flexibility/divided attention, and motor functions.
Word Context	This subtest is a measure of verbal abstract thinking and deductive reasoning. The examinee is required to discover the meanings of a made-up word based on its use in five clue sentences, which progressively provide more detailed information about the target word's meaning.
Color-Word Interference – Condition 4 (Inhibition/Switching)	Assesses the inhibition of the natural inclination to respond in a certain way in order to respond in accordance with a set of defined rules. It is similar to the *Stroop Color-Word Test*, which measures the Stroop Effect.
Verbal Fluency – Condition 1 (Letter Fluency)	This is a test that measures the fluency of verbal responses. It assesses the ability to quickly produce verbal responses in accordance with a set of rules and under a specified time constraint.
Design Fluency – Condition 1 (Filled Dots)	This subtest provides a measure of fluency in the spatial domain. The examinee is required to produce as many differing designs as possible using a series of dots and rules as a guide within a delineated time period.

Validity studies have also shown that the D-KEFS has moderate sensitivity for differentiating among some clinical populations, including fetal alcohol syndrome (FAS), schizophrenia, Alzheimer's disease, Huntington's disease, Parkinson's disease, frontal lobe lesions, and other groups with various types of brain damage (Delis et al., 2001c; Homack et al., 2005; Shunk et al., 2006). According to the intercorrelations reported in the *Technical Manual*, the tests of the D-KEFS have low relationships between them, indicating that they are each measuring distinct aspects of executive functioning (Delis et al., 2001c).

Methods to assess the reliability of the D-KEFS included evaluating the internal consistency with split-half coefficients, test-retest reliability, and alternate form reliability (Delis et al., 2001c; Homack et al., 2005; Shunk et al., 2006). Significant variability was observed for the split-half reliability coefficients across tests, age groups, and among conditions for each test. These split-half coefficients ranged from low (.10) to high (.90). Similarly, the test-retest reliability coefficients showed variability, ranging from low (.24) to high (.76). However, the majority of these fell in the moderate range. Alternate form reliability was found to show a pattern much like that of the split-half coefficients, with a wide range of correlations, but with the majority falling in the moderate range. The test authors argue that such variability in performance on the D-KEFS is expected, given the range of complex and interdependent neurocognitive processes being measured.

The authors of the D-KEFS acknowledge that the instructions for the tests are lengthy and complex; however, an explanation for this was provided in the *Examiner's Manual* (Delis, Kaplan, & Kramer, 2001b). The need for complex task instructions is argued to be because the D-KEFS was developed specifically for assessing higher-level neurocognitive abilities. Homack and colleagues (2005) add that, "while normal individuals may find the instructions unnecessarily complex and repetitive, these instructions may prove helpful for more impaired patients" (p. 602). However, while the D-KEFS is said to measure executive functions in both verbal and nonverbal modalities, the issue of linguistic demand due to task instructions was not addressed. Furthermore, there is no explanation provided in the D-KEFS manuals as to recommendations for its use in the assessment of CLD individuals.

Scores derived. The scores that can be obtained from the D-KEFS include scaled scores and cumulative percentile ranks for each test, as well as process scores, which include contrast scores, combined scaled scores, response accuracy, error rates, and response latency (Delis et al., 2001b; Homack et al., 2005; Maricle & Avirett, 2012). Because this assessment battery is comprised of nine stand-alone tests, it does not allow for the calculation of an overall performance score (Shunk et al., 2006). Instead, the raw scores obtained for each of the subtests can be converted into individual standard scores and process scores may also be obtained. To assist in the scoring process, the D-KEFS includes a Scoring Assistant software program (Delis et al.).

The D-KEFS subtests included in this study each provide a variety of scores, all of which explain distinct facets of performance (Delis et al., 2001b). Depending on the score, the examiner may be reviewing information pertaining to different neurocognitive processes. For the purposes of this study, it was necessary to include only the subtest scores that were most relevant and useful in discriminating performance for CLD individuals. Scaled scores from Trail-Making Test Condition 4, Design Fluency – Filled

Dots, Color-Word Interference Condition 4 (Inhibition/Switching), Word Context, and Verbal Fluency – Letter Fluency were included in this study.

WJ III COG NU

The *Woodcock-Johnson III Tests of Cognitive Abilities Normative Update* (WJ III COG NU; Woodcock, McGrew, & Mather, 2001, 2007) is a comprehensive test battery designed to assess general intellectual ability for individuals two to 95 years of age. It is comprised of 20 norm-referenced tests across two batteries: standard and extended. The *Woodcock-Johnson III Diagnostic Supplement to the Tests of Cognitive* Abilities (WJ III NU DS; Woodcock et al., 2003, 2007) provides an additional 11 co-normed tests that may be used to increase the diagnostic utility of the WJ III COG NU. However, the WJ III NU DS was not included in this study due to an insufficient volume of cases with its test scores in the archival sample from which the data was culled. The content of the WJ III COG NU is based on the CHC theory, and most of the tests measure broad and narrow abilities that fall within this framework. However, some tests measure other aspects of cognitive functioning. The broad CHC factors that this test assesses include Fluid Reasoning (*Gf*), Comprehension Knowledge (*Gc*), Visual Processing (*Gv*), Working Memory (*Gsm*), Processing Speed (*Gs*), Auditory Processing (*Ga*), and Long-Term Storage and Retrieval (*Glr*).

Fluid Reasoning (*Gf*) measures the ability to solve problems using new information by drawing inferences, forming concepts, or classifying and relating particular objects or entities. Tasks that assess *Gf* may require one to form abstract concepts or rules to solve a problem, compared to concrete, predictable rules.

Comprehension Knowledge (*Gc*) provides a measure of the overall pool of general information acquired through various sources (i.e., family, school, and daily life). It reveals the depth and breadth of acquired knowledge and understanding about topics such as people, facts, everyday life, and the world around oneself.

Visual Processing (*Gv*) is the ability to mentally manipulate visual images, analyze visual information, or detect discrepancies with visual stimuli. It is the capacity to process and organize visual information, as well as interpret or reassemble information perceived visually. Working Memory (*Gsm*) encompasses the mental capacity and ability to transform information that is stored for short periods of time. Further, it is the ability to successfully retain and recall information that has been stored and altered in some way.

Processing Speed (*Gs*) measures the ability to rapidly and fluently scan, recognize, and identify differences or similarities between stimuli. Efficient processing speed seems to be necessary for smooth execution of all other cognitive functions (Miller, 2013). Auditory Processing (*Ga*) measures one's ability to comprehend, discriminate, synthesize, and process auditory stimuli. Long-Term Storage and Retrieval (*Glr*) provides a measure of the ability to learn new concepts or store novel information in long-term memory for later retrieval on an associated task. Table 3 includes a list and descriptions of the WJ III COG NU tests that were selected for inclusion in the current study.

Table 3

WJ III COG NU Tests Selected for Inclusion in this Study and Their Corresponding Descriptions

Test	Description
Verbal Comprehension	A measure of lexical knowledge and language development that involves object recognition and re-identification, semantic activation, access, and matching, and verbal analogical reasoning.
Visual-Auditory Learning	A measure of associative memory that requires paired-associative encoding via directed spotlight attention, as well as storage and retrieval.
Spatial Relations	A measure of visualization and spatial relations requiring visual feature detection, matching, and manipulation of visual images in space.
Sound Blending	Assesses auditory processing via the synthesis of acoustic, phonological elements in immediate awareness, matching the sequence of elements to stored lexical entries, and lexical activation and access.
Concept Formation	A test of fluid reasoning involving rule-based categorization, switching, and induction/inference.
Visual Matching	Assesses processing speed requiring speeded visual perception and matching.
Numbers Reversed	A measure of working memory via holding a span of numbers in immediate awareness while reversing the sequence.
Incomplete Words	Assesses auditory processing involving the analysis of a sequence of acoustic, phonological elements in immediate awareness, as well as activation of a stored representation of the word from an incomplete set of phonological features.
Auditory Working Memory	A measure of working memory requiring the recoding of acoustic, verbalizable stimuli held in immediate awareness.
General Information	This is a measure of comprehension-knowledge using semantic activation and access to declarative generic knowledge.

(Continued)

Retrieval Fluency	This measure of long-term storage and retrieval provides an assessment of ideational fluency and naming facility via recognition, fluent retrieval, and oral production of examples of a semantic category.
Picture Recognition	A test of visual-spatial thinking through the formation of iconic memories and matching of visual stimuli to stored representations.
Analysis-Synthesis	This is a measure of general sequential reasoning and quantitative reasoning using deduction and algorithmic reasoning.
Decision Speed	A measure of semantic processing speed through the location and circling pairs of pictures in a row that are conceptually the most similar.
Memory for Words	A measure of short-term memory that provides assessment of auditory memory span through the formation of echoic memories and verbalizable span of echoic store.
Rapid Picture Naming	This is a test to assess processing speed and naming facility requiring speed/fluency of retrieval and oral production of recognized objects.
Pair Cancellation	A test of processing speed requiring controlled, focal attention/concentration and vigilance.

Psychometrics and standardization. Information presented in the *Technical Manual* for the WJ III series (WJ III COG NU, WJ III NU DS, and the *Woodcock-Johnson III Tests of* Achievement, WJ III ACH NU; Woodcock, McGrew, & Mather, 2001) provides an overview of the standardization procedures used during its development, as well as demographic information of the sample obtained for the normative update (McGrew, Schrank, & Woodcock, 2007). This sample of 8,782 participants (ages 12 months to 90+ years) was recalculated to be based on the 2005 U.S. Census statistics (U.S. Census Bureau, 2005) with the intention to utilize a nationally representative sample across the demographic variables of geographic region (Northeast, Midwest, South, and West), community size (Urbanized Area > 50,000, Urban Cluster ≥ 25,000, Rural Area < 2,500), sex, race, Hispanic and non-Hispanic, type of school (public, private, home), education of parents, and native versus foreign born. A clinical sample of 3,702 participants was also provided in the *Technical Manual* with clinical data from 11 special population groups: Anxiety Spectrum Disorders, Attention-Deficit/Hyperactivity Disorder, Autism Spectrum Disorders, Depressive Spectrum Disorders, Giftedness, Head Injury, Language Disorders, Mathematics Disorder, Mental Retardation, Reading Disorders, and Written Language Disorders.

The purpose for the normative update of the WJ III (referred hereafter as WJ III NU) was to provide a more current source of comparison to the U.S. population (McGrew et al., 2007). Part of the process included a recalculation of standardization norm data and the addition of clinical data for special populations using updated statistical procedures, but some changes were also made with regard to test items. The *Technical Manual* states that bias and sensitivity reviews were conducted to eliminate certain items that did not seem to be fair for those of cultural or linguistic minority backgrounds. This step was intended to help the test authors in improving construct validity and reduce measurement of extraneous or construct-irrelevant variance.

Extensive research regarding the validity of the WJ III NU has demonstrated that it reliably conforms to the CHC theory of cognitive abilities from which its foundation is derived (McGrew et al., 2007; Schrank, Miller, Wendling, & Woodcock, 2010). Test and cluster score intercorrelations reported in the WJ III NU *Technical Manual*, as well as comparisons between the WJ III NU and other measures of intellectual functioning, provide substantial support for convergent, discriminant, and construct validity of this measure. What remains a weakness for the WJ III NU is its lack of validity studies for the use of it with clinical groups of neurological or neurodevelopmental disorders (Schrank et al., 2010).

Reliability statistics for the WJ III NU were obtained through the use of the split-half procedure with the Spearman-Brown correction and Rasch analysis procedures (McGrew et al., 2007). Median test reliabilities reported in the *Technical Manual* generally fall in the high range, with the majority of the reliability coefficients being .80 or higher. For the WJ III NU cluster scores, reliability coefficients show even stronger median reliabilities, with most falling at .90 or higher. Test-retest reliability correlations reported in the *Technical Manual* also indicate strong reliability for the WJ III NU.

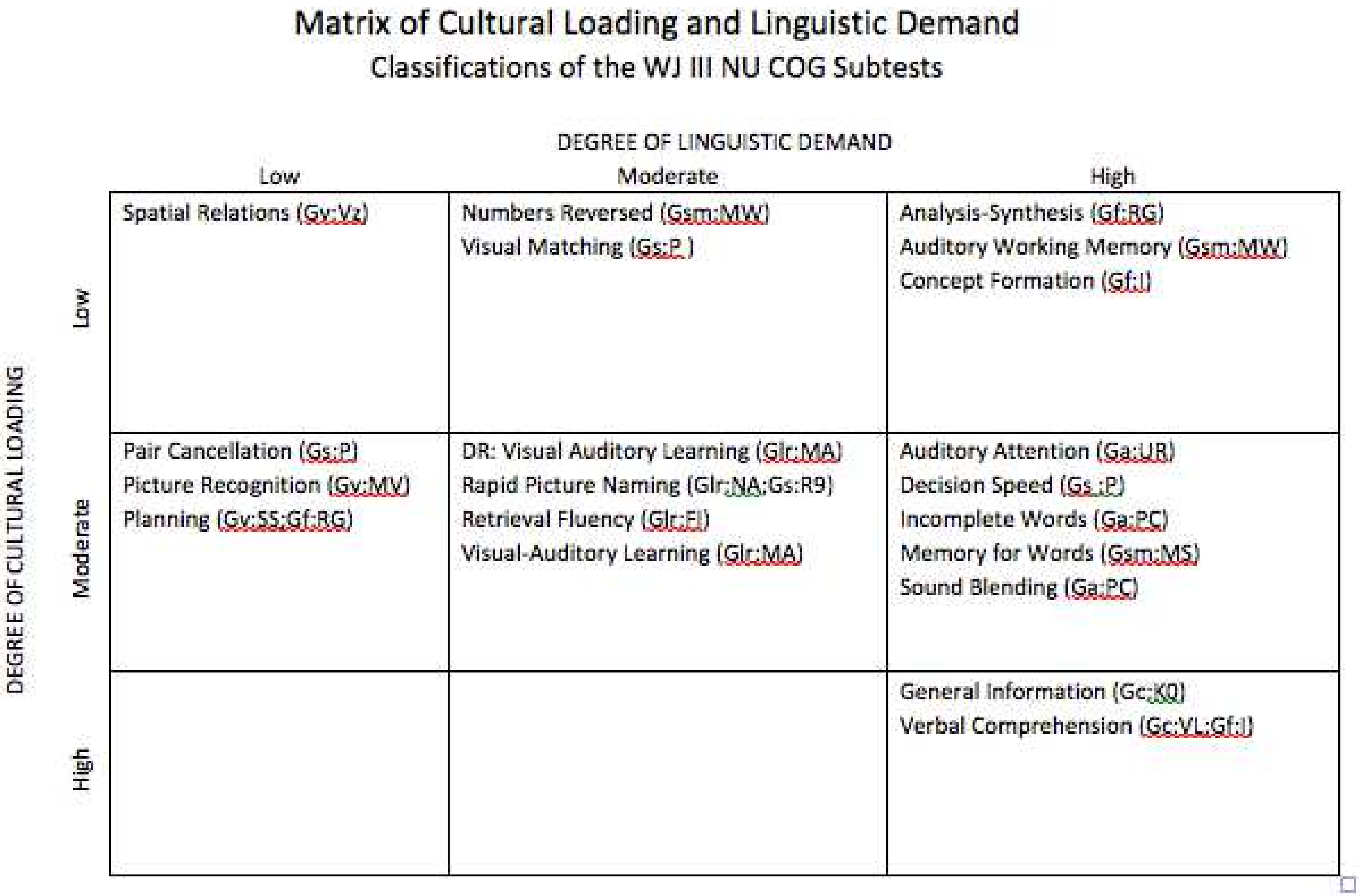
The authors of the WJ III NU recognize the issues relating to assessment with CLD populations. As such, they have offered several tools and suggestions for approaching a more valid assessment of their cognitive abilities. For instance, examiners have been afforded the option to give credit for correct answers given from their CLD examinees in languages other than English (Schrank et al., 2010). There is also a test battery option for English-dominant bilingual individuals. Seven subtests were selected based on their low language demand to collectively provide this General Intellectual Ability-Bilingual (GIA-Bil) option. Furthermore, the *Batería III Woodcock-Muñoz* (Muñoz-Sandoval, Woodcock, McGrew, & Mather, 2004, 2007) is a parallel Spanish version to the WJ III NU that may be used in the assessment of Spanish-speaking individuals. While these supplemental devices may assist examiners in assessment with reduced linguistic demands or bias, there is still the issue of addressing other culturally-related bias, such as level of acculturation.

Scores derived. Scores and profiles that can be obtained from the WJ III NU include: standard scores, percentile ranks, W scores, age equivalents, grade equivalents, Relative Proficiency Index (RPI) scores, CALP levels, *z* scores, T scores, normal curve equivalents (NCE), Stanines, and percentile rank/standard score profiles, or age/grade profiles (McGrew et al., 2007). All scores are obtained through the use of a computerized scoring program that is included with the WJ III NU, called the Woodcock-Johnson III Normative Update Compuscore and Profiles Program (Compuscore; Schrank & Woodcock, 2008); hand-scoring for the WJ III NU is not available. Another scoring option for examiners is the Woodcock Interpretation and Instructional Interventions Program (WIIIP; Schrank, Wendling, & Woodcock, 2008), which performs the same tasks as the Compuscore program, as well as the addition of narrative reports that link scores obtained to educational interventions and accommodations. The standardized scores used in this study were drawn from all subtests in the Standard and Extended Batteries with the exception of the Planning subtest, due to lack of sufficient data in the archival sample. As the purpose of this study was to differentiate performance differences between CLD individuals, cases with neurocognitive profiles consistent with an intellectual disability were excluded from the data analysis.

Procedures

Armed with the empirical evidence that tests of intelligence and neurocognitive functioning are inherently culturally loaded and linguistically biased, Flanagan and colleagues (2007, 2013) created the Cultural-Language Test Classifications (C-LTC) and the Cultural-Language Interpretive Matrix (C-LIM). The aim was to develop a system for organizing the degree to which performance on specific subtests might be influenced by cultural and linguistic factors, as well as a method for determining how to interpret findings when assessing diverse individuals.

The categories of cultural loading (low, moderate, high) and linguistic demand (low, moderate, high), as defined by Flanagan and colleagues (2007, 2013), were used as a frame of reference for the statistical analyses performed in this study. This enabled the researcher to evaluate any differences in performance along these dimensions. Figures 2, 3, and 4 show the C-LTC classifications for the WJ III COG NU, NEPSY-II, and D-KEFS, respectively. These C-LTC classifications assisted in examining whether observed differences in performance across ethnicity groups of the sample were more evident in those subtests deemed high for cultural loading and linguistic demand.

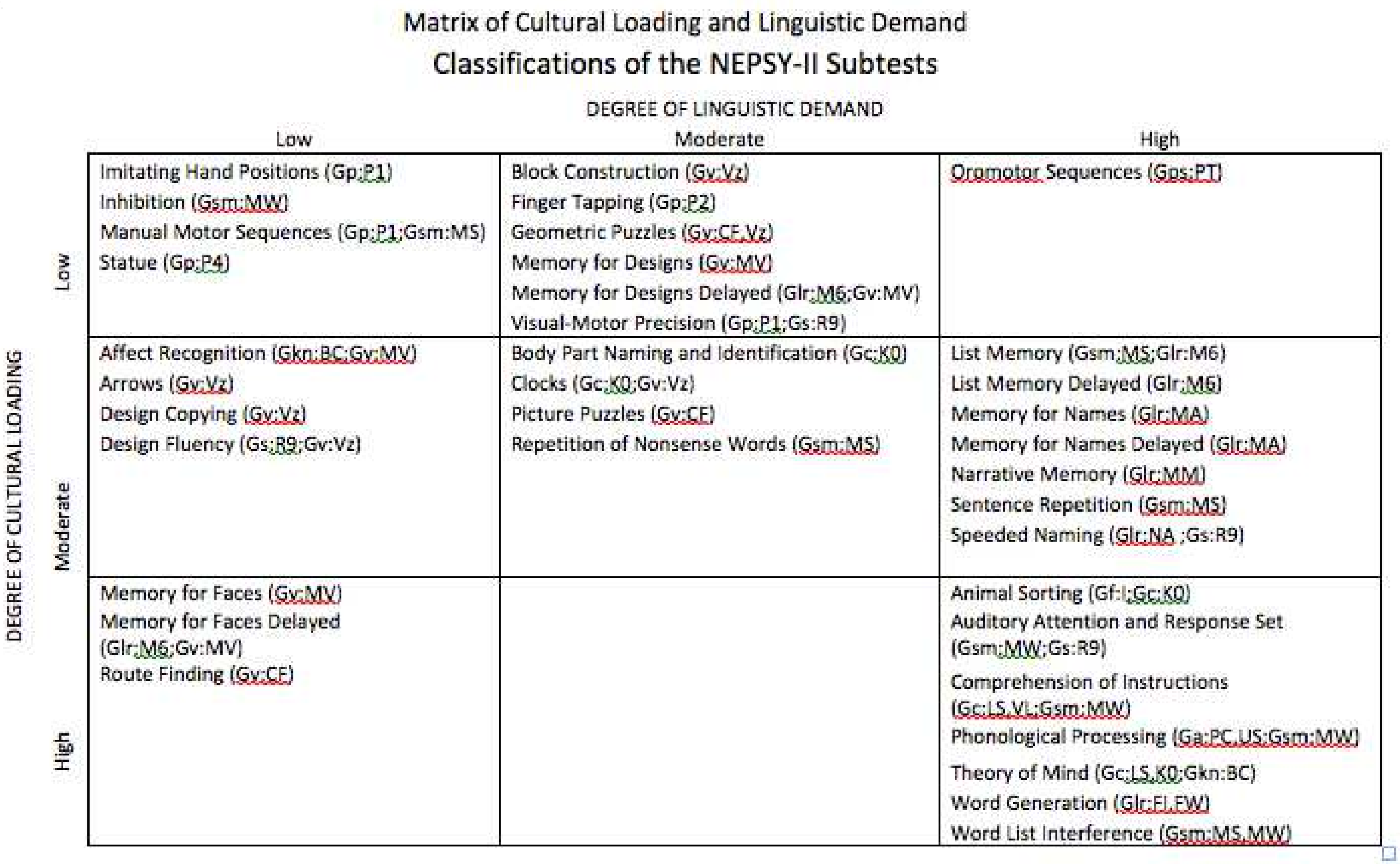


Matrix of Cultural Loading and Linguistic Demand

Classifications of the WJ III NU COG Subtests

		DEGREE OF LINGUISTIC DEMAND		
		Low	Moderate	High
DEGREE OF CULTURAL LOADING	Low	Spatial Relations (Gv:Vz)	Numbers Reversed (Gsm:MW) Visual Matching (Gs:P)	Analysis-Synthesis (Gf:RG) Auditory Working Memory (Gsm:MW) Concept Formation (Gf:I)
	Moderate	Pair Cancellation (Gs:P) Picture Recognition (Gv:MV) Planning (Gv:SS;Gf:RG)	DR: Visual Auditory Learning (Glr:MA) Rapid Picture Naming (Glr:NA;Gs:R9) Retrieval Fluency (Glr:FI) Visual-Auditory Learning (Glr:MA)	Auditory Attention (Ga:UR) Decision Speed (Gs :P) Incomplete Words (Ga:PC) Memory for Words (Gsm:MS) Sound Blending (Ga:PC)
	High			General Information (Gc:K0) Verbal Comprehension (Gc:VL;Gf:I)

Figure 2. Culture-Language Test Classifications: Woodcock-Johnson III Normative Update Tests of Cognitive Abilities (Flanagan et al., 2007)



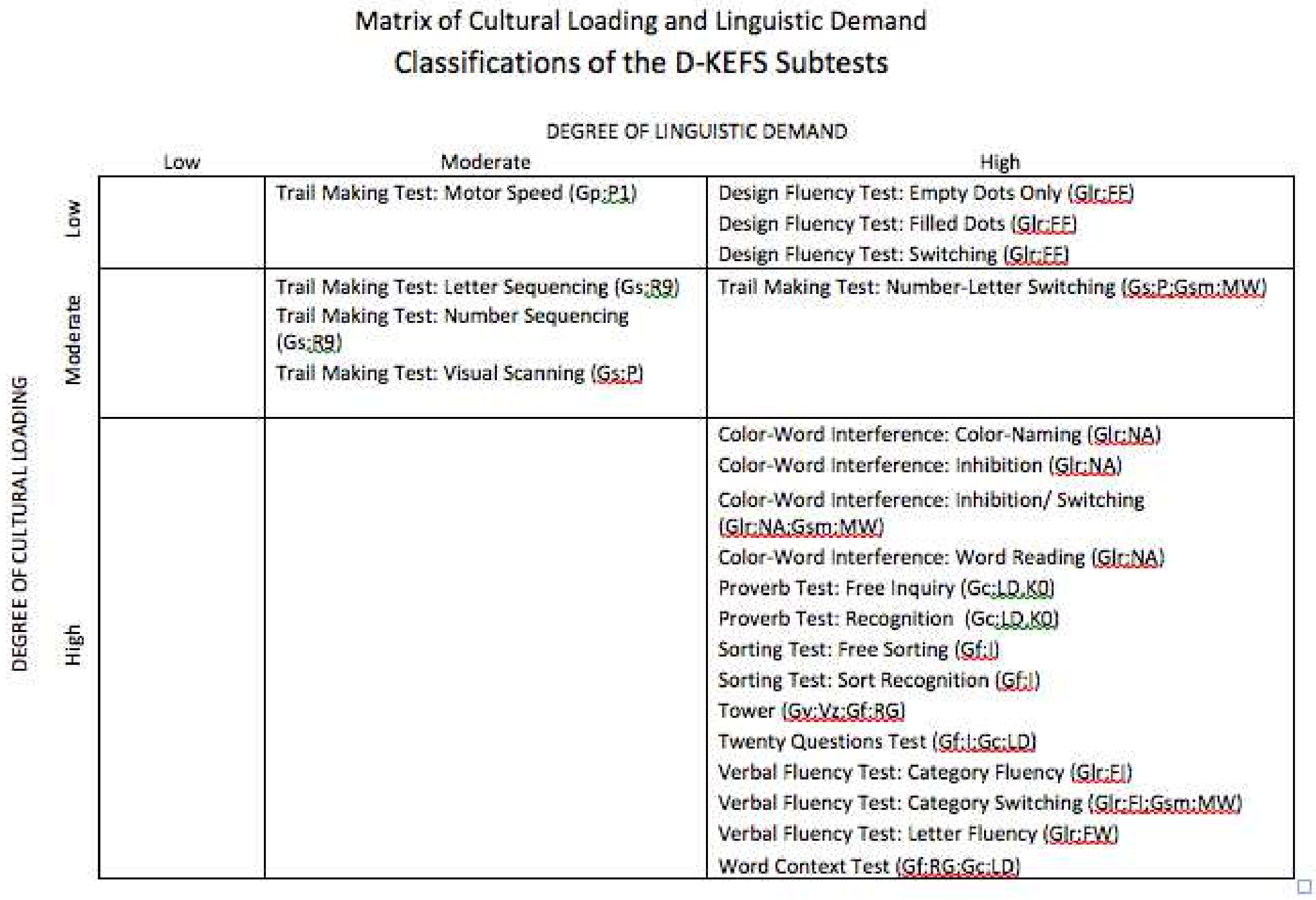
Matrix of Cultural Loading and Linguistic Demand

Classifications of the NEPSY-II Subtests

DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING	Low	Moderate	High
Low	Imitating Hand Positions (Gp:P1) Inhibition (Gsm:MW) Manual Motor Sequences (Gp:P1;Gsm:MS) Statue (Gp:P4)	Block Construction (Gv:Vz) Finger Tapping (Gp:P2) Geometric Puzzles (Gv:CF,Vz) Memory for Designs (Gv:MV) Memory for Designs Delayed (Glr:M6;Gv:MV) Visual-Motor Precision (Gp:P1;Gs:R9)	Oromotor Sequences (Gps:PT)
Moderate	Affect Recognition (Gkn:BC;Gv:MV) Arrows (Gv:Vz) Design Copying (Gv:Vz) Design Fluency (Gs:R9;Gv:Vz)	Body Part Naming and Identification (Gc:K0) Clocks (Gc:K0;Gv:Vz) Picture Puzzles (Gv:CF) Repetition of Nonsense Words (Gsm:MS)	List Memory (Gsm:MS;Glr:M6) List Memory Delayed (Glr:M6) Memory for Names (Glr:MA) Memory for Names Delayed (Glr:MA) Narrative Memory (Glr:MM) Sentence Repetition (Gsm:MS) Speeded Naming (Glr:NA ;Gs:R9)
High	Memory for Faces (Gv:MV) Memory for Faces Delayed (Glr:M6;Gv:MV) Route Finding (Gv:CF)		Animal Sorting (Gf:I;Gc:K0) Auditory Attention and Response Set (Gsm:MW;Gs:R9) Comprehension of Instructions (Gc:LS,VL;Gsm:MW) Phonological Processing (Ga:PC,US;Gsm:MW) Theory of Mind (Gc:LS,K0;Gkn:BC) Word Generation (Glr:FI,FW) Word List Interference (Gsm:MS,MW)

Figure 3. Culture-Language Test Classifications: NEPSY-II (Flanagan et al., 2013)



Matrix of Cultural Loading and Linguistic Demand

Classifications of the D-KEFS Subtests

			DEGREE OF LINGUISTIC DEMAND	
		Low	Moderate	High
DEGREE OF CULTURAL LOADING	Low		Trail Making Test: Motor Speed (Gp:P1)	Design Fluency Test: Empty Dots Only (Glr:FF) Design Fluency Test: Filled Dots (Glr:FF) Design Fluency Test: Switching (Glr:FF)
	Moderate		Trail Making Test: Letter Sequencing (Gs:R9) Trail Making Test: Number Sequencing (Gs:R9) Trail Making Test: Visual Scanning (Gs:P)	Trail Making Test: Number-Letter Switching (Gs:P:Gsm:MW)
	High			Color-Word Interference: Color-Naming (Glr:NA) Color-Word Interference: Inhibition (Glr:NA) Color-Word Interference: Inhibition/ Switching (Glr:NA:Gsm:MW) Color-Word Interference: Word Reading (Glr:NA) Proverb Test: Free Inquiry (Gc:LD,K0) Proverb Test: Recognition (Gc:LD,K0) Sorting Test: Free Sorting (Gf:I) Sorting Test: Sort Recognition (Gf:I) Tower (Gv:Vz:Gf:RG) Twenty Questions Test (Gf:I:Gc:LD) Verbal Fluency Test: Category Fluency (Glr:FI) Verbal Fluency Test: Category Switching (Glr:FI:Gsm:MW) Verbal Fluency Test: Letter Fluency (Glr:FW) Word Context Test (Gf:RG:Gc:LD)

Figure 4. Culture-Language Test Classifications: Delis-Kaplan Executive Function System (Flanagan et al., 2013)

Research Questions and Hypotheses

A review of relevant literature demonstrates a need to investigate the impact of cultural loading and linguistic demand on the measures utilized by mental health practitioners to assess their clients. The following research questions were posed:

1. Do differences in performance exist based upon ethnicity for the three standardized (WJ III COG NU, NEPSY-II, D-KEFS) measures of neurocognitive functioning?
2. How much of the difference in performance is attributed to ethnicity?

Based upon previous literature and research, it was expected that the measures investigated in this study would reveal statistically significant differences in performance across ethnicity groups. Further, it was expected that those who reported being of minority group status (African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic) would demonstrate a significantly lower performance than those who reported being of majority group status (Caucasian/White). A literature search and relevant research leads to the hypothesis that the differences in performance based on ethnicity would be more evident in neurocognitive measures high in cultural loading and linguistic demand (Cormier et al., 2011). Specifically, the observed differences across ethnicity groups were expected to be found for those subtests that are rated high for cultural loading and linguistic demand, according to Flanagan and colleagues' (2007, 2013) C-LTC classifications for the WJ III COG NU, NESPY-II, and D-KEFS.

Data Analyses

For successful investigation of the aforementioned research questions, the appropriate statistical analyses must be selected for use with the available data. Based upon the nature of the research design and questions, a two-pronged approach using tests of differences was necessary. Before any statistical analyses were computed, preliminary procedures were carried out to determine the sample size, effect size, and alpha level that would yield the most powerful statistical outcome (Schumacker & Lomax, 2004).

Power and Error

A power analysis was performed prior to (*a priori*) determining participants for this study (Mertler & Vannatta, 2010). This determined the number of participants needed for the sample, based upon the alpha level, degrees of freedom, and desired or expected effect size (derived from existing literature), in order to have adequate power in the statistical analysis (Schumacker & Lomax, 2004). Power indicates the probability that a statistically significant difference can be revealed in the analysis if such a difference exists. This probability is directly related to the probability of having a Type II error; the lower the power, the greater the chance of having a Type II error. However, the greater the effect size, the lower the power needed to identify significance if it exists. All of these elements should be addressed using a balancing approach. Conducting the power analysis *a priori* assisted in the decisions regarding sample size, so that the analysis was able to accurately discover any effect without it being due to a false positive (Type I error). An issue of this sort would have affected the inferences that were drawn from the results. For this study, the number of participants needed to ensure adequate power was 251.

The final area for concern is regarding statistical conclusion validity. Attempts to control for threats to this included the conduction of a power analysis *a priori* to ensure adequate power in the sample, choosing the most appropriate statistical procedure for the research question, pre-screening the data to detect any violated assumptions of the data, and avoiding any over-interpretations of the findings of the study (Schumacker & Lomax, 2004).

Statistical Methods

Preliminary Descriptive Statistics

After pre-screening the data, preliminary descriptive statistics were computed to determine the means, standard deviations, frequency distributions and bivariate correlations for each of the neurocognitive subtests selected from the WJ III COG NU, NEPSY-II, and D-KEFS. Descriptive statistics were also calculated and reported for the demographic variables of the mixed clinical sample selected for inclusion in the study. Following the summary of descriptive statistics, the data were then statistically analyzed to evaluate the research questions previously identified.

Primary Statistical Analyses

To answer the first research question (RQ1), "do differences in performance exist based upon ethnicity for three standardized measures of neurocognitive functioning," a multivariate analysis of variance (MANOVA) was computed. A MANOVA involves an assessment for group differences across a single independent variable (IV) for a linear combination of dependent variables (DVs) (Mertler & Vannatta, 2010). In this statistical procedure, there was one categorical IV (IV: ethnicity) and the neurocognitive subtests from the WJ III COG NU, NEPSY-II, and D-KEFS were the quantitative DVs that collectively formed the new variable (y-hat) that was compared with the IV. The use of this type of analysis enabled the researcher to maximize the group differences for the IV, because a separate, weighted y-hat was created for each comparison to be made. As such, this approach revealed differences that may not have been detected in separate, univariate ANOVAs.

RQ2: How much of the difference in performance is attributed to ethnicity? The best approach to answer this research question was to utilize a discriminant function analysis (DFA) as a post-hoc procedure following the MANOVA (Enders, 2003; Hadzi-Pavlovic, 2010). This statistical procedure produced beta weights and structure coefficients to examine the variance explained by each DV. It is a more powerful post-hoc test than others frequently used (i.e., Tukey's honestly significant post-hoc analysis) and it provides a more detailed look at the contributed variance by each DV within the y-hat with regard to the IV.

Conclusion

In summary, the issues related to neurocognitive assessment of CLD individuals are in need of investigation. Performance on these measures of neurocognitive functioning is influenced by the level of acculturation and language proficiency of the individual being evaluated. To investigate the effect of language demand and cultural loading on neurocognitive performance for CLD individuals, a study was conducted to determine if there were differences in performance on three specific instruments, frequently used in psychological practice.

To answer the research questions, data were drawn from an archival set of case studies submitted by students of the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program. A mixed clinical sample of children and adolescents with neurological and developmental disabilities comprised this sample. To analyze the data, a MANOVA and a DFA were conducted using SPSS. It was expected that this study would reveal statistically significant differences in performance across ethnicity groups. Furthermore, the differences in performance based on ethnicity were expected to be more pronounced for subtests that are rated high for cultural loading and linguistic demand based on the C-LTC developed by Flanagan and colleagues (2007, 2013).

CHAPTER IV

RESULTS

This chapter will present the findings of the current study. Descriptive statistics will be summarized, as well as results from the primary statistical analyses. Data will be displayed using tables and figures, and accompanying narrative explanations of the results will serve to describe any significant findings. A discussion of these findings in relation to the aforementioned research questions, as well as general implications of this study, can be found in the subsequent chapter.

Preliminary Statistical Analyses

Assumptions

Statistical analyses were computed once confidence was assured that the data did not violate any of the corresponding assumptions. Thus, before computing any statistical analyses for this study, the data were pre-screened for multivariate normality, linearity, homogeneity of variance/covariance, univariate and multivariate outliers, absence of multicollinearity, and independence of observations (Mertler & Vannatta, 2010). These assumptions were tested in a number of ways, including the use of residual plots and reviewing the results from Box's M in SPSS. This reveals whether the variance among the group of DVs is similar enough for comparative purposes, while ensuring that there is not a statistically significant difference. If the correlation among DVs was high, this would demonstrate that the subtests included in the analysis were sharing too much of the variance. The overlap would suggest that they are essentially explaining the same variance within the y-hat. One solution to this problem would be to remove one of the DVs. This was not needed in the current study.

The meeting of these assumptions is critical, as the methods used for computing the primary statistical analyses, one-way MANOVA and DFA, are dependent upon them. If the data had violated one of the assumptions, multivariate normality for example, then steps would have been taken to transform the data or alternative statistical procedures would have been used to evaluate the research questions (Schumacker & Lomax, 2004). The data were screened for univariate outliers by looking at bivariate scatter plots among each pair of variables. Multivariate outliers were screened for by looking at the Mahalahobis distance and running a statistical analysis in SPSS.

The assumption of independence of observations is one that should be evaluated when collecting data, entering data, and before analyzing data (Mertler & Vannatta, 2010). This reflects confidence that each observation was independent and that there are no repeats in the data (i.e., no one participated in the study more than once). After the data were entered, it was screened for the presence of repeated cases manually. Additionally, data were assessed for independence of observations by looking for possible clerical or administration errors.

Typically, in the event that there are not enough cases with subtest scores present to ensure adequate power for any single subtest, said subtest would be excluded. However, the data used in this study had already been screened to determine how much of it was missing, and statistical software was used to impute the data when less than

10% was missing at random (MAR). Specifically, multiple imputation (MI) was used, which allows for the inclusion of more complete case files. Through this procedure, each missing value is imputed is replaced by *m*>1 plausible values based on a Monte Carlo simulation technique. For this data set, each missing value was imputed five times using Lisrel 8.80 and then analyzed and combined by way of Rubin's (1987) method to produce estimates and confidence intervals that incorporate missing data uncertainty. While the use of imputed data may be helpful in the analysis of incomplete case files, it comes with the caveat that some of the data is the product of a mathematical prediction based on existing data. In fact, for some subtests that are given to participants only sparingly, a significant portion of the scores (>50%) had to be imputed. For example, the Planning subtest from the WJ III COG NU was excluded from this study for this reason.

Descriptive Statistics

Preliminary descriptive statistics were conducted to determine the means, standard deviations, and frequencies of the demographic variables included in this study. Of the 911 cases in the archival data sample, 387 (42.5%) were missing ethnicity data. The remaining 520 cases (*n* = 520) were included in the primary statistical analyses. Table 4 displays the frequencies and percentages for the demographic variables included in this study. Ethnicity in the final sample consisted of 70.2% Caucasian/White, 10.4% African-American/Black, 11.2% Asian-American/Pacific Islander, and 8.3% Latino/Hispanic. The sample was comprised of 54.6% males and 45.4% females. While the proportion of males to females is ideal, the disproportionate number of

Caucasian/White participants in comparison to all of the other ethnicity groups is not ideal.

This study utilized a sample of children and adolescents with various clinical diagnoses. The frequencies and percentages of these clinical diagnoses are displayed in Table 5. However, clinical diagnoses for 30.8% of the sample was either not reported or unknown. The most frequently observed diagnoses in this sample were learning disability (LD; 17.7%) and Attention Deficit/Hyperactivity Disorder (ADHD; 12.7%). Also included in this table and in the data analysis are pairs of comorbid clinical diagnoses, such as LD and ADHD, which were reported in the data sample. The frequencies of the clinical diagnoses reported in this sample are not necessarily representative of the frequency with which they are observed in the general population, as this sample is comprised of data from cases that met specific criteria for this study. For example, cases with missing ethnicity data and those with clinical diagnoses consistent with an intellectual disability were excluded from the study.

Table 4

Frequencies and Percentages of the Demographic Variables

Variable	*N*	%
Ethnicity		
Caucasian/White	365	70.2
African-American/Black	54	10.4
Asian-American/Pacific Islander	58	11.2
Latino/Hispanic	43	8.3
Gender		
Male	284	54.6
Female	236	45.4

Table 5

Frequencies and Percentages of the Clinical Diagnoses

Clinical Diagnosis	*N*	%
Learning Disability	92	17.7
Language Disability	8	1.5
Neurological Impairment (Acquired)	29	5.6
ADHD	66	12.7
Autism Spectrum	27	5.2
Emotional Disability	21	4.0
General Medical (OHI)	17	3.3
Deaf	4	0.8
Other (Multiple Disabilities)	26	5.0
LD/ADHD (Comorbid)	42	8.1
Neurological Impairment/ADHD (Comorbid)	2	0.4
Autism/ADHD (Comorbid)	4	0.8
ED/ADHD (Comorbid)	17	3.3
General Medical/ADHD (Comorbid)	5	1.0
Not Reported	160	30.8

Note. ADHD = Attention Deficit/Hyperactivity Disorder, OHI = Other Health Impairment, LD = Learning Disability, ED = Emotional Disability

The overall means and standard deviations (*SD*) for the subtest scores selected for use as dependent variables in this study are listed in Table 6. Standard Scores are used for the subtests of the WJ III COG NU, with a mean of 100 and *SD* of 15. For both the NEPSY-II and D-KEFS subtests, Scaled Scores are used, with a mean of 10 and *SD* of 3. Scores are described as being in the average range if they fall within one standard deviation of the mean. All of the means fell in the average range for the subtests of the WJ III COG NU, while only 20 of the 22 NEPSY-II subtests and four of the five D-KEFS subtests had means that fell in the average range. For the NEPSY-II, subtests with mean scores that fell below average were Geometric Puzzles (*M* = 5.2) and Inhibition

Switching –Combined (*M* = 7.0). The D-KEFS subtest with a mean score below average was the Trail Making Test – Condition 4 (*M* = 6.4).

In Tables 7 and 8, bivariate correlations are presented for the WJ III COG NU, NEPSY-II, and D-KEFS subtests selected for this study. These correlations were computed to test for multicollinearity. Subtest correlations between the WJ III COG NU and the NEPSY-II ranged from *r* = -.23 (Retrieval Fluency and Memory for Names – Delayed) to *r* = .32 (Verbal Comprehension and Comprehension of Instructions). Of the 374 correlations calculated among the subtests of the WJ III COG NU and the NEPSY-II, 55 were significant at the *p* < .05 level and 134 were significant at the *p* < .01 level. For the NEPSY-II, the Comprehension of Instructions subtest had the highest frequency of significant correlations with the WJ III COG NU. The Auditory Working Memory subtest from the WJ III COG NU had the highest frequency of significant correlations with the NEPSY-II.

Table 6

Means and Standard Deviations of the Subtests

Subtest	*N*	Mean	*(SD)*
WJ III COG NU			
Pair Cancellation	519	94.6	(10.4)
Visual Matching	520	86.1	(11.9)
Spatial Relations	520	98.5	(10.9)
Picture Recognition	519	99.4	(11.1)
Sound Blending	519	104.2	(15.3)
Incomplete Words	520	96.2	(13.1)
Rapid Picture Naming	520	85.2	(11.2)
Memory for Words	519	92.8	(12.9)

(Continued)

Visual Auditory Learning	518	88.7	(15.5)
Retrieval Fluency	520	91.6	(12.1)
Numbers Reversed	516	90.5	(14.1)
Auditory Working Memory	517	94.5	(15.1)
Verbal Comprehension	519	95.6	(13.1)
General Information	519	93.2	(13.8)
Concept Formation	519	96.2	(14.1)
Analysis Synthesis	520	97.4	(13.2)
Decision Speed	518	93.7	(15.5)
NEPSY-II			
Finger Tapping – Dominant Hand	520	8.9	(2.7)
Imitating Hand Positions	520	8.0	(2.2)
Visuomotor Precision	482	8.8	(3.3)
Response Set	477	7.5	(4.3)
Block Construction	520	8.1	(2.5)
Arrows	520	8.7	(3.1)
Geometric Puzzles	520	5.2	(3.4)
Picture Puzzles	520	8.8	(2.6)
Speeded Naming	519	7.8	(2.9)
Comprehension of Instructions	519	7.7	(2.9)
Phonological Processing	520	7.6	(2.7)
Word List Interference	500	8.6	(2.3)
Narrative Memory – Free Recall	519	10.2	(3.5)
Memory for Faces	508	9.6	(2.7)
Memory for Faces – Delayed	520	9.1	(2.5)
Memory for Designs	520	8.9	(2.9)
Memory for Designs – Delayed	520	8.2	(3.0)
Memory for Names	519	8.2	(2.9)
Memory for Names – Delayed	520	7.8	(2.7)
Animal Sorting – Combined	505	8.2	(2.9)
Design Fluency	520	8.1	(2.5)
Inhibition Switching – Combined	520	7.0	(2.3)
D-KEFS			
Trail Making Test – Condition 4	518	6.4	(3.4)
Design Fluency – Condition 1	520	8.9	(1.5)
Color Word Interference – Condition 4	518	7.7	(2.9)
Word Context	520	8.0	(2.9)
Verbal Fluency – Condition 1	518	8.6	(2.8)

Table 7

Bivariate Correlations between the Selected Subtests from the WJ III COG NU and the NEPSY-II and D-KEFS

	WJ III COG NU																
	PC	VM	SR	PR	SB	IW	RPN	MW	VAL	RF	NR	AWM	VC	GI	CF	AS	DS
NEPSY-II																	
FT	.09*	.14**	.07	.11*	.03	.03	.07	.04	.07	.26**	-.03	-.02	.14**	.10*	.09	.00	.02
HP	.06	.12**	.05	.17**	.05	.06	-.01	-.02	.07	.21**	.07	.13**	.11*	.03	.08	.02	.00
VMP	-.11*	.07	.07	.04	.09	.06	.17**	.19**	-.01	-.02	.04	.03	.07	.10*	.09	.06	.08
RS	-.01	.07	-.01	.06	.02	.09*	.05	.01	.07	-.07	.15**	.13**	.09*	.04	.05	.10*	.18**
BC	.13**	.13**	.16**	.15**	.10*	.04	.15**	.12**	.10*	.17**	.03	.14**	.15**	.18**	.22**	.19**	.07
AR	.06	.06	.13**	.06	.12**	.08	.25**	.04	.11*	-.02	.11*	.08	.14**	.13**	.13**	.20**	.15**
GP	.11*	.04	.05	.04	.05	.07	.13**	.03	.04	.06	.10*	.12**	.06	.12**	.12**	.17**	.18**
PPZ	.02	.03	.10*	.06	.06	.11*	.15**	.06	-.00	.18**	.12**	.15**	.08	.10*	.16**	.11*	.06
SN	.04	.04	.03	.02	.18**	.23**	.22**	.20**	.14**	.09*	.10*	.14**	.19**	.22**	.07	.16**	.10*
CI	.11*	.17**	.20**	.16**	.18**	.10*	.15**	.21**	.19**	.19**	.16**	.21**	.32**	.20**	.26**	.23**	.07
PP	.15**	.19**	.01	.11*	.11*	.08	.18**	.04	.16**	.15**	.05	.02	.15**	.06	.05	.13**	.11*
WLI	.02	-.05	-.06	-.04	-.03	.09*	.03	.07	-.03	-.07	.07	.03	.11*	.10*	.01	.05	.01
NMFR	.07	.02	.11*	.01	.07	-.01	.02	.04	.12**	-.05	.05	.12**	.20**	.21**	.12**	.17**	.06
MF	.08	.12**	.05	.02	.09*	.14**	.01	.09*	.06	.05	.19**	.20**	.07	.06	.12**	.09*	.04
MFD	.04	.07	.10*	.07	.16**	.16**	.04	.24**	.10*	.00	.19**	.16**	.08	.10*	.11*	.16**	.15**
MD	.06	.06	.08	.07	.03	-.07	.20**	.19**	.08	.04	.07	.19**	.09*	.12**	.08	.22**	.09*
MDD	.09**	.10*	.14**	.15**	.04	-.06	.11*	.14**	.15**	.19**	.10*	.16**	.12**	.12**	.12**	.19**	.11**

(Continued)

MN	-.01	.02	.04	.02	.13**	.14**	.08	.14**	.11**	-.19**	.13**	.15**	.21**	.12**	.14**	.11*	.07
MND	-.12**	-.13**	-.03	-.05	.06	-.02	-.03	.13**	.11*	-.23**	.11*	.15**	.04	-.04	.07	.03	-.12**
ASC	.06	-.07	.02	.01	.18**	.02	.04	.08	.21**	-.03	.13**	.27**	.14**	.11*	.19**	.22**	.09*
DF	-.06	.01	.03	.03	.09*	.03	-.02	.12**	.01	.11*	.06	.10*	.10*	.10*	.15**	.09	.03
ISC	.01	.02	.01	.04	.09*	.03	.20**	-.04	.07	.07	.10*	.10*	.02	.04	-.04	-.03	-.04
D-KEFS																	
TMT-4	.20**	.02	.09*	.09*	.02	.10*	-.06	.10*	.12**	.06	.13**	.07	.15**	.10*	.18**	.18**	.14**
DF-1	.09*	.11*	.22**	.15**	.13**	-.05	-.02	.02	.11*	.17**	.10*	.22**	.07	.07	.13**	.04	.01
CWI-4	.08	.00	-.01	.03	.06	.07	.12**	-.05	.01	-.01	.09*	.04	.04	-.02	.02	.02	.04
WC	.05	.16**	.15**	.11*	.16**	-.02	.15**	.13**	.20**	.09*	.13**	.14**	.19**	.11*	.16**	.18**	.06
VF-1	.07	.16**	.10*	.11*	.19**	.09*	.15**	.17**	.05	.18**	.16**	.20**	.14**	.10*	.16**	.15**	.10*

Note. PC = Pair Cancellation; VM = Visual Matching; SR = Spatial Relations; PR = Picture Recognition; SB = Sound Blending; IW = Incomplete Words; RPN = Rapid Picture Naming; MW = Memory for Words; VAL = Visual Auditory Learning; RF = Retrieval Fluency; NR = Numbers Reversed; AWM = Auditory Working Memory; VC = Verbal Comprehension; GI = General Information; CF = Concept Formation; AS = Analysis Synthesis; DS = Decision Speed; FT = Finger Tapping – Dominant Hand; HP = Imitating Hand Positions; VMP = Visuomotor Precision; RS = Response Set; BC = Block Construction; AR = Arrows; GP = Geometric Puzzles; PPZ = Picture Puzzles; SN = Speeded Naming; CI = Comprehension of Instructions; PP = Phonological Processing; WLI = Word List Interference; NMFR = Narrative Memory – Free Recall; MF = Memory for Faces; MFD = Memory for Faces – Delayed; MD = Memory for Designs; MDD = Memory for Designs – Delayed; MN = Memory for Names; MND = Memory for Names – Delayed; ASC = Animal Sorting – Combined; DF = Design Fluency; ISC = Inhibition Switching – Combined; TMT-4 = Trail Making Test – Condition 4; DF-1 = Design Fluency – Condition 1; CWI-4 = Color Word Interference – Condition 4; WC = Word Context; VF-1 = Verbal Fluency – Condition 1.
* *p* ≤ .05. ** *p* ≤ .01. *** *p* ≤ .001

Table 8

Bivariate Correlations between the Selected Subtests from the NEPSY-II and the D-KEFS

	D-KEFS				
	TMT-4	DF-1	CWI-4	WC	VF-1
NEPSY-II					
FT	-.06	.03	-.01	.16**	.08
HP	.04	.20**	-.08	-.07	-.00
VMP	-.08	-.09	-.10*	-.01	-.05
RS	.05	-.13**	-.03	.00	.06
BC	.18**	.09*	.02	-.02	.13**
AR	.05	.00	.01	.12**	.05
GP	.09	.08	-.01	.01	.10*
PPZ	.02	.02	-.14**	-.11**	.23**
SN	.07	.00	.07	.05	-.01
CI	.02	.06	-.05	.17**	.32**
PP	-.03	-.11*	.13**	.14**	.11**
WLI	.13**	-.07	-.30**	-.33**	.17**
NMFR	-.04	.01	-.12**	.04	.03
MF	-.03	.04	-.10*	-.01	.10*
MFD	.05	.00	-.19**	-.07	.05
MD	-.04	.09*	-.11*	.03	.11*
MDD	-.02	.07	-.05	-.00	.15**
MN	-.03	.01	-.02	.03	.00
MND	-.01	.00	-.04	.08	-.06
ASC	.17**	.10*	-.04	.07	.16**
DF	.05	.03	-.17**	-.01	.06
ISC	.02	.03	.16**	.04	.16**

Note. FT = Finger Tapping – Dominant Hand; HP = Imitating Hand Positions; VMP = Visuomotor Precision; RS = Response Set; BC = Block Construction; AR = Arrows; GP = Geometric Puzzles; PPZ = Picture Puzzles; SN = Speeded Naming; CI = Comprehension of Instructions; PP = Phonological Processing; WLI = Word List Interference; NMFR = Narrative Memory – Free Recall; MF = Memory for Faces; MFD = Memory for Faces – Delayed; MD = Memory for Designs; MDD = Memory for Designs – Delayed; MN = Memory for Names; MND = Memory for Names – Delayed; ASC = Animal Sorting – Combined; DF = Design Fluency; ISC = Inhibition Switching – Combined; TMT-4 = Trail Making Test – Condition 4; DF-1 = Design Fluency – Condition 1; CWI-4 = Color Word Interference – Condition 4; WC = Word Context; VF-1 = Verbal Fluency – Condition 1.
* $p \leq .05$. ** $p \leq .01$. *** $p \leq .001$

Subtest correlations between the WJ III COG NU and the D-KEFS ranged from *r* = -.06 (Rapid Picture Naming and Trail Making Test – Condition 4) to *r* = .22 (Spatial Relations and Design Fluency – Condition 1; Auditory Working Memory and Design Fluency – Condition 1). Of the 85 correlations calculated among the subtests of the WJ III COG NU and the D-KEFS, 17 were significant at the $p < .05$ level and 35 were significant at the $p < .01$ level. Concept Formation from the WJ III COG NU had the highest frequency of significant correlations at the $p < .01$ level with the D-KEFS. The Word Context subtest from the D-KEFS had the highest frequency of significant correlations with the WJ III COG NU.

Bivariate correlations between the subtests selected for this study from the D-KEFS and the NEPSY-II ranged from *r* = -.33 (Word Context and Word List Interference) to *r* = .32 (Verbal Fluency – Condition 1 and Comprehension of Instructions). There were 110 correlations calculated among the subtests of the D-KEFS and the NEPSY-II, of which 10 were significant at the $p < .05$ level and 26 were significant at the $p < .01$ level. Verbal Fluency – Condition 1 was the D-KEFS subtest that had the highest frequency of significant correlations with the NEPSY-II at the $p < .01$ level. The Word List Interference subtest from the NEPSY-II had the highest frequency of significant correlations with the D-KEFS at the $p < .01$ level.

Primary Statistical Analyses

Following the preliminary statistical analyses, the primary statistical analyses of MANOVA and DFA were conducted to answer the research questions previously described in the preceding chapters:

1. Do differences in performance exist based upon ethnicity for the three standardized (WJ III COG NU, NEPSY-II, D-KEFS) measures of neurocognitive functioning?
2. How much of the difference in performance is attributed to ethnicity?

The statistical method of MANOVA allowed for the comparison of the IV (ethnicity) to a linear combination of the DVs (y-hat), which were comprised of the subtests selected from the WJ III COG NU, NEPSY-II, and D-KEFS. It was the best way to approach this question, because it allowed for the evaluation of group differences across a number of DVs in a single statistical procedure (Grice & Iwasaki, 2007; Jaccard & Guilamo-Ramos, 2002). This procedure reveals whether there is a statistically significant difference in performance based on ethnicity. While the use of several analyses of variances (ANOVAs) could answer the same question, the use of a MANOVA allowed for the reduction of the probability of a Type I error through experiment-wise control of the alpha level (Grice & Iwasaki, 2007). By evaluating the significance of the MANOVA, using Wilks' Lambda as the omnibus *F* test, it was determined whether there was a significant difference between the IV and the linear combination of DVs (Jaccard & Guilamo-Ramos, 2002).

Table 9

Means and Standard Deviations of the Subtests by Ethnicity

	Caucasian/White		African-American/Black		Asian-American/Pacific Islander		Latino/Hispanic	
Subtest	Mean	*(SD)*	Mean	*(SD)*	Mean	*(SD)*	Mean	*(SD)*
WJ III COG NU								
Pair Cancellation	94.80	(9.33)	93.10	(7.93)	96.30	(7.70)	95.87	(7.91)
Visual Matching	85.51	(12.60)	83.37	(10.40)	87.31	(10.73)	85.90	(6.38)
Spatial Relations*	100.08	(11.47)	95.20	(8.28)	97.40	(10.67)	99.77	(8.04)
Picture Recognition*	99.66	(10.90)	95.90	(14.06)	102.98	(10.22)	98.87	(12.75)
Sound Blending*	106.89	(14.24)	95.78	(15.42)	103.12	(10.61)	100.00	(22.80)
Incomplete Words	96.30	(12.97)	94.37	(13.38)	95.49	(9.13)	90.84	(15.54)
Rapid Picture Naming*	86.21	(11.37)	81.24	(10.79)	83.18	(13.22)	84.61	(9.13)
Memory for Words	93.02	(11.97)	89.32	(14.89)	91.20	(12.15)	88.55	(14.83)
Visual Auditory Learning	89.98	(14.37)	85.70	(16.62)	87.94	(19.02)	86.00	(17.05)
Retrieval Fluency	92.32	(12.67)	89.78	(10.67)	90.98	(10.97)	90.42	(5.37)
Numbers Reversed*	92.89	(13.44)	87.27	(13.67)	87.67	(12.63)	92.97	(13.17)
Auditory Working Memory*	96.83	(14.13)	90.71	(11.46)	95.86	(12.93)	89.16	(14.15)
Verbal Comprehension*	98.74	(12.32)	88.46	(12.49)	93.55	(13.36)	87.06	(12.89)
General Information*	96.41	(13.15)	88.17	(14.70)	89.84	(11.81)	85.06	(13.40)
Concept Formation*	97.97	(13.79)	91.85	(11.59)	93.55	(17.28)	92.32	(9.90)
Analysis Synthesis*	98.36	(13.14)	92.39	(11.23)	95.14	(12.98)	95.00	(11.78)
Decision Speed	93.84	(16.61)	87.83	(13.38)	96.27	(12.10)	95.13	(13.31)
NEPSY-II								
Finger Tapping – Dominant Hand	9.04	(2.60)	8.34	(2.85)	8.92	(2.48)	9.29	(2.75)

(Continued)

Imitating Hand Positions	8.09	(2.28)	7.27	(1.98)	8.53	(2.92)	8.29	(1.99)
Visuomotor Precision	8.77	(3.11)	8.20	(3.43)	7.84	(3.53)	9.16	(3.97)
Response Set*	7.13	(4.47)	8.85	(3.77)	7.27	(4.09)	8.84	(3.54)
Block Construction	8.09	(2.68)	7.85	(2.18)	8.54	(1.80)	7.48	(2.19)
Arrows*	8.89	(3.06)	7.15	(3.01)	9.14	(3.18)	8.06	(2.89)
Geometric Puzzles*	4.78	(3.59)	6.32	(2.59)	5.59	(2.81)	6.19	(2.73)
Picture Puzzles	8.83	(2.71)	8.54	(1.89)	8.55	(1.90)	9.32	(1.83)
Speeded Naming*	7.95	(3.12)	6.93	(2.47)	7.00	(2.84)	6.90	(3.04)
Comprehension of Instructions	7.84	(3.04)	7.10	(3.00)	7.18	(2.30)	6.94	(1.82)
Phonological Processing	7.72	(3.03)	7.05	(2.39)	7.65	(2.18)	8.16	(1.90)
Word List Interference	8.50	(2.42)	8.15	(2.24)	7.90	(2.17)	9.23	(1.93)
Narrative Memory – Free Recall	10.50	(3.47)	10.66	(3.52)	10.39	(3.01)	9.71	(2.96)
Memory for Faces	9.50	(2.70)	10.10	(3.26)	9.61	(2.83)	10.10	(2.33)
Memory for Faces – Delayed	9.07	(2.60)	8.41	(2.41)	9.24	(2.33)	9.97	(3.29)
Memory for Designs	8.72	(2.90)	7.90	(3.32)	8.65	(2.23)	8.81	(2.79)
Memory for Designs – Delayed	8.36	(2.97)	7.76	(2.81)	8.04	(2.33)	8.55	(2.64)
Memory for Names	8.24	(3.11)	8.20	(2.53)	8.04	(2.83)	7.68	(2.43)
Memory for Names – Delayed	7.71	(2.61)	8.02	(2.79)	7.65	(2.16)	8.42	(3.11)
Animal Sorting – Combined	8.31	(2.91)	7.90	(2.72)	7.88	(2.85)	8.00	(2.74)
Design Fluency	8.00	(2.61)	7.90	(2.62)	8.35	(1.75)	7.68	(1.64)
Inhibition Switching – Combined	7.42	(2.23)	7.27	(2.09)	6.51	(1.85)	7.19	(2.51)
D-KEFS								
Trail Making Test – Condition 4	6.25	(3.50)	5.28	(3.51)	6.63	(3.17)	6.23	(2.86)
Design Fluency – Condition 1	9.07	(1.57)	8.90	(1.11)	9.04	(1.46)	9.13	(1.02)
Color Word Interference – Condition 4	8.06	(2.94)	7.55	(2.99)	7.97	(3.08)	6.74	(2.84)
Word Context*	8.37	(2.95)	8.51	(2.73)	7.14	(2.32)	7.32	(3.80)
Verbal Fluency – Condition 1	8.70	(3.00)	8.66	(2.23)	8.33	(1.98)	8.68	(2.53)

The presence of a statistically significant difference indicates that a post-hoc test should be used to locate the source of the most variance explained among the DVs. Instead of using several univariate ANOVAs, a DFA was conducted as a post-hoc procedure. This choice was based, again, upon the premise that multiple univariate ANOVAs lead to inflated Type I error (Grice & Iwasaki, 2007). A DFA allows a researcher to use beta weights and structure coefficients to examine the variance explained by each DV (Mertler & Vannatta, 2010). It is a more powerful post-hoc test and it provides a more detailed look at the contributed variance by each DV within the y-hat with regard to the IV.

A presentation of the means and *SD*s for each of the subtests included in this study across the ethnicity categories may be found in Table 9. All of the means for the Caucasian/White subgroup fell in the average range for the WJ III COG NU subtests. For the African-American/Black subgroup, all of the WJ III COG NU subtests fell in the average range with the exception of Visual Matching (*M* = 83.37, *SD* = 10.40) and Rapid Picture Naming (*M* = 81.24, *SD* = 10.79). All of the WJ III COG NU subtests fell in the average range with the exception of Rapid Picture Naming for both the Asian-American/Pacific Islander (*M* = 83.13, *SD* = 13.22) and Latino/Hispanic (*M* = 84.61, *SD* = 9.13) subgroups.

For the NEPSY-II subtests, the Caucasian/White subgroup fell in the average range for all subtests except Geometric Puzzles (*M* = 4.78, *SD* = 3.59). The African-American/Black subgroup fell in the average range for all subtests of the NEPSY-II except for Geometric Puzzles (*M* = 6.32, *SD* = 2.59) and Speeded Naming (*M* = 6.93, *SD*

= 2.47). All of the means for the Asian-American/Pacific Islander subgroup fell in the average range for the NEPSY-II subtests with the exception of Geometric Puzzles (*M* = 5.59, *SD* = 2.81), Speeded Naming (*M* = 7.00, *SD* = 2.84), and Inhibition Switching – Combined (*M* = 6.51, *SD* = 1.85). The Latino/Hispanic subgroup had means in the average range for all of the NEPSY-II subtests except for Geometric Puzzles (*M* = 6.19, *SD* = 2.73), Speeded Naming (*M* = 6.90, *SD* = 3.04), and Comprehension of Instructions (*M* = 6.94, *SD* = 1.82).

For the D-KEFS subtests, the Caucasian/White subgroup had means in the average range for all subtests except for Trail Making Test – Condition 4 (*M* = 6.25, *SD* = 3.50). Similarly, all D-KEFS subtest means were in the average range except for the Trail Making Test – Condition 4 for the African-American/Black (*M* = 5.28, *SD* = 3.51) and Asian-American/Pacific Islander (*M* = 6.63, *SD* = 3.17) subgroups. For the Latino/Hispanic subgroup, all of the D-KEFS subtests fell in the average range except for Trail Making Test – Condition 4 (*M* = 6.23, *SD* = 2.86) and Color Word Interference – Condition 4 (*M* = 6.74, *SD* = 2.84).

Table 10

F and P Values for the WJ III COG NU, NEPSY-II, and D-KEFS Subtests by Ethnicity

Subtest	*F*	*P*
WJ III COG NU		
Pair Cancellation	1.08	.357
Visual Matching	.84	.472
Spatial Relations	2.94	.033*
Picture Recognition	2.95	.033*
Sound Blending	8.14	<.001*
Incomplete Words	1.82	.143
Rapid Picture Naming	2.90	.035*
Memory for Words	2.06	.105
Visual Auditory Learning	1.45	.229
Retrieval Fluency	.78	.504
Numbers Reversed	3.80	.010*
Auditory Working Memory	4.73	.003*
Verbal Comprehension	15.17	<.001*
General Information	12.04	<.001*
Concept Formation	4.23	.006*
Analysis Synthesis	3.36	.019*
Decision Speed	2.48	.061
NEPSY-II		
Finger Tapping – Dominant Hand	1.02	.386
Imitating Hand Positions	2.38	.069
Visuomotor Precision	1.63	.182
Response Set	3.08	.027*
Block Construction	1.26	.290
Arrows	4.64	.003*
Geometric Puzzles	4.06	.007*
Picture Puzzles	.79	.501
Speeded Naming	3.05	.029*
Comprehension of Instructions	1.96	.119
Phonological Processing	1.02	.384
Word List Interference	2.32	.075
Narrative Memory – Free Recall	.57	.634
Memory for Faces	.90	.440

(Continued)

Memory for Faces – Delayed	2.15	.094
Memory for Designs	1.01	.387
Memory for Designs – Delayed	.74	.528
Memory for Names	.37	.774
Memory for Names – Delayed	.82	.481
Animal Sorting – Combined	.55	.652
Design Fluency	.53	.665
Inhibition Switching – Combined	2.41	.067
D-KEFS		
Trail Making Test – Condition 4	1.28	.283
Design Fluency – Condition 1	.18	.912
Color Word Interference – Condition 4	2.03	.109
Word Context	3.48	.016*
Verbal Fluency – Condition 1	.26	.858

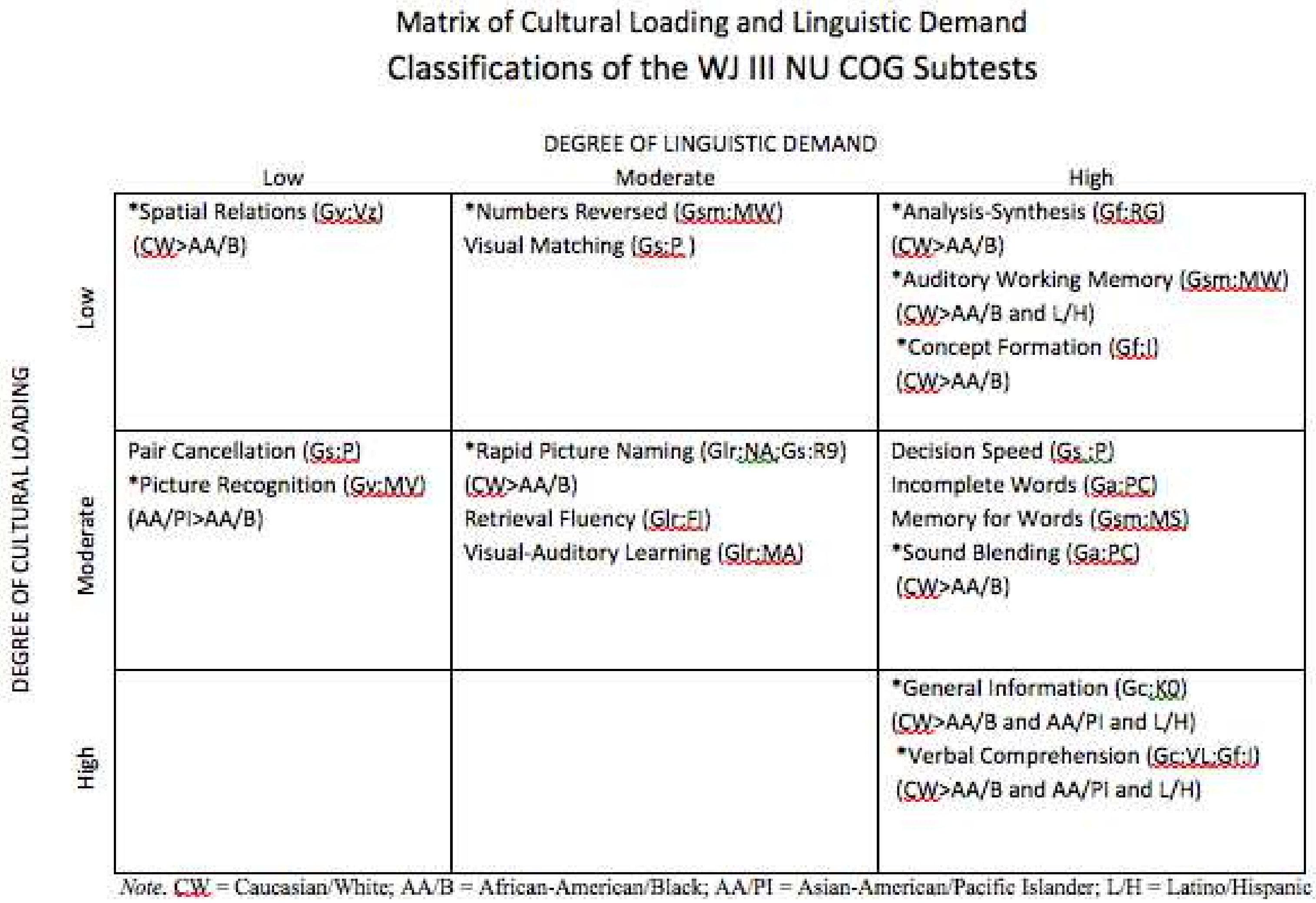
Table 10 includes the results from the MANOVA, indicating the multivariate effects and corresponding significance levels for all selected subtests from the WJ III COG NU, NEPSY-II, and D-KEFS across the independent variable, ethnicity. The MANOVA showed that there was a statistically significant multivariate effect between the ethnicity categories, $F(3, 387) = 1.83, p < .001, \eta^2 = .197$. Among the WJ III COG NU subtests, there were statistically significant differences on Spatial Relations, $F(3, 387) = 2.94, p = .033, \eta^2 = .022$, Picture Recognition, $F(3, 387) = 2.95, p = .033, \eta^2 = .022$, Sound Blending, $F(3, 387) = 8.14, p < .001, \eta^2 = .059$, Rapid Picture Naming, $F(3, 387) = 2.90, p = .035, \eta^2 = .022$, Numbers Reversed, $F(3, 387) = 3.80, p = .010, \eta^2 = .029$, Auditory Working Memory, $F(3, 387) = 4.73, p = .003, \eta^2 = .035$, Verbal Comprehension, $F(3, 387) = 15.17, p < .001, \eta^2 = .105$, General Information, $F(3, 387) = 12.04, p < .001, \eta^2 = .085$, Concept Formation, $F(3, 387) = 4.23, p = .006, \eta^2 = .032$, and Analysis Synthesis, $F(3, 387) = 3.36, p = .019, \eta^2 = .025$.

Tukey's honestly significant difference (HSD) post-hoc comparisons indicated that for the Spatial Relations subtest, children who were Caucasian/White (*M* = 100.08, *SD* = 11.47) had scores that were significantly higher than those who were African-American/Black (*M* = 95.20, *SD* = 8.28). On Picture Recognition, African-American/Black children (*M* = 95.90, *SD* = 14.06) scored significantly lower than Asian-American/Pacific Islander children (*M* = 102.98, *SD* = 10.22). For Sound Blending, African-American/Black children (*M* = 95.78, *SD* = 15.42) performed significantly lower than Caucasian/White children (*M* = 106.89, *SD* = 14.24). African-American/Black children (*M* = 81.24, *SD* = 10.79) also performed significantly lower than Caucasian/White children (*M* = 86.21, *SD* = 11.37) on the Rapid Picture Naming subtest. On Auditory Working Memory, both African-American/Black (*M* = 90.71, *SD* = 11.46) and Latino/Hispanic children (*M* = 89.16, *SD* = 14.15) had scores that were significantly lower than Caucasian/White children (*M* = 96.83, *SD* = 14.13).

On the Verbal Comprehension subtest, Caucasian/White children (*M* = 98.74, *SD* = 12.32) performed significantly better than African-American/Black (*M* = 88.46, *SD* = 12.49), Asian-American/Pacific Islander (*M* = 93.55, *SD* = 13.36), and Latino/Hispanic children (*M* = 87.06, *SD* = 12.89). Similarly, on the General Information subtest, Caucasian/White children (*M* = 96.41, *SD* = 13.15) outperformed the African-American/Black (*M* = 88.17, *SD* = 14.70), Asian-American/Pacific Islander (*M* = 89.84, *SD* = 11.81), and the Latino/Hispanic children (*M* = 85.06, *SD* = 13.40). On Concept Formation, African-American/Black children (*M* = 91.85, *SD* = 11.59) had scores that were significantly lower than Caucasian/White children (*M* = 97.97, *SD* = 13.79), and

African-American/Black children (*M* = 92.39, *SD* = 11.23) also scored significantly lower than Caucasian/White children (*M* = 98.36, *SD* = 13.14) on the Analysis Synthesis subtest. For Numbers Reversed, there were no significant differences between ethnicity groups at the univariate level.

For the NEPSY-II subtests, there were statistically significant differences on Response Set, $F(3, 387) = 3.08$, $p = .027$, $\eta^2 = .023$, Arrows, $F(3, 387) = 4.64$, $p = .003$, $\eta^2 = .035$, Geometric Puzzles, $F(3, 387) = 4.06$, $p = .007$, $\eta^2 = .031$, and Speeded Naming, $F(3, 387) = 3.05$, $p = .029$, $\eta^2 = .023$. For Response Set and Speeded Naming, Tukey's HSD post-hoc comparisons revealed that there were no significant differences between ethnicity groups at the univariate level. On the Arrows subtest, African-American/Black children (*M* = 7.15, *SD* = 3.01) had significantly lower scores than both Caucasian/White children (*M* = 8.89, *SD* = 3.06) and Asian-American/Pacific Islander children (*M* = 9.14, *SD* = 3.18). For Geometric Puzzles, Caucasian/White children (*M* = 4.78, *SD* = 3.59) performed significantly lower than African-American/Black children (*M* = 6.32, *SD* = 2.59). For the D-KEFS subtests, there were statistically significant differences on Word Context, $F(3, 387) = 3.48$, $p = .016$, $\eta^2 = .026$. Tukey's HSD post-hoc analysis showed that Asian-American/Pacific Islander children (*M* = 7.14, *SD* = 2.32) performed significantly lower than Caucasian/White children (*M* = 8.37, *SD* = 2.95). Figures 5, 6, and 7 illustrate the significant findings between ethnicity groups within the C-LTC classification tables for the WJ III COG NU, NEPSY-II, and D-KEFS, respectively. Only subtests included in this study are listed.



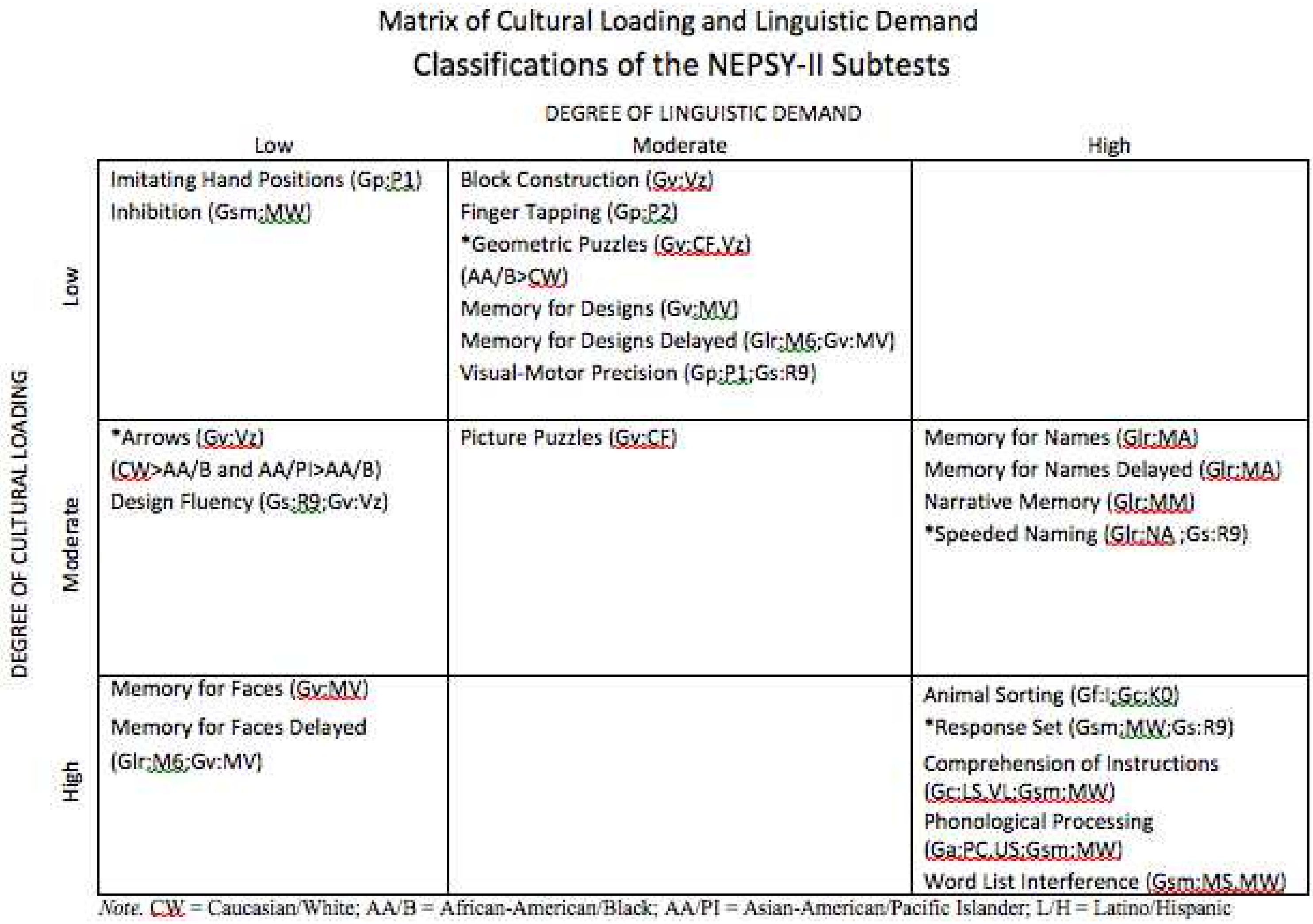
Matrix of Cultural Loading and Linguistic Demand

Classifications of the WJ III NU COG Subtests

		DEGREE OF LINGUISTIC DEMAND		
		Low	Moderate	High
DEGREE OF CULTURAL LOADING	Low	*Spatial Relations (Gv:Vz) (CW>AA/B)	*Numbers Reversed (Gsm:MW) Visual Matching (Gs:P)	*Analysis-Synthesis (Gf:RG) (CW>AA/B) *Auditory Working Memory (Gsm:MW) (CW>AA/B and L/H) *Concept Formation (Gf:I) (CW>AA/B)
	Moderate	Pair Cancellation (Gs:P) *Picture Recognition (Gv:MV) (AA/PI>AA/B)	*Rapid Picture Naming (Glr:NA;Gs:R9) (CW>AA/B) Retrieval Fluency (Glr:FI) Visual-Auditory Learning (Glr:MA)	Decision Speed (Gs.:P) Incomplete Words (Ga:PC) Memory for Words (Gsm:MS) *Sound Blending (Ga:PC) (CW>AA/B)
	High			*General Information (Gc:K0) (CW>AA/B and AA/PI and L/H) *Verbal Comprehension (Gc:VL;Gf:I) (CW>AA/B and AA/PI and L/H)

Note. CW = Caucasian/White; AA/B = African-American/Black; AA/PI = Asian-American/Pacific Islander; L/H = Latino/Hispanic

Figure 5. Significant findings between ethnicity groups for C-LTC Classifications of the WJ III COG NU



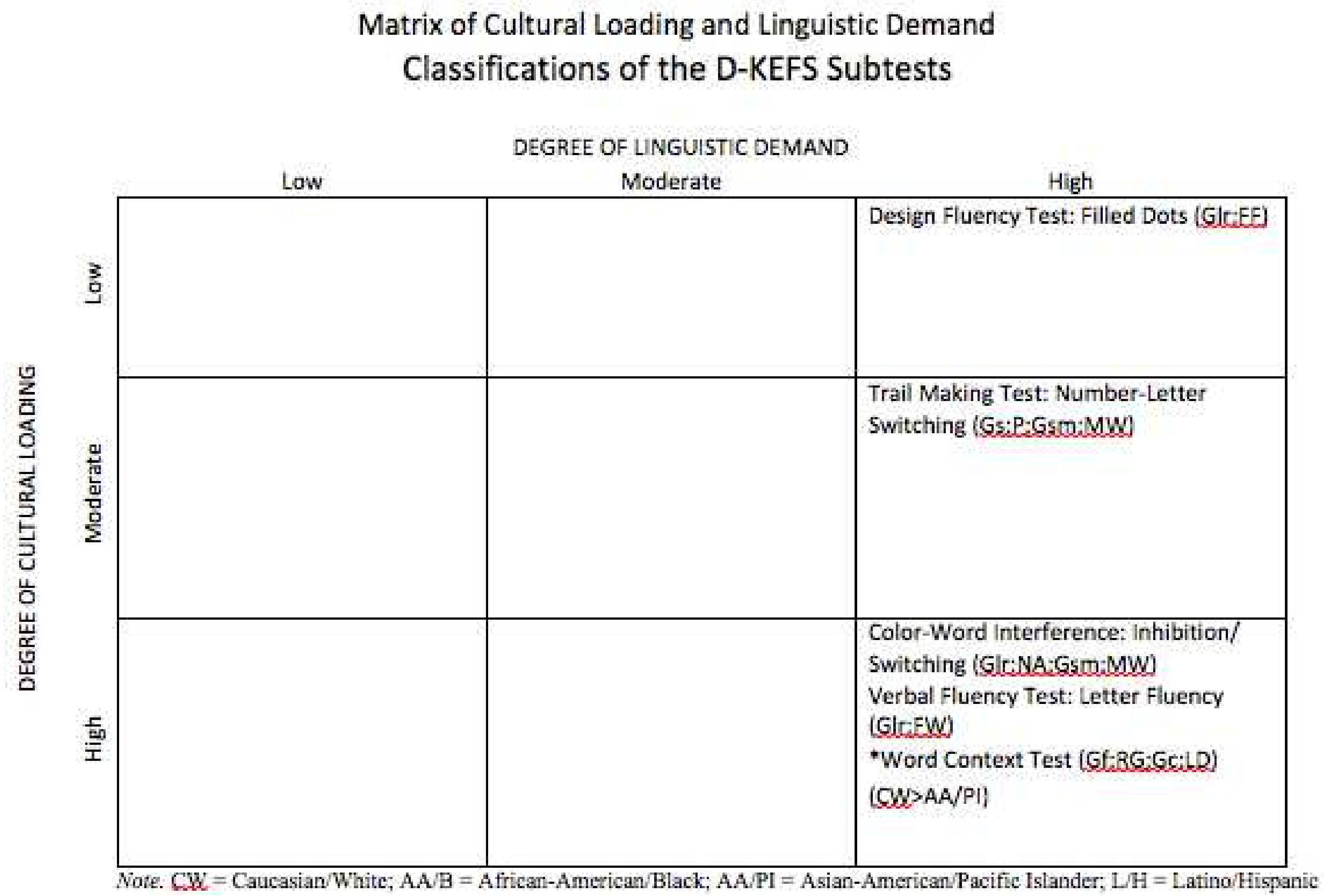
Matrix of Cultural Loading and Linguistic Demand

Classifications of the NEPSY-II Subtests

DEGREE OF CULTURAL LOADING	DEGREE OF LINGUISTIC DEMAND: Low	Moderate	High
Low	Imitating Hand Positions (Gp:P1) Inhibition (Gsm:MW)	Block Construction (Gv:Vz) Finger Tapping (Gp:P2) *Geometric Puzzles (Gv:CF,Vz) (AA/B>CW) Memory for Designs (Gv:MV) Memory for Designs Delayed (Glr:M6;Gv:MV) Visual-Motor Precision (Gp:P1;Gs:R9)	
Moderate	*Arrows (Gv:Vz) (CW>AA/B and AA/PI>AA/B) Design Fluency (Gs:R9;Gv:Vz)	Picture Puzzles (Gv:CF)	Memory for Names (Glr:MA) Memory for Names Delayed (Glr:MA) Narrative Memory (Glr:MM) *Speeded Naming (Glr:NA ;Gs:R9)
High	Memory for Faces (Gv:MV) Memory for Faces Delayed (Glr:M6;Gv:MV)		Animal Sorting (Gf:I;Gc:K0) *Response Set (Gsm:MW;Gs:R9) Comprehension of Instructions (Gc:LS,VL;Gsm:MW) Phonological Processing (Ga:PC,US;Gsm:MW) Word List Interference (Gsm:MS,MW)

Note. CW = Caucasian/White; AA/B = African-American/Black; AA/PI = Asian-American/Pacific Islander; L/H = Latino/Hispanic

Figure 6. Significant findings between ethnicity groups for C-LTC Classifications of the NEPSY-II



Matrix of Cultural Loading and Linguistic Demand

Classifications of the D-KEFS Subtests

DEGREE OF LINGUISTIC DEMAND

DEGREE OF CULTURAL LOADING

	Low	Moderate	High
Low			Design Fluency Test: Filled Dots (Glr:FF)
Moderate			Trail Making Test: Number-Letter Switching (Gs:P:Gsm:MW)
High			Color-Word Interference: Inhibition/ Switching (Glr:NA:Gsm:MW) Verbal Fluency Test: Letter Fluency (Glr:FW) *Word Context Test (Gf:RG:Gc:LD) (CW>AA/PI)

Note. CW = Caucasian/White; AA/B = African-American/Black; AA/PI = Asian-American/Pacific Islander; L/H = Latino/Hispanic

Figure 7. Significant findings between ethnicity groups for C-LTC Classifications of the D-KEFS

Although some of the scales showed overall multivariate effects for ethnicity, the absence of significant differences at the univariate level for a few of these suggests that this may be due to Type I error. It also could mean that the differences are only significant when in the context of the other subscales included in the analysis. As previously addressed in the methodology section, a DFA is a more powerful post-hoc procedure that does not carry with it the risk of inflated Type I error associated with using multiple univariate ANOVAs as performed for the MANOVA post-hoc analysis. The structure coefficients and standardized function coefficients obtained with the DFA provide more detailed information regarding the attribution of variance for each DV within the y-hat with regard to the IV (Grice & Iwasaki, 2007; Mertler & Vannatta, 2010). As such, a DFA was performed to better analyze the results of the MANOVA.

In the DFA, variables previously used as the DVs (i.e., subtests) became the IVs, or predictor variables, while the IV previously used (ethnicity) becomes the DV. This procedure seeks to determine how well participants may be classified into the four ethnicity categories based upon the scores obtained on the selected subtests. The DFA generated two statistically significant discriminant functions, $\Lambda = .517$, $\chi^2(138, N = 391) = 240.92$, $p < .001$ and $\Lambda = .701$, $\chi^2(90, N = 391) = 129.75$, $p = .004$, respectively. Ethnicity group was found to account for 26.2% of the variance for the first function and 18.7% of the variance for the second function. Correlation coefficients and standardized function coefficients for each discriminant function are listed in Tables 11 and 12. Also included in each of these tables are the C-LTC ratings for each of the subtests across both dimensions, cultural loading and linguistic demand.

Table 11

Correlation Coefficients and Standardized Function Coefficients for the First Discriminant Function Generated

Subtest	Correlation Coefficients	Standardized Function Coefficients	Cultural Loading	Linguistic Demand
WJ III COG NU				
Pair Cancellation	.009	-.263	Mod	Low
Visual Matching	.028	-.078	Low	Mod
Spatial Relations	.204*	.045	Low	Low
Picture Recognition	.077	-.140	Mod	Low
Sound Blending	.408*	.320	Mod	High
Incomplete Words	.160*	-.094	Mod	High
Rapid Picture Naming	.228*	.074	Mod	Mod
Memory for Words	.208*	-.019	Mod	High
Visual Auditory Learning	.177*	-.199	Mod	Mod
Retrieval Fluency	.130*	.079	Mod	Mod
Numbers Reversed	.212*	.008	Low	Mod
Auditory Working Memory	.293*	.193	Low	High
Verbal Comprehension	.568*	.342	High	High
General Information	.499*	.283	High	High
Concept Formation	.302*	.080	Low	High
Analysis Synthesis	.263*	.038	Low	High
Decision Speed	.086	.189	Mod	High
NEPSY-II				
Finger Tapping – Dominant Hand	.079	-.028	Low	Mod
Imitating Hand Positions	.067	.057	Low	Low
Visuomotor Precision	.083	.050	Low	Mod
Response Set	-.234*	-.356	High	High
Block Construction	.056	.015	Low	Mod
Arrows	.241*	.124	Mod	Low
Geometric Puzzles	-.297*	-.438	Low	Mod
Picture Puzzles	.016	-.098	Mod	Mod
Speeded Naming	.251*	.105	Mod	High
Comprehension of Instructions	.202*	.173	High	High
Phonological Processing	.050	-.050	High	High
Word List Interference	.027	.372	High	High

(Continued)

Narrative Memory – Free Recall	.043	-.302	Mod	High
Memory for Faces	-.132*	-.205	High	Low
Memory for Faces – Delayed	-.004	-.002	High	Low
Memory for Designs	.095	.307	Low	Mod
Memory for Designs – Delayed	.078	-.316	Low	Mod
Memory for Names	.061	.148	Mod	High
Memory for Names – Delayed	-.098	-.283	Mod	High
Animal Sorting – Combined	.101*	.110	High	High
Design Fluency	.017	-.065	Mod	Low
Inhibition Switching – Combined	.122	-.005	Low	Low
D-KEFS				
Trail Making Test – Condition 4	.079	.033	Mod	High
Design Fluency – Condition 1	.031	.033	Low	High
Color Word Interference – Condition 4	.166*	.095	High	High
Word Context	.140	.172	High	High
Verbal Fluency – Condition 1	.032	-.175	High	High

Table 12

Correlation Coefficients and Standardized Function Coefficients for the Second Discriminant Function Generated

Subtest	Correlation Coefficients	Standardized Function Coefficients	Cultural Loading	Linguistic Demand
WJ III COG NU				
Pair Cancellation	-.101	-.056	Mod	Low
Visual Matching	-.114	-.222	Low	Mod
Spatial Relations	.091	.255	Low	Low
Picture Recognition	-.251*	-.295	Mod	Low
Sound Blending	-.057	.048	Mod	High
Incomplete Words	-.089	-.158	Mod	High
Rapid Picture Naming	.076	.069	Mod	Mod
Memory for Words	-.038	-.248	Mod	High
Visual Auditory Learning	-.019	.007	Mod	Mod
Retrieval Fluency	.003	.050	Mod	Mod
Numbers Reversed	.196	.406	Low	Mod
Auditory Working Memory	-.158	-.359	Low	High

(Continued)

Verbal Comprehension	-.084	-.341	High	High
General Information	.015	.262	High	High
Concept Formation	.041	.128	Low	High
Analysis Synthesis	.031	.064	Low	High
Decision Speed	-.131	-.046	Mod	High
NEPSY-II				
Finger Tapping – Dominant Hand	.019	.042	Low	Mod
Imitating Hand Positions	-.148	-.072	Low	Low
Visuomotor Precision	.190*	.193	Low	Mod
Response Set	.136	.230	High	High
Block Construction	-.191*	-.110	Low	Mod
Arrows	-.200	-.369	Mod	Low
Geometric Puzzles	.014	-.031	Low	Mod
Picture Puzzles	.107	.077	Mod	Mod
Speeded Naming	.067	-.061	Mod	High
Comprehension of Instructions	.031	.003	High	High
Phonological Processing	.023	-.076	High	High
Word List Interference	.225*	.533	High	High
Narrative Memory – Free Recall	-.032	-.191	Mod	High
Memory for Faces	.058	.110	High	Low
Memory for Faces – Delayed	.012	.029	High	Low
Memory for Designs	-.016	-.061	Low	Mod
Memory for Designs – Delayed	.060	.206	Low	Mod
Memory for Names	-.007	-.139	Mod	High
Memory for Names – Delayed	.100*	.173	Mod	High
Animal Sorting – Combined	.050	.175	High	High
Design Fluency	-.130*	-.195	Mod	Low
Inhibition Switching – Combined	.231*	.355	Low	Low
D-KEFS				
Trail Making Test – Condition 4	-.112	-.195	Mod	High
Design Fluency – Condition 1	.007	.190	Low	High
Color Word Interference – Condition 4	-.118	-.067	High	High
Word Context	.184	.534	High	High
Verbal Fluency – Condition 1	.082*	-.041	High	High

For the first discriminant function, the subtests most associated included Spatial Relations, Sound Blending, Incomplete Words, Rapid Picture Naming, Memory for Words, Visual Auditory Learning, Retrieval Fluency, Numbers Reversed, Auditory Working Memory, Verbal Comprehension, General Information, Concept Formation, Analysis Synthesis, Response Set, Arrows, Geometric Puzzles, Speeded Naming, Comprehension of Instructions, Memory for Faces, Animal Sorting – Combined, and Color Word Interference – Condition 4. The subtests of Picture Recognition, Visuomotor Precision, Block Construction, Word List Interference, Memory for Names – Delayed, Design Fluency, Inhibition Switching – Combined, and Verbal Fluency Condition 1 were all most associated with the second discriminant function. Classification results revealed that 93.3% of the Caucasian/White group was correctly classified, 39.0% of the African-American/Black group were correctly classified, 34.7% of the Asian-American/Pacific Islander group were correctly classified, and 38.7% of the Latino/Hispanic group were correctly classified.

Summary

The purpose of this chapter was to present the results of the statistical procedures that were conducted to answer this study's two research questions. Archival data were drawn from a set of case studies submitted by students of the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program. The archival data supplied a mixed clinical sample of children and adolescents (*n* = 520) with various neurological and developmental disabilities. Descriptive statistics regarding the demographic characteristics for this sample were presented, as well as bivariate correlations for the selected subtests from the three neurocognitive measures included in this study (WJ III COG NU, NEPSY-II, D-KEFS). The means and standard deviations for the selected subtests were presented both as overall averages and separately by the four ethnicity groups.

Following the discussion of preliminary statistical analyses, the primary statistical analyses were described and results presented. To examine whether performance differences existed between ethnicity groups on three standardized measures of neurocognitive ability, a MANOVA was conducted using ethnicity as the IV and the selected subtests from the test batteries as the DVs. Results of this analysis indicated that there were, indeed, statistically significant differences in performance across the four ethnicity groups in the sample utilized. The post-hoc procedure used to further investigate these differences was a DFA. Results of the DFA were presented, revealing two statistically significant discriminant functions in which ethnicity group accounted for 26.2% and 18.7% of the variance, respectively.

CHAPTER V

DISCUSSION

The purpose of this chapter is to summarize the findings and implications of the study that has been conducted. It will begin with a reiteration of the purpose and goals of the study, followed by a discussion of conclusions that have been made for the research questions based on the results obtained. Implications for the field of psychology are proposed, as well as limitations of the study and recommendations for future research. This chapter will conclude with an overview of the study and its significance for the field of school psychology.

Purpose of the Study

There is evidence both in psychological practice and research literature that the tools and methods used to assess the neurocognitive abilities of CLD children are tainted with issues of validity. Performance on measures of neurocognitive ability is influenced by the level of acculturation and language proficiency of the individual being evaluated. Historically, the effect of language on cognitive assessment has been acknowledged and methods developed to mediate the influence on performance include administering the test in the examinee's native language, accepting answers in the examinee's native language, re-standardizing existing measures of neurocognitive abilities to include normative samples from non-English speaking populations, development of alternative forms, use of "nonverbal" tests and procedures, and development of tests in languages other than English. However, each approach brings with it various issues that challenge the reliability and validity of the test scores derived. For example, Laing and Kamhi (2003) point out that even with adjustments to normative samples, CLD children will continue to appear as though they perform below the mean of their age-matched peers due to factors such as unfamiliarity with test content or limited English proficiency (LEP). Furthermore, the cultural loading of neurocognitive tests has been even more neglected in research and practice.

Problems that arise from assessment of CLD individuals can include inappropriate test procedures, misinterpretations of test scores derived and test performance observations, inappropriate decision-making regarding referrals to special education, disproportionality in the identification of CLD students for special education programming, and inappropriate or ineffective intervention strategies employed (Ford et al., 2008; Jones et al., 2013; Laing & Kamhi, 2003; Skiba et al., 2008). However, multicultural competency among psychologists and other professionals working with CLD individuals remains considerably limited. Issues stemming from this area of weakness have been overlooked and/or trivialized by researchers and practitioners for decades (Skiba et al., 2008). It is speculated that reasons for this include a lack of specific knowledge for methods to ameliorate the issues, belief that current alternative procedures for cross-cultural assessment yield valid and reliable scores, inadequate methodologies applied in multicultural research studies, or the failure to recognize the significance of certain cultural and linguistic differences (Byrne et al., 2009; Ortiz, 2006; Vazquez-Nuttall et al., 2007). The consensus among recent researchers is that the underrepresentation of CLD students in gifted education and overrepresentation in special education may be largely attributed to these multicultural competency issues (Ford et al., 2008; Jones et al., 2013; Skiba et al., 2008).

While there have been a variety of approaches investigated and implemented with the intention of addressing the problems associated with the assessment of CLD individuals, the issue remains largely unresolved. However, one approach to the assessment of CLD individuals is a recent development by Flanagan and colleagues (2007, 2013) that has had a promising appearance for meeting the needs of this specialized type of assessment. The C-LTC and C-LIM comprise a framework for designing and interpreting an assessment protocol for a CLD individual. Classifications on the C-LTC are rated along two dimensions, cultural loading and linguistic demand, for a wide selection of commonly used neurocognitive assessment batteries. The C-LIM provides a software program that analyzes the influence of cultural loading and linguistic demand on test performance based on scores that the examiner inputs manually. The result is a determination by this program of the degree to which the test performance is influenced by cultural and linguistic differences or an actual reflection of intellectual ability.

While the C-LTC and C-LIM appear to be a promising approach to addressing issues with the assessment of CLD individuals, there is limited research evaluating its validity. The purpose of the current study was to investigate its validity by examining whether differences in performance exist across ethnicity groups for three standardized neurocognitive assessment batteries and then looking to see if differences were most pronounced for those subtests rated by the C-LTC to be high in cultural loading and linguistic demand. By conducting this study, the overall goal was to contribute useful information regarding the assessment of CLD individuals to the existing research literature, as well as provide practitioners with additional knowledge that may support better assessment practices and educational outcomes for CLD individuals.

Summary of Results

To investigate the effect of language demand and cultural loading on neurocognitive performance for CLD individuals, a study was conducted to review three specific instruments that are frequently used in psychological practice: WJ III COG NU, NEPSY-II, and D-KEFS. Two research questions were presented:

1. Do differences in performance exist based upon ethnicity for the three standardized (WJ III COG NU, NEPSY-II, D-KEFS) measures of neurocognitive functioning?
2. How much of the difference in performance is attributed to ethnicity?

To answer these research questions, data were drawn from an archival set of case studies submitted by students of the KIDS, Inc.'s School Neuropsychology Post-Graduate Certification Program. A mixed clinical sample (*n* = 520) of children and adolescents with neurological and developmental disabilities comprised this sample, and only cases that met the specified criteria for this study were included. To analyze the data, a MANOVA and a DFA were conducted using SPSS. It was expected that this study would reveal statistically significant differences in performance across ethnicity groups. Furthermore, the differences in performance based on ethnicity were expected to be more pronounced for subtests that are rated high for cultural loading and linguistic demand based on the C-LTC developed by Flanagan and colleagues (2007, 2013).

Preliminary Statistical Analyses

Prior to running the primary statistical analyses, the data were screened for normality, multicollinearity, and descriptive statistics were calculated. Information regarding these procedures was presented in both table and descriptive format. There were no issues found related to normality or multicollinearity in the sample used for this study. Means and standard deviations for the subtests were also provided. In terms of missing data, these data were screened prior to acquiring it for use in this study. Variables missing only 10% or less of values were imputed using multiple imputation (MI). Of the 911 cases in the archival data sample, 42.5% were missing ethnicity data. These cases were excluded from the study and the remaining 520 cases (*n* = 520) were included in the primary statistical analyses.

Of the cases included in the study, there were a larger percentage who were Caucasian/White, as compared to African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic. However, the sample was comprised of roughly the same number of males and females. While the proportion of males to females is ideal, the disproportionately high number of Caucasian/White participants in comparison to all of the other ethnicity groups is not ideal. This lack of proportionate representation of all ethnicity groups is particularly problematic in this study because ethnicity is the primary cultural variable that was used to investigate the research questions. As such, generalizations of the results should be made with awareness of the limitations that are associated. Further discussion of this limitation of the study will be included in a later section of this chapter.

The sample used in this study was comprised of children and adolescents with various clinical diagnoses, including LD, language disability, neurological impairment – acquired, ADHD, autism spectrum disorder, ED, general medical/OHI, and deaf/hearing impairment. Also included in the data analysis were pairs of comorbid clinical diagnoses, such as LD and ADHD, which were reported in the data sample. Clinical diagnoses for cases that were either not reported or unknown were included in the study. Those with clinical diagnoses consistent with an intellectual disability were excluded from the study. The frequencies of the clinical diagnoses reported in this sample are not necessarily representative of the frequency with which they are observed in the general population, as this sample is comprised of data from cases that met specific criteria for this study. For example, cases with missing ethnicity data were excluded from the study.

Overall means and standard deviations for the subtests selected for inclusion in this study from the WJ III COG NU, NEPSY-II, and D-KEFS were reported. All of the means fell in the average range for the subtests of the WJ III COG NU, while only 20 of the 22 NEPSY-II subtests and four of the five D-KEFS subtests had means that fell in the average range. For the NEPSY-II, subtests with mean scores that fell below average were Geometric Puzzles and Inhibition Switching –Combined. The D-KEFS subtest with a mean score below average was the Trail Making Test – Condition 4. Given that this study was conducted using a sample of children and adolescents with neurological and developmental disorders, is it surprising that more variability in performance was not observed.

Finally, preliminary statistical analyses for this study included bivariate correlations between the selected subtests for all three of the neurocognitive assessment batteries used in this study. Of the 85 correlations calculated among the subtests of the WJ III COG NU and the D-KEFS, 52 were found to be statistically significant. Approximately half of the 374 correlations calculated among the subtests of the WJ III COG NU and the NEPSY-II were found to be statistically significant. There were 110 correlations calculated among the subtests of the D-KEFS and the NEPSY-II, of which 36 were found to be statistically significant. These bivariate correlations were used to test for multicollinearity, which was not observed in the present study.

Primary Statistical Analyses

The purpose of this study was to investigate the influence of cultural loading and linguistic demand of neurocognitive assessment batteries on the test performance of CLD individuals. Using an archival sample of children and adolescents with mixed clinical disorders, a MANOVA was conducted. This procedure was used in determining whether differences in performance existed across four ethnicity groups (Caucasian/White, African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic) on selected subtests from three commonly used neurocognitive assessment batteries (WJ III COG NU, NEPSY-II, and D-KEFS). Though the use of ethnicity as a discriminating variable for cultural and linguistic differences was knowingly problematic, it was expected that differences in performance would, indeed, be observed. Results of the

MANOVA indicated that this hypothesis was correct, as there was a statistically significant multivariate effect across ethnicity groups.

The second research question to be answered was the degree to which ethnicity group could be attributed to the differences found. This information was drawn from effect size derived from the MANOVA, as well as the results from the follow-up DFA post-hoc procedure. The effect size for ethnicity on the MANOVA was 19.7%, and for the two statistically significant discriminant functions identified in the DFA, the variance was explained by 26.2% and 18.7%, respectively. These values are close enough that it could be concluded that ethnicity accounts for approximately 18.7-26.2% of the variability in test performance for this sample of children and adolescents.

What remains to be discussed is the influence of cultural loading and linguistic demand on these observed differences in performance based on the C-LTC and C-LIM framework developed by Flanagan and colleagues (2007, 2013). By examining the subtests that were found to have statistically significant differences across ethnicities in the DFA, it was possible to evaluate whether those that were rated high for cultural loading and linguistic demand were the more frequently discriminating variables.

With regard to the WJ III COG NU subtests included in this study, 13 were statistically significant predictors for the first discriminant function and one was a statistically significant predictor for the second discriminant function. The two subtests rated high for both cultural loading and linguistic demand, Verbal Comprehension and General Information, were significant predictors for the first discriminant function. The remaining 11 subtests that were significant predictors for the first discriminant function were all rated moderate or high for linguistic demand with the exception of Spatial Relations, which is rated low for both linguistic demand and cultural loading.

The absence of the influence of linguistic demand and cultural loading in the presence of significant performance differences on the Spatial Relations subtests suggests that there is a true difference in ability among ethnicity groups. This interpretation is guided by the C-LTC ratings provided by Flanagan and colleagues (2007, 2013); however, another explanation could be that the C-LTC ratings are not accurate for this subtest, the disproportionate sample sizes across ethnicity groups could have skewed the results, or some other variable that is yet to be seen is responsible for the significant differences.

For the WJ III COG NU subtests found to be statistically significant predictors for the two discriminant functions, the majority of them have C-LTC ratings of only low or moderate cultural loading. It could be the case that the subtests of the WJ III COG NU are simply not as plagued with cultural loading, or it might be that the linguistic demands of the subtests just weigh much more heavily on the performance for CLD individuals. A noteworthy observation of the DFA is that only one of the WJ III COG NU subtests and very few subtests from the NEPSY-II and D-KEFS were statistically significant predictors for the second discriminant function. However, a greater number of NEPSY-II and D-KEFS subtests were statistically significant predictors for the first discriminant function. Similar to the findings regarding the significant subtests from the WJ III COG NU, those significant subtests from the NEPSY-II and D-KEFS were also generally more highly rated for cultural loading and linguistic demand based on the C-LTC. It would seem that the first discriminant function could be influenced by these dimensions of cultural loading and linguistic demand, while the second discriminant function is influenced by some other unknown variable.

Interestingly, all of the D-KEFS subtests included in this study are rated high for linguistic demand and three out of the five are high for cultural loading, but only one subtest was found to be statistically significant for predicting group membership on each discriminant function based on ethnicity. This opens up for interpretation in a few directions. One explanation for this might be that the C-LTC ratings are not accurate for these subtests. Statistical significance might also not have been found due to the limitations of the sample used in this study. However, past research points to an explanation that the use of ethnicity as the sole discriminating demographic variable is insufficient for revealing the performance differences that may exist for CLD individuals.

Implications

Findings of this study must be interpreted with regard to the limitations it possesses; however, the conclusions drawn from the results may have some important implications for the field of psychology. Specifically, this study adds to a steadily growing body of school psychology research pertaining to finding more appropriate and valid assessment practices for CLD individuals. First and foremost, the completion of this study will serve to bring needed attention to the issues associated with cross-cultural assessment and research methods. This effect is expected regardless of the actual results of the study. For too long has this topic been minimized and addressed in superficial, unscientific methods. The presence of this research may also inspire further research into the methods and tools used to assess CLD individuals.

The significant findings of this study suggest that many neurocognitive subtests rated moderate or high for cultural loading and linguistic demand by Flanagan and colleagues (2007, 2013) on the C-LTC and C-LIM result in diminished performance by CLD individuals, as defined by those in the minority ethnicity groups included in the sample. While ethnicity group is a cultural variable limited by its mixed composition of various other cultural subgroups, these findings appear to suggest that the C-LTC ratings have some validity for use in CLD individuals. However, there were a number of instances in which subtests rated highly for both cultural loading and linguistic demand were not found to be statistically significantly different across ethnicity in this study. For example, many of the D-KEFS subtests are rated high on both dimensions, but only one was a significant predictor for each of the two discriminant functions identified in the DFA. This may call into question the accuracy of ratings for some subtests on the C-LTC. Another possibility is that the quality of the sample, in terms of demographic and cultural representation, affected the ability to reveal differences in performance that might have otherwise been observed.

The completion of this study should also serve to reinforce the importance of gathering a thorough background and cultural/language proficiency profile of every individual for which neurocognitive assessment is intended. Should an examiner have information that would suggest the examinee has a different cultural or linguistic background that might affect test performance, careful selection of tests for an assessment protocol is critical. This study emphasizes the negative implications that may occur when tests with high cultural loading and linguistic demand are used with CLD individuals. The C-LTC and C-LIM framework developed by Flanagan and colleagues (2007, 2013), though still in need of additional research to substantiate its validity, may serve as a useful tool in test selection for CLD individuals. However, findings of this study indicate that caution should still be exercised in the interpretation of neurocognitive test scores of CLD individuals, even with the use of tools like the C-LTC and C-LIM.

Methodological Issues/Limitations

As with any other research study, the one conducted here was subject to a variety of methodological issues and limitations. There were several issues of reliability and validity that were addressed prior to and after conducting this study. Concerns regarding the reliability of assessment batteries that measure neurocognitive functioning influenced the decision regarding the measures to include in the analysis (Jackson et al., 2006). Each of the scales used to assess the DVs must have demonstrated good test-retest, interrater, and alternative form (if applicable) reliability within the associated literature. Typically, Cronbach's alphas of 0.7 and higher can indicate adequate reliability (Frost et al., 2007). Unfortunately, there was no way to control for the reliable and accurate administration of the measures to the children in the sample, due to the archival data gathering method (Jackson et al., 2006). The researcher did not have the ability to ensure that the scales/measures were used reliably across all participants.

Reliability of a measure does not guarantee that it measures the construct it purports to measure (Frost et al., 2007). The measures included should also have demonstrated good construct, content, convergent, and divergent validity within the associated literature. As such, the construct validity should show that the scale is measuring the construct that it is intended to measure, reliably, of course. Content validity involves the degree to which an instrument measures a broad, representative sample of the construct it intends to measure; low content validity is indicative of narrow stimulus sampling (Jackson et al., 2006). Convergent validity suggests that the scale measures the construct just as well as another scale that is deemed to be measuring the same thing. Finally, divergent validity allows the researcher to be confident that the scale does not measure some unrelated construct, as demonstrated by low correlations with some other scale measuring something different.

Based upon the fact that this study used preexisting groups (no random assignment) of participants, there were some aspects of the study for which there was little or no control. For this reason, it was difficult to have confidence that the constructs were not due to some inherent differences in the sample (i.e., cultural differences). Thus, internal validity was threatened. Also, the external validity of this study was limited, due to the specific population that was targeted for the sample. For this reason, the ability for the results of this study to be generalized was constrained to the population for which the sample was drawn.

There are a number of limitations that are inherent in a research design such as this. The nonexperimental nature and use of archival data precluded the researcher from the opportunity to control for various threats to internal and external validity (Jackson et al., 2006). To begin, the researcher was limited by the demographic characteristics and scores that were reported in the data set. Reliance on the accurate administration, scoring, and data entry by the professionals who performed the evaluations was imperative (Mertler & Vannatta, 2010). Further, there was no control over the accuracy or scope of information the data set includes about the sample's demographic characteristics.

One of the major limitations to the present study is with regard to the demographic information available to use as variables to discriminate among CLD individuals in the sample. The only cultural variables included in the archival data were the ethnicity and gender of the participant. As the literature review pointed out, using ethnicity as a cultural variable is problematic due to the tendency for ethnic groupings to be comprised of individuals with diverse linguistic or other cultural backgrounds (Flanagan et al., 2013; Gasquoine, 2009). With the limited range of demographic cultural variables included in the archival data used in this study, ethnicity was selected to be the sole discriminating independent variable. Not only does this choice neglect the influence of language and language proficiency, but ethnicity categories are plagued with inconsistent definitions and do not account for other more significant cultural idiosyncrasies that may influence test performance (Flanagan et al., 2013; Gasquoine, 2009). Furthermore, some individuals may ascribe to more than one ethnic category, and multiple ethnicity group membership may bring with it its own cultural distinctiveness. For this reason, as well as the limited sample size, those with multiple ethnicities were excluded from this study. It was suspected that should ethnicity not prove to differentiate performance among participants in this study, it would be as a result of cultural heterogeneity in each group.

Primary language spoken by the participants, as well as level of English-language proficiency, was important information that could have augmented this study; however, this was information that the archival data did not possess. Furthermore, some form of demographic variables pertaining to the participants' level of acculturation (i.e., immigration generation status, years residing in the U.S.) would have provided the opportunity to include yet another independent variable in which to look for performance differences on tests of neurocognitive functioning. This study suggests that information regarding a client's language and acculturation should be collected prior to testing CLD individuals, because these factors are more informative than ethnicity alone.

In addition to the limitation of using ethnicity as the sole cultural variable in this study, the disproportionate ratio of Caucasian/White participants to African-American/Black, Asian-American/Pacific Islander, and Latino/Hispanic participants affects the generalizability of the findings. The smaller sample sizes for three of the ethnicity groups included translates to a lack of representativeness for those populations (Williams & Cottle, 2011). Furthermore, even with a sample comprised of proportionate groups of each ethnicity group, the influence of language and acculturation remains unaddressed.

Limitations with regard to the clinical sample may additionally be due to the potential that there are inaccuracies in clinical diagnosis of participants. This would limit the researcher's ability to make inferences based on comparisons made between clinical sample groups. There was also a substantial portion of the archival data that was missing ethnicity descriptors of participants. Reasons for this are unknown to the researcher;

however, the identities of said participants may be related to this fact that the information was not obtained. Nonetheless, this unfortunate information gap could have an influence on the outcome of this study. Furthermore, by not investigating the possibility of differences between clinical groups, the researcher cannot know whether this is a significant factor influencing differences in performance on these test batteries. Future research into the ratio of different ethnic groups within specific clinical diagnoses might provide useful information that might explain patterns of performance observed in CLD individuals.

In addition to the limited cultural information of the archival sample used in this study, the amount of data that needed to be imputed contributes to the caution that must be taken with regard to interpretations of results. Missing data for some cases and variables were able to be imputed if less than 10% of values were absent; however, there were some subtests that had to be excluded altogether due to the absence of a much larger percentage of values. Furthermore, because this data was retrieved after it had already been imputed, information regarding patterns of missing data was not available. As such, there is no way to know if patterns of missing data had any influence on the results obtained in this study, which might be the case if there were significantly more missing data for participants of minority group status. Access to this information would have proved useful for guiding interpretations of this study's results.

The inclusion of Flanagan and colleagues' (2007, 2013) Culture-Language Interpretive Matrix (C-LIM) for this study confines the present investigation to the cultural loading and linguistic demand paradigm that is the underlying framework. The research supporting the validity and reliability for the C-LIM is relatively limited, primarily due to its development only dating back less than ten years. Should there be significant flaws in the theory or some structural element pertaining to this framework, the findings of the present study would be similarly affected.

Recommendations for Future Research

To further evaluate the validity of the C-LTC matrices, researchers should consider using confirmatory factor analysis (CFA), as it may assist in determining model fit based on sample test data (Mertler & Vannatta, 2010). It is commonly used to assess theories involving one or more latent variables believed to exist based on some set of observable variables. This approach would be similar to the research conducted in the dissertation of Damien Cormier (2012). Cormier used path analysis and structural equation modeling to investigate the validity of the C-LIM and C-LTC using a cross-sectional approach with the WJ III NU normative sample. The results of this study showed that some of the C-LIM classifications were upheld; however, it was proposed by Cormier that the subtests of the WJ III COG NU undergo a re-classification based upon the results of his study.

Should a similar study to the current one be conducted in the future using ethnicity group as a demographic variable, it might be useful to determine if differences in neurocognitive performance persist even after removing the variance associated with cultural loading and linguistic demand. The information this endeavor might reveal could support the notion that the use of ethnicity groups as a discriminating demographic variable for CLD individuals is not useful for revealing the many nuances inherent in these populations. One such approach to evaluating how much difference observed may be attributed to each dimension of cultural loading and linguistic demand could be the use of a path analysis. This statistical procedure would also allow the researcher to determine if the exclusion of variance due to high levels of cultural loading and linguistic demand would continue to reveal differences in performance across ethnicity groups for a sample of CLD individuals.

Another area of research that could provide some validity for the C-LTC might be in examining the intercorrelations of subtests classified within each cell of the classification tables. For example, one might be interested to see if subtests rated low for both cultural loading and linguistic demand have strong correlations among other similarly rated subtests. This information would lend support to the classifications made by Flanagan and colleagues (2013) and be a good next step in validating use of the C-LTC in practice.

Conclusion

The contribution of this proposed study to the field of psychology is the possibility for a better understanding of the effects of culture and language on test performance with CLD individuals. Additionally, this knowledge may serve to enlighten practitioners currently conducting cross-cultural intelligence evaluations with regard to the interpretations and recommendations being made in practice. Of course, another possible outcome from this study may be the foundation for further research into the issues associated with neurocognitive assessment of CLD individuals. This study was necessary for progress in the development of more culturally fair assessment practices in the field of school psychology.

REFERENCES

Aud, S., Wilkinson-Flicker, S., Kristapovich, P., Rathbun, A., Wang, X., & Zhang, J. (2013). The Condition of Education 2013 (NCES 2013-037). U.S. Department of Education, National Center for Education Statistics. Washington, DC. Retrieved [2/5/14] from http://nces.ed.gov/pubsearch.

Baron, I. S. (2004). Test review: Delis-Kaplan Executive Function System. *Child Neuropsychology, 10*(2), 147-152.

Benjamin, L. T., Jr. (2009). *A history of psychology: Original sources and contemporary research.* Malden, MA: Blackwell.

Berg, E. A. (1948). A simple objective technique for measuring flexibility in thinking. *Journal of General Psychology, 39*, 15–22.

Berry, J. W. (1997). Immigration, acculturation, and adaption. *Applied Psychology: An International Review, 46*(1), 5-68.

Binet, A. & Simon, T. (1905). Méthodes nouvelles pour le diagnostic du niveau intellectual des anormaux [New methods for the diagnosis of the intellectual level of abnormals]. *L'Année Psychologique, 11*, 191-336.

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

Brooks, B. L., Sherman, E., & Strauss, E. (2010). Test review: NEPSY-II: A Developmental Neuropsychological Assessment, Second Edition. *Child Neuropsychology, 16*, 80-101. doi:10.1080/09297040903146966

Burnham, W. H. (1888). Memory, historically and experimentally considered. I. An historical sketch of the older conceptions of memory. *The American Journal of Psychology, 2*(1), 39-90.

Byrne, B. M., Oakland, T., Leong, F. T. L., van de Vijver, F. J. R., Hambleton, R. K., & Cheung, F. M. (2009). A critical analysis of cross-cultural research and testing practices: Implications for improved education and training in psychology. *Training and Education in Professional Psychology, 3*(2), 94-105. doi:10.1037/a0014516

Carter, R., Aldridge, S., Page, M., & Parker, S. (2009). *The human brain book: An illustrated guide to its structure, function, and disorders.* New York, NY: Dorling Kindersley Limited.

Chen, F. F. (2008). What happens if we compare chopstick with forks? The impact of making inappropriate comparisons in cross-cultural research. *Journal of Personality and Social Psychology, 95*(5), 1005-1018. doi:10.1037/a0013193

Cohen, M. J. (1997). *Children's Memory Scale*. San Antonio, TX: Harcourt.

Cohen, R. J. & Swerdlik (2005). *Psychological testing and assessment: An introduction to tests and measurement* (6th ed.). New York, NY: McGraw-Hill.

Cormier, D. C. (2012). *The influences of linguistic demand and cultural loading on cognitive test scores* (Doctoral dissertation). Available from ProQuest Dissertations and Theses database. (Order No. 3513323)

Cormier, D. C., McGrew, K. S., & Evans, J. J. (2011). Quantifying the "degree of linguistic demand" in spoken intelligence test directions. *Journal of Psychoeducational Assessment, 29*(6), 515-533. doi:10.1177/0734282911405962

Dalen, K., Jellestad, F., & Kamaloodien, K. (2007). The translation of the NEPSY-II to Afrikaans, some ethical reflections. *Cognition, Brain, & Behavior, 11*(3), 609-620.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001a). *Delis-Kaplan Executive Function System*. San Antonio, TX: Psychological Corporation.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001b). *Delis-Kaplan Executive Function System: Examiner's manual*. San Antonio, TX: Psychological Corporation.

Delis, D. C., Kaplan, E., & Kramer, J. H. (2001c). *Delis-Kaplan Executive Function System: Technical manual*. San Antonio, TX: Psychological Corporation.

Delis, D. C., Kramer, J., Kaplan, E., & Ober, B. A. (2000). *California Verbal Learning Test – Second Edition.* San Antonio, TX: Psychological Corporation.

Duckworth, A. L., Quinn, P. D., & Tsukayama, E. (2012). What *No Child Left Behind* leaves behind: The roles of IQ and self-control in predicting standardized achievement test scores and report card grades. *Journal of Educational Psychology, 104*(2), 439-451. doi:10.1037/a0026280

Dugbartey, A. T. (2011). Review of the Delis-Kaplan Executive Function System. Retrieved from http://ezproxy.twu.edu:2182/sp-3.4.1b/ovidweb.cgi?&S=GAPFFPNKEODDGACHNCCLKBOBKMKPAA00&Complete+Reference=S.sh.14%7c1%7c1

Elliott, C. D. (2007). *Differential Ability Scales* (2nd ed.). San Antonio, TX: Psychological Corporation.

Enders, C. K. (2003). Performing multivariate group comparisons following a statistically significant MANOVA. *Measurement and Evaluation in Counseling and Development, 36*, 40-56.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2007). *Essentials of cross-battery assessment* (2nd ed.). Hoboken, NJ: Wiley.

Flanagan, D. P., Ortiz, S. O., & Alfonso, V. C. (2013). *Essentials of cross-battery assessment* (3rd ed.). Hoboken, NJ: Wiley.

Floyd, R. G. & Kranzler, J. H. (2012). Processing approaches to interpretation of information from cognitive ability tests. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 422-435). New York, NY: Guilford Press.

Ford, D. Y., Moore, J. L., III, Whiting, G. W., & Grantham, T. C. (2008). Conducting cross-cultural research: Controversy, cautions, concerns, and considerations. *Roeper Review, 30*, 82-92. doi:10.1080/02783190801954924

Fraine, N. & McDade, R. (2009). Reducing bias in psychometric assessment of culturally and linguistically diverse students from refugee backgrounds in Australian schools: A process approach. *Australian Psychologist, 44*(1), 16-26. doi:10.1080/00050060802582026

Frost, M. H., Reeve, B. B., Liepa, A. M., Stauffer, J. W., Hays, R. D. & the Mayo/FDA Patient-Reported Outcomes Consensus Meeting Group (2007). What is sufficient evidence for the reliability and validity of patient-reported outcome measures? *Value in Health, 10*(2), 94-105.

Garrett, B. (2009). *Brain & behavior: An introduction to biological psychology* (2nd ed.). Thousand Oaks, CA: Sage.

Gasquoine, P. G. (2009). Race-norming of neuropsychological tests. *Neuropsychology Review, 19*, 250-262. doi:10.1007/s11065-009-9090-5

Goddard, H. H. (1913). The Binet tests in relation to immigration. *Journal of Psycho-Asthenics, 18*, 105-107.

Gottfredson, L. & Saklofske, D. H. (2009). Intelligence: Foundations and issues in assessment. *Canadian Psychology, 50*(3), 183-195. doi:10.1037/a0016641

Gravetter, F. J. & Forzano, L. B. (2009). *Research methods for the behavioral sciences* (3rd ed.). Boston, MA: Wadsworth, Cengage Learning.

Grice, J. W. & Iwasaki M. (2007). A truly multivariate approach to MANOVA. *Applied Multivariate Research, 12*(3), 199-226.

Guthrie, R. V. (2004). *Even the rat was white: A historical view of psychology*. Boston, MA: Pearson Education, Inc.

Hadzi-Pavlovic, D. (2010). Multivariate analysis of variance. *Acta Neuropsychiatrica, 22*, 257-258. doi:10.1111/j.1601-5215.2010.00488.x

Heaton, R. K., Chelune, G. J., Talley, J. L., Kay, G. G., & Curtiss, G. (1993). *Wisconsin Card Sorting Test Manual*. Odessa, FL: Psychological Assessment Resources.

Helms, J. E. (2006). Fairness is not validity or cultural bias in racial-group assessment: A quantitative perspective. *American Psychologist, 61*(8), 845-859. doi:10.1037/0003-066X.61.8.845

Homack, S., Lee, D., & Riccio, C. A. (2005). Test review: Delis-Kaplan Executive Function System. *Journal of Clinical and Experimental Neuropsychology, 27*, 599-609. doi:10.1080/13803390490918444

Humes, K. R., Jones, N. A., & Ramirez, R. R. (2011). Overview of race and Hispanic origin: 2010. *2010 Census Briefs, United States Census Bureau.* Retrieved from http://www.census.gov/prod/cen2010/briefs/c2010br-02.pdf

Individuals with Disabilities Education Act of 1997, Pub. L. 105-117.

Individuals with Disabilities Education Improvement Act of 2004, Pub. L. 108-446.

Jääskeläinen, M. (1998). Bridging intellectual and cultural gaps: Adler and Aristotle. *The Journal of Individual Psychology, 54*(3), 324-335.

Jaccard, J. & Guilamo-Ramos, V. (2002). Analysis of variance frameworks in clinical child and adolescent psychology: Issues and recommendations. *Journal of Clinical Child and Adolescent Psychology, 31*(1), 130-146.

Jackson, S. L., Mitchell, & Jolley (2006). *Research methods and statistics: A critical thinking approach*. Mason, OH: Thomson Wadsworth.

Jones, J. M., Sander, J. B., & Booker, K. W. (2013). Multicultural competency building: Practical solutions for training and evaluating student progress. *Training and Education in Professional Psychology, 7*(1), 12-22. doi:10.1037/a0030880

Kamphaus, R. W., Winsor, A. P., Rowe, E. W., & Kim, S. (2012). A history of intelligence test interpretation. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 422-435). New York, NY: Guilford Press.

Kempert, S., Saalbach, H., & Hardy, I. (2011). Cognitive benefits and costs of bilingualism in elementary school students: The case of mathematical word problems. *Journal of Educational Psychology, 103*(3), 547-561. doi:10.1037/a0023619

Korkman, M., Kirk, U., & Kemp, S. (1998). *NEPSY: A Developmental Neuropsychological Assessment*. San Antonio, TX: Psychological Corporation.

Korkman, M., Kirk, U., & Kemp, S. (2007). *NEPSY-II: A Developmental Neuropsychological Assessment: Clinical and Interpretive Manual*. San Antonio, TX: Psychological Corporation.

Laing, S. P. & Kamhi, A. (2003). Alternative assessment of language and literacy in culturally and linguistically diverse populations. *Language, Speech, and Hearing Services in Schools, 34*(1), 44-55. doi:10.1044/0161-1461(2003/005)

Lakin, J. M. & Lai, E. R. (2012). Multi-group generalizability analysis of verbal, and nonverbal ability tests for culturally and linguistically diverse students. *Educational and Psychological Measurement, 72*(1), 139-158.

Luria, A. R. (1980). *Higher cortical functions in man* (2nd ed.). New York: Basic Books.

Malda, M., van de Vijver, F. J. R., & Temane, Q. M. (2010). Rugby versus soccer in South Africa: Content familiarity contributes to cross-cultural differences in cognitive test scores. *Intelligence, 38*, 582-595. doi:10.1016/j.intell.2010.07.004

Maricle, D. E. & Avirett, E. (2012). The role of cognitive and intelligence tests in the assessment of executive functions. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 422-435). New York, NY: Guilford Press.

Matthews, R. N., Riccio, C. A., & Davis, J. L. (2012). The NEPSY-II. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (3rd ed., pp. 422-435). New York, NY: Guilford Press.

McGrew, K. S. (2005). The Cattell-Horn-Carroll theory of cognitive abilities. In D. P. Flanagan & P. L. Harrison (Eds.), *Contemporary intellectual assessment: Theories, tests, and issues* (2nd ed., pp. 136-181). New York, NY: The Guilford Press.

McGrew, K. S., Schrank, F. A., & Woodcock, R. W. (2007). Technical Manual. *Woodcock-Johnson III Normative Update.* Rolling Meadows, IL: Riverside.

McGrew, K. S. & Woodcock, R. W. (2001). Technical Manual. *Woodcock-Johnson III Tests of Cognitive Abilities*. Rolling Meadows, IL: Riverside.

Mertler, C. A. & Vannatta, R. A. (2010). *Advanced and multivariate statistical methods: Practical application and interpretation* (4th ed.). Glendale, CA: Pyrczak Publishing.

Meyer, L. & Melchert, T. P. (2011). Examining the content of mental health intake assessments from a biopsychosocial perspective. *Journal of Psychotherapy Integration, 21*(1), 70-89.

Miller, D. C. (2007). *Essentials of school neuropsychological assessment*. Hoboken, NJ: Wiley.

Miller, D. C. (2010). *Best practices in school neuropsychology: Guidelines for effective practice, assessment, and evidence-based intervention*. Hoboken, NJ: John Wiley & Sons, Inc.

Miller, D. C. (2013). *Essentials of school neuropsychological assessment* (2nd ed.). Hoboken, NJ: Wiley.

Mpofu, E. & Ortiz, S. O. (2009). Equitable assessment practices in diverse contexts. In E. L. Grigorenko (Ed.) *Multicultural psychoeducational assessment* (pp. 41-76). New York, NY: Springer Publishing Company.

Muñoz-Sandoval, A. F., Woodcock, R. W., McGrew, K. S., & Mather, N. (2004, 2007). *Batería III Woodcock-Muñoz: Pruebas de habilidades cognitivas.* Rolling Meadows, IL: Riverside.

Nampija, M., Apule, B., Lule, S., Akurut, H., Muhangi, L., Elliott, A. M., & Alcock, K. J. (2010). Adaptation of Western measures of cognition for assessing 5-year-old semi-urban Ugandan children. *British Journal of Educational Psychology, 80*(1), 15-30.

No Child Left Behind Act (NCLB) of 2001, Pub. L. 107-115, 20 U.S.C. 6301 (2001).

Oakland, T. (2009). How universal are test development and use? In E. L. Grigorenko (Ed.) *Multicultural psychoeducational assessment* (pp. 1-40). New York, NY: Springer Publishing Company.

Olvera, P. & Gomez-Cerrillo, L. (2011). A bilingual (English & Spanish) psychoeducational assessment model grounded in Cattell-Horn-Carroll (CHC) theory: A cross battery approach. *Contemporary School Psychology,* 15, 117-127.

Ortiz, S. O. (2006). Multicultural issues in school psychology practice: A critical analysis. *Journal of Applied School Psychology, 22*(2), 151-167.

Ortiz, S. O. & Flanagan, D. P. (1998). *Gf-Gc* cross-battery interpretation and selective cross-battery assessment: Referral concerns and the needs of the culturally and linguistically diverse populations. In K. S. McGrew & D. P. Flanagan, *The intelligence test desk reference (ITDR): Gf-Gc cross-battery assessment* (pp. 401-444). Boston, MA: Allyn & Bacon.

Ortiz, S. O., Ochoa, S. H., & Dynda, A. M. (2012). Testing with culturally and linguistically diverse populations: Moving beyond the verbal-performance dichotomy into evidence-based practice. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed., pp. 526-552). New York, NY: Guilford Press.

Packard, N. & Chen, C. (2005). From medieval mnemonics to a social construction of memory. *American Behavioral Scientist, 48*, 1297-1319. doi:10.1177/0002764205277010

Rembis, M. A. (2004). “I ain’t been reading while on parole”: Experts, mental tests, and Eugenic Commitment Law in Illinois, 1890-1940. *History of Psychology, 7*(3), 225-247. doi:10.1037/1093-4510.7.3.225

Repovš, G. & Baddeley, A. (2006). The multi-component model of working memory: Explorations in experimental cognitive psychology. *Neuroscience, 139*, 5-21. doi:10.1016/j.neuroscience.2005.12.061

Reschly, D. J. (2009). Minority special education disproportionality: Findings and misconceptions. Minorities in Special Education [Briefing Report], 57-66.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys*. New York, NY: John Wiley & Sons, Inc.

Schaefer, R. T. (2006). *Sociology matters* (2nd ed.). New York, NY: McGraw-Hill.

Schölmerich, A., Leyendecker, B., Citlak, B., Caspar, U., & Jäkel, J. (2008). Assessment of migrant and minority children. *Journal of Psychology, 216*(3), 187-194. doi:10.1027/0044-3409.216.3.187

Schrank, F. A., Miller, D. C., Wendling, B. J., & Woodcock, R. W. (2010). *Essentials of WJ III Cognitive Abilities assessment* (2nd ed.). Hoboken, NJ: Wiley.

Schrank, F. A., Wendling, B. J., & Woodcock, R. W. (2008). *Woodcock Interpretation and Instructional Interventions Program* (WIIIP, Version 1.0) [Computer software]. Rolling Meadows, IL: Riverside.

Schrank, F. A. & Woodcock, R. W. (2008). *Woodcock-Johnson III Normative Update Compuscore and Profiles Program* (Version 3.1) [Computer software]. Rolling Meadows, IL: Riverside.

Schumacker, R. E. & Lomax, R. G. (2004). *A beginner's guide to structural equation modeling* (2nd ed.). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Shifrer, D., Muller, C., & Callahan, R. (2011). Disproportionality and learning disabilities: Parsing apart race, socioeconomic status, and language. *Journal of Learning Disabilities, 44*, 246-257. doi:10.1177/0022219410374236

Shunk, A. W., Davis, A. S., & Dean, R. S. (2006). Test review: Dean C. Delis, Edith Kaplan & Joel H. Kramer, *Delis Kaplan Executive Function System (D-KEFS)*, The Psychological Corporation, San Antonio, TX, 2001. $415.00 (complete kit). *Applied Neuropsychology, 13*(4), 275-279. doi:10.1207/s15324826an1304_9

Skiba, R. J., Simmons, A. B., Ritter, S., Gibb, A. C., Rausch, M. K., Cuadrado, J., & Chung, C. (2008). Achieving equity in special education: History, status, and current challenges. *Exceptional Children, 74*(3), 264-288.

Society for the Psychological Study of Social Issues (SPSSI) (1938). *Statement on racial psychology.* Washington, D.C.: American Psychological Association.

Steele, C. M. & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology, 69*(5), 797-811.

Sternberg, R. J. (2007). Who are the bright children? The cultural context of being and acting intelligent. *Educational Researcher, 36*(3), 148-155. doi:10.3102/0013189X07299881

Sternberg, R. J. (2012). The triarchic theory of successful intelligence. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed., pp. 156-177). New York, NY: Guilford Press.

Tan, M. T., Aljughaiman, A. M., Elliott, J. G., Kornilov, S. A., Ferrando-Prieto, M., Bolden, D. S., …Grigorenko, E. L. (2009). Considering language, culture, and cognitive abilities: The international translation and adaptation of the Aurora Assessment Battery. In E. L. Grigorenko (Ed.) *Multicultural psychoeducational assessment* (pp. 443-468). New York, NY: Springer Publishing Company.

U.S. Census Bureau (2005). *Census 2000 Summary File 3 United States.* Washington, DC: Author.

Van de Vijver, F. & Phalet, K. (2004). Assessment in multicultural groups: The role of acculturation. *Applied Psychology: An International Review, 53*(2), 215-236.

Van de Vijver, F. & Tanzer, N. K. (2004). Bias and equivalence in cross-cultural assessment: An overview. *Revue Européenne de Psychologie Appliquée, 54*(2), 119-135. doi:10.1016/j.erap.2003.12.004

Vazquez-Nuttall, E., Li, C., Dynda, A. M., Ortiz, S. O., Armengol, C. G., Walton, J. W., & Phoenix, K. (2007). Cognitive assessment of culturally and linguistically diverse students. In G. B. Esquivel, E. C. Lopez, & S. Nahari (Eds.) *Handbook of multicultural school psychology: An interdisciplinary perspective* (pp. 265-288). Mahwah, NJ: Lawrence Erlbaum Associates, Inc.

Walker, A. J., Batchelor, J., & Shores, A. (2009). Effects of education and cultural background on performance on WAIS-III, WMS-III, WAIS-R and WMS-R measures: Systematic review. *Australian Psychologist, 44*(4), 216-223.

Wang, Q. (2008). Emotion knowledge and autobiographical memory across the preschool years: A cross-cultural longitudinal investigation. *Cognition, 108*(1), 117-135.

Wasserman, J. D. (2012). A history of intelligence assessment: The unfinished tapestry. In D. P. Flanagan & P.L. Harrison (Eds.) *Contemporary intellectual assessment: Theories, tests and issues* (3rd ed., pp. 3-55). New York, NY: Guilford Press.

Wechsler, D. (2001). *Wechsler Individual Achievement Test* (2nd ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2003). *Wechsler Intelligence Scale for Children* (4th ed.). San Antonio, TX: Psychological Corporation.

Wechsler, D. (2005). *Wechsler Intelligence Scale for Children – Fourth Edition Spanish.* San Antonio, TX: Psychological Corporation.

Wechsler, D. & Naglieri, J. A. (2006). *Wechsler Nonverbal Scale of Ability.* San Antonio, TX: Psychological Corporation.

Wicherts, J. M., Dolan, C. V., & Hessen, D. J. (2005). Stereotype threat and group differences in test performance: A question of measurement invariance. *Journal of Personality and Social Psychology, 89*(5), 696-716. doi:10.1037/0022-3514.89.5.696

Williams, J. M. & Cottle, C. C. (2011). A correction for recruitment bias in norms derived from meta-analysis. *Psychological Assessment*. doi:10.1037/a0023651

Woodcock, R. W., McGrew, K. S., & Mather, N. (2001, 2007). *Woodcock-Johnson III*. Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W., McGrew, K. S., Mather, N., & Schrank, F. A. (2003, 2007). *Woodcock-Johnson III Diagnostic Supplement to the Tests of Cognitive Abilities.* Rolling Meadows, IL: Riverside Publishing.

Woodcock, R. W. & Muñoz-Sandoval, A. F. (2001). *Woodcock-Muñoz Language Survey Normative Update*. Chicago, IL: Riverside Publishing.

Yerkes, R. M. (1921). Development of a substitute group test for illiterates and foreigners. In R. M. Yerkes (Ed.), *Memoirs of the National Academy of Sciences: Vol. 15: Psychological examining in the United States Army* (pp. 363-377). Washington, DC: U.S. Government Printing Office. doi:10.1037/10619-011